Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES

Volume 45

A Festschrift for Herman Rubin

Anirban DasGupta, Editor

Institute of Mathematical Statistics
Beachwood, Ohio, USA

ISBN 0-940600-61-7

Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES

A Festschrift for Herman Rubin

Anirban DasGupta, Editor

Volume 45

Institute of Mathematical Statistics
Lecture Notes–Monograph Series

Series Editor:
Joel Greenhouse

The production of the *Institute of Mathematical Statistics Lecture Notes–Monograph Series* is managed by the IMS Societal Office: Julia A. Norton, Treasurer and Elyse Gustafson, Executive Director.

Library of Congress Control Number: 2004109892

International Standard Book Number 0-940600-61-7

Copyright © 2004 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

Contents

Preface v

Contributors vi

List of publications viii

Some reminiscences of my friendship with Herman Rubin
Herman Chernoff 1

Evaluating improper priors and the recurrence of symmetric Markov chains: An overview
Morris L. Eaton 5

Estimation in restricted parameter spaces: A review
Eric Marchand and William E. Strawderman 21

A Rubinesque theory of decision
J. B. Kadane, Mark J. Schervish and Teddy Seidenfeld 45

On the distribution of the greatest common divisor
Persi Diaconis and Paul Erdös 56

Versions of de Finetti's Theorem with applications to damage models
C. R. Rao and D. N. Shanbhag 62

A short history of stochastic integration and mathematical finance: The early years, 1880–1970
Robert Jarrow and Philip Protter 75

Non-linear filtering with Gaussian martingale noise: Kalman filter with fBm noise
L. Gawarecki and V. Mandrekar 92

Self-similar processes, fractional Brownian motion and statistical inference
B.L.S. Prakasa Rao 98

Some properties of the arc-sine law related to its invariance under a family of rational maps
Jim Pitman and Marc Yor 126

On time changing continuous martingales to Brownian motion
Burgess Davis 138

On counts of Bernoulli strings and connections to rank orders and random permutations
Jayaram Sethuraman and Sunder Sethuraman 140

Chebyshev polynomials and G-distributed functions of F-distributed variables
Anirban DasGupta and L. Shepp 153

Zeroes of infinitely differentiable characteristic functions
Herman Rubin and Thomas M. Sellke 164

On the characteristic function of Pearson type IV distributions
Wei-Liem Loh 171

Characterizations, Sub and resampling, and goodness of fit
L. Brown, Anirban DasGupta, John Marden and Dimitris Politis 180

Notes on the bias-variance trade-off phenomenon
Jeesen Chen 207

Combining correlated unbiased estimators of the mean of a normal distribution
Timothy Keller and Ingram Olkin 218

An asymptotic minimax determination of the initial sample size in a two-stage sequential procedure
Michael Woodroofe 228

Estimating gradient trees
Ming-Yen Cheng, Peter Hall and John A. Hartigan 237

Conservative bounds on extreme P-values for testing the equality of two probabilities based on very large sample sizes
Herman Chernoff 250

Detecting a target in very noisy data from multiple looks
Jiashun Jin 255

r-scan extremal statistics of inhomogeneous Poisson processes
Samuel Karlin and Chingfer Chen 287

On the strong consistency, weak limits and practical performance of the ML estimate and Bayesian estimates of a symmetric domain in R^k
Wen-Chi Tsai and Anirban DasGupta 291

Maximum likelihood estimation for the contact process
Marta Fiocco and Willem R. van Zwet 309

On the "Poisson boundaries" of the family of weighted Kolmogorov statistics
Leah Jager and Jon A. Wellner 319

A theorem on compatibility of systems of sets with applications
A. Goswami and B.V. Rao 332

A question of geometry and probability
Richard A. Vitale 337

Generalized Accept–Reject sampling schemes
George Casella, Christian P. Robert and Martin T. Wells 342

Scalable mining for classification rules in relational databases
Min Wang, Bala Iyer and Jeffrey Scott Vitter 348

A simple proof of a condition for cointegration
T. W. Anderson 378

Forecasting NBA basketball playoff outcomes using the weighted likelihood
Feifang Hu and James V. Zidek 385

Distributions of failure times associated with non-homogeneous compound Poisson damage processes
S. Zacks 396

Conversations with Herman Rubin
Mary Ellen Bock 408

Preface

First and foremost, I would like to express my deep gratitude to the distinguished authors of this volume, to Joel Greenhouse and Rick Vitale, the Editors of the IMS Lecture Notes and Monograph Series, to Elyse Gustafson, to the referees of the papers, and to the people at Mattson Publishing Services and VTEX who worked dedicatedly and diligently in order to make this a reality. This book is a team effort. I must express my very special thanks to one person: Geri Mattson. I could write a poetic paragraph to thank her; let me simply say she is unreal. I am also thankful to Kristy Brewer, Rebekah Holmes, Teena Seele and Norma Lucas for editorial support at Purdue University. They were smilingly helpful at all times. Finally, I thank my teachers and friends T. Krishnan, B.V. Rao, Jim Pitman and Jenifer Brown for their help and support.

The quality of the papers in this Festschrift, to my pride, joy, and satisfaction, is very high. I looked through every paper in this volume. A large number of the papers are very original. Some open a window to a major area through a well presented review. The articles reflect the main themes Herman Rubin has contributed to over half a century. I am so thankful that the authors gave quality papers; it was magnanimous.

When, in March, 2003, I approached the IMS with a proposal for a Festschrift for Herman Rubin, two emotions were playing in my brain. I have had an affectionate relationship with Herman for about a quarter century now. But I was also mindful that the man is a rare scholar, in the finest tradition of that word. Herman could have had a few hundred more papers if he had insisted on getting credit for all he did for the rest of us, without ever asking for it. Now that's unselfish. I am honored and I am delighted that the IMS Festschrift for Herman, a fully deserved honor, even belated, is now in print. I speak for a community when I wish Herman a happy, healthy, and intellectually fulfilling long life. Herman, as Herman Chernoff said, is indeed a national treasure.

Anirban DasGupta,
Purdue University

Editor

Contributors to this volume

T. W. Anderson, Stanford University
Mary Ellen Bock, Purdue University
L. Brown, University of Pennsylvania
George Casella, University of Florida
Chingfer Chen, Stanford University
Jeesen Chen, University of Cincinnati
Ming-Yen Cheng, National Taiwan University
Herman Chernoff, Harvard University
Anirban DasGupta, Purdue University
Burgess Davis, Purdue University
Persi Diaconis, Stanford University
Morris L. Eaton, University of Minnesota
Paul Erdös
Marta Fiocco, Leiden University
L. Gawarecki, Kettering University
A. Goswami, Indian Statistical Institute
Peter Hall, Australian National University
John A. Hartigan, Yale University
Feifang Hu, University of Virginia
Bala Iyer, IBM Silicon Valley Lab
Leah Jager, University of Washington
Robert Jarrow, Cornell University
Jiashun Jin, Purdue University
J. B. Kadane, Carnegie Mellon University
Samuel Karlin, Stanford University
Timothy Keller, National Agricultural Statistics Service, USDA
Wei-Liem Loh, National University of Singapore
V. Mandrekar, Michigan State University
Eric Marchand, University of New Brunswick
John Marden, University of Illinois at Urbana-Champaign
Ingram Olkin, Stanford University
Jim Pitman, University of California, Berkeley
Dimitris Politis, University of California, San Diego
Philip Protter, Cornell University
B. L. S. Prakasa Rao, Indian Statistical Institute
B. V. Rao, Indian Statistical Institute
C. R. Rao, Pennsylvania State University
Christian P. Robert, Université Paris 9 – Dauphine
Herman Rubin, Purdue University
Mark J. Schervish, Carnegie Mellon University
Teddy Seidenfeld, Carnegie Mellon University
Thomas M. Sellke, Purdue University
Jayaram Sethuraman, Florida State University
Sunder Sethuraman, Iowa State University
D. N. Shanbhag, Pennsylvania State University

L. Shepp, Rutgers University
William E. Strawderman, Rutgers University
Wen-Chi Tsai, National Cheng-Chi University
Willem R. van Zwet, University of Leiden
Richard A. Vitale, University of Connecticut
Jeffrey Scott Vitter, Purdue University
Min Wang, Watson Research Center, IBM
Jon A. Wellner, University of Washington
Martin T. Wells, Cornell University
Michael Woodroofe, University of Michigan
Marc Yor, Université Pierre et Marie Curie – Boîte
S. Zacks, SUNY, Binghamton
James V. Zidek, University of British Columbia

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) viii–xv
© Institute of Mathematical Statistics, 2004

Publications of Herman Rubin

On the distribution of the serial correlation coefficient. *Ann. Math. Statist.* **16** (1945) 211–215.

A constant-utility index of the cost of living. *Rev. Econ. Studies* **XV** (1947–48) 84–87. With L. R. Klein.

Estimation of the parameters of a single equation in a complete system of stochastic equation. *Ann. Math. Statist.* **20** (1949) 46–63. With T. W. Anderson.

The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.* **21** (1950) 370–382. With T. W. Anderson.

Measuring the equation systems of dynamic economics. In *Cowles Commission Monograph* **10** (T. C. Koopmans, ed.) 53–237, Wiley, New York, 1950. With T. C. Koopmans and R. B. Leipnik.

Consistency of maximum-likelihood estimates in the explosive case. In *Cowles Commission Monograph* **10** (T. C. Koopmans, ed.) 356–364, Wiley, New York, 1950.

Note on random coefficients. In *Cowles Commission Monograph* **10** (T. C. Koopmans, ed.) 419–422, Wiley, New York, 1950.

A Bayes approach to a quality control model. *Ann. Math. Statist.* **23** (1952) 114–125. With M. A. Girshick.

Postulates for generalizations of Hilbert space. *Proc. Amer. Math. Soc.* **4** (1953) 611–616. With M. H. Stone.

Asymptotic properties of limited information estimates under generalized conditions. In *Cowles Commission Monograph* **14** (Wm. C. Hood and T. C. Koopmans, eds.) 200–212, Yale Univ. Press, New Haven, 1953. With H. Chernoff.

On distribution-free statistics. *Ann. Math. Statist.* **25** (1954) 593–598. With Z. W. Birnbaum.

Transformations of relativistic particle mechanics. *Pacific J. Math.* **4** (1954) 563–601. With P. C. Suppes.

On two-place predicates and fitting sequences. *J. Symb. Logic* **20** (1955) 121–122. With P. C. Suppes.

The maximum value of a Fourier–Stieltjes transform. *Math. Scand.* **3** (1955) 97–102. With E. Hewitt.

Estimates of bounded relative error in particle counting. *Ann. Math. Statist.* **26** (1955) 276–285. With M. A. Girshick and R. Sitgreaves.

Limit-preserving embeddings of partially ordered sets in directed sets. *Proc. Amer. Math. Soc.* **7** (1956) 812–813. With J. R. Isbell.

The estimation of the location of a discontinuity in density. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, Vol. 1 (J. Neyman, ed.) 19–37, Univ. of Calif. Press, Berkeley, 1956. With H. Chernoff.

Statistical inference in factor analysis. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, Vol. 5 (J. Neyman, ed.) 111–150, Univ. of Calif. Press, Berkeley, 1956. With T. W. Anderson.

Uniform convergence of random functions with applications to statistics. *Ann. Math. Statist.* **27** (1956) 201–204.

The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Statist.* **27** (1956) 272–300. With S. Karlin.

Distributions possessing a monotone likelihood ratio. *J. Amer. Statist. Assoc.* **51** (1956) 637–643. With S. Karlin.

Estimation of regression line with both variables subject to error under an unusual identification condition. *Ann. Math. Statist.* **29** (1958) 606–608.

A note on convexity in Euclidean n-space. *Proc. Amer. Math. Soc.* **9** (1958) 522–523. With O. Wesler.

Generating functions and the semigroup theory of branching Markov process. *Proc. Natl. Acad. Sci.* **44** (1958) 1057–1060. With A. T. Bharucha-Reid.

On the foundations of quantum mechanics. In *Axiomatic Method with Special Reference to Geometry and Physics.* Proc. Internat. Symp. at Univ. of Calif., Berkeley (L. Henkin, P. Suppes and A. Tarski, eds.) 333–340, North Holland, Amsterdam, 1959.

A new form of the generalized continuum hypothesis. *Bull. Amer. Math. Soc.* **65** (1959) 282–283.

Estimating the parameters of a differential process. *Ann. Math. Statist.* **30** (1959) 641–658. With H. G. Tucker.

The estimation of discontinuities in multivariate densities and related problems in stochastic processes. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, Vol. I (J. Neyman, ed.) 563–574, Univ. of Calif. Press, Berkeley, 1960.

Partnership games with secret signals prohibited. *Michigan Math. J.* **8** (1961) 71–76. With Martin Fox.

Characterization of the Wishart distribution. *Ann. Math. Statist.* **33** (1962) 1272–1280. With Ingram Olkin.

Likelihood ratios of differential processes. *Ann. Math. Statist.* **34** (1963) 1115 (Abstract).

Elementary characterization of the geometric and exponential distribution. *Amer. Math. Monthly* **71**(7) (1964) 828 (Abstract).

Admissibility of quantile estimates of a single location parameter. *Ann. Math. Statist.* **35** (1964) 1019–1030. With Martin Fox.

On a characterization of the Poisson distribution. *Sankhyā A* **26** (1964) 295–298. With C. R. Rao.

Multivariate beta distributions and independence properties of the Wishart distribution. *Ann. Math. Statist.* **35** (1964) 261–269. With Ingram Olkin.

Supports of convolutions of identical distributions. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab. II*, Part I (J. Neyman, ed.) 415–422, Univ. of Calif. Press, Berkeley, 1965/66.

An elementary concentration inequality. *Sankhyā A* **27** (1965) 415–416.

Probabilities of moderate deviations. *Sankhyā A* **27** (1965) 325–346. With J. Sethuraman.

Bayes risk efficiency. *Sankhyā A* **27** (1965) 347–356. With J. Sethuraman.

An elementary treatment of the amount of information in an experiment. *Sankhyā A* **28** (1966) 97–98.

Some properties of quantile estimates. *Z. Wahrsch. Verw. Gebiete* **4** (1966) 352–353. With Martin Fox.

The 'secretary' problem. *Ann. Math. Statist.* **37**(2) (1966) 544 (Abstract).

The Halmos–Savage theorem in the non-dominated case. *Ann. Math. Statist.* **37**(5) (1966) 1431 (Abstract).

A property of the log-likelihood-ratio process for Gaussian processes. *Ann. Inst. Stat. Math.* **20** (1968) 311–314. With B. L. S. Prakasa Rao.

A theorem on n-tuples which is equivalent to the well-ordering theorem. *Notre Dame J. Formal Logic* **8** (1967) 48–50. Corregendum. *Notre Dame J. Formal Logic* **11** (1970) 220. With J. E. Rubin.

Domains of attraction for the subcritical Galton–Watson branching process. *J. Appl. Probab.* **5** (1968) 216–219. With D. Vere-Jones.

The effect of auto-regressive dependence on a nonparametric test. *IEEE Trans. Inform. Theory* **IT-13** (1967) 311–313. With S. S. Wolff and J. L. Gastwirth.

On robust linear estimators. *Ann. Math. Statist.* **40** (1969) 24–39. With J. L. Gastwirth.

Representation of nonnegative continuous functions on product spaces. *Bull. Amer. Math. Soc.* **74** (1968) 332–333.

Functions of processes with Markovian states. *Ann. Math. Statist.* **39** (1968) 938–946. With Martin Fox.

Functions of processes with Markovian states II. *Ann. Math. Statist.* **40** (1969) 865–869. With Martin Fox.

A new approach to the foundations of probability. In *Foundations of Mathematics*. Symposium commemorating Kurt Gödel, Columbus, Ohio, 1966 (J. Bulloff, T. Holyoke and S. Hahn, eds.) 46–50, Springer, New York, 1969.

Functions of processes with Markovian states III. *Ann. Math. Statist.* **41** (1970) 472–479. With Martin Fox.

Decision theoretic approach to some multivariate problems. In *Multivariate Analysis II*. Proc. Second Internat. Symp., Dayton, Ohio, 1968 (P. R. Krishnaiah, ed.) 507–513, Academic Press, New York, 1969.

Decision-theoretic evaluation of some nonparametric methods. In *Nonparametric Techniques in Statistical Inference*. Proc. Symp. Indiana University, 1969 (M. L. Puri, ed.) 579–583, Cambridge Univ. Press, London, 1970.

A characterization based on the absolute difference of two i.i.d. random variables. *Ann. Math. Statist.* **41** (1970) 2113–2122. With P. S. Puri.

Effect of dependence on the level of some one-sample tests. *J. Amer. Statist. Assoc.* **66** (1971) 816–820. With J. L. Gastwirth.

On extreme stable laws and some applications. *J. Appl. Probab.* **8** (1971) 794–801. With M. L. Eaton and C. Morris.

Occam's Razor needs new blades. In *Foundations of Statistical Inference*. Proc. Symposium, Waterloo, Ont., 1970 (V. P. Godambe and D. A. Sprott, eds.) 372–374, Holt, Rinehart and Winston of Canada, Ltd., Toronto, Montreal, 1971.

Optimization problems in simulation. In *Optimizing Methods in Statistics*. Proc. Symposium, Columbus, Ohio, 1971 (J. S. Rustagi, ed.) 29–32, Academic Press, New York, 1971.

On large sample properties of certain non-parametric procedures. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, Vol. I (L. Le Cam, J. Neyman and E. Scott, eds.) 429–435, Univ. of Calif. Press, Berkeley, 1972.

A decision-theoretic approach to the problem of testing a null hypothesis. In *Statistical Decision Theory and Related Topics*. Proc. Symp. W. Lafayette, Indiana, 1970 (S. S. Gupta and J. Yackel, eds.) 103–108, Academic Press, New York and London, 1971.

The relationship between two weak forms of the axiom of choice. *Fund. Math.* **80** (1973) 75–79. With P. Howard and J. E. Rubin.

Probabilistic proof of the normal convergence theorem. *Ann. Probab.* **1** (1973) 867–869. With D. Root.

On a characterization of the family of distributions with constant multivariate failure rates. *Ann. Probab.* **2** (1974) 738–740. With P. S. Puri.

Special case of the distribution of the median. *Ann. Statist.* **3** (1975) 251–256. With S. R. Paranjape.

Asymptotic distribution theory of the empiric c.d.f. for mixing processes. *Ann. Statist.* **3** (1975) 809–824. With J. L. Gastwirth.

The behavior of robust estimators on dependent data. *Ann. Statist.* **3** (1975) 1070–1100. With J. L. Gastwirth.

Some non-standard problems of inference in stochastic processes. In *Proceedings of the Summer Research Institute on Statistical Inference for Stochastic Processes*. Bloomington, In., 1974 (M. L. Puri, ed.) 203–209, Academic Press, New York, 1975.

The relationship between convergence of random variables and convergence of distribution on arbitrary topological spaces. *Notices of AMS* **20** A-536, 1973.

Estimate of ultimate recoverable oil and gas in the United States. (1975). Report prepared for the Federal Energy Administration included in their interim report to Congress. With G. P. McCabe.

A method for computing random multinomial probabilities from random Poisson probabilities. *Bull. Inst. Math. Statist.* **4** (1975) 168 (Abstract).

On selecting a subset containing the best population – a Bayesian approach. *Ann. Statist.* **5** (1977) 969–983. With P. K. Goel.

Robust Bayesian estimation. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 351–356, Academic Press, New York, 1977.

The approximation of symmetric X-ray peaks by Pearson Type VII distributions. *J. Appl. Cryst.* **10** (1977) 66–68. With M. M. Hall, Jr., V. G. Veeraraghavan and P. G. Winchell.

The finite memory secretary problem. *Ann. Probab.* **5**(4) (1977) 627–635. With S. M. Samuels.

Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8** (1980) 165–170. With R. Vitale.

Computation of penalty-maximum-likelihood density estimators. *J. Statist. Comput. Simulat.* **10** (1979) 65–78. With J. K. Ghorai.

Asymptotic theory of estimation in nonlinear stochastic differential equations. *Sankhyā A* **43** (1981) 170–189. With B. L. S. Prakasa Rao.

Infinite precision methods of generating random variables with specified distributions. *Notices AMS* (1977) A-249.

Computational complexity of the generation of non-uniform random variables. *Notices AMS* (1978) A-12.

Asymptotic regular almost compactness and convergence of nets of probability measures. *Notices AMS* (1978) A-152.

Estimating a possibly rational mean. In *Statistical Decision Theory and Related Topics III*, Vol. 2 (S. S. Gupta and J. O. Berger, eds.) 263–268, Academic Press, New York, 1982.

Bayes risk consistency of nonparametric Bayes density estimators. *Austral. J. Statist.* **1** (1982) 51–66. With J. K. Ghorai.

Convergence rates of large deviations for point estimators. *Statist. Probab. Lett.* **1** (1983) 197–202. With A. Rukhin.

A note on the behavior of sample statistics when the population mean is infinite. *Ann. Probab.* **12** (1984) 256–261. With Jeesen Chen.

Drawing a random sample from a density selected at random. *Comp. Statist. & Data Anal.* **4** (1986) 219–227. With Jeesen Chen.

On the distributions of sums of symmetric random variables and vectors. *Ann. Probab.* **14** (1986) 247–259. With Thomas Sellke.

Bounds for the difference between median and mean for Gamma and Poisson distributions. *Statist. Probab. Lett.* **4** (1986) 281–283. With Jeesen Chen.

How non-uniform can a uniform sample be? *Probab. Theory and Related Fields* **73** (1986) 245–254. With Jeesen Chen and Burgess Davis.

Another look at Poisson processes. *Sankhyā A* **49** (1987) 133–137. With W.-J. Huang and P. S. Puri.

Bayesian estimation subject to minimaxity of a multivariate normal mean in the case of a common unknown variance. In *Statistical Decision Theory and Related Topics IV*, Vol. 1 (S. S. Gupta and J. O. Berger, eds.) 326–348, Springer-Verlag, New York, 1988. With A. DasGupta.

A weak system of axioms for "rational" behavior and the non-separability of utility from prior. *Statistics and Decisions* **5** (1987) 47–58.

Robustness in generalized ridge regression and related topics. In *Third Valencia Symposium on Bayesian Statistics* **3** (J. Bernardo, M. DeGroot, D. Lindley and A. F. M. Smith, eds.) 403–410, Clarendon Press, Oxford, 1988.

Some results on robustness in testing. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) 371–78, Springer-Verlag, New York, 1988.

Some stochastic processes related to density functions. *J. Theoret. Probab.* **1** (1988) 227–237. With J. Chen.

Exact computation of the asymptotic efficiency of maximum likelihood estimators of a discontinuous signal in a Gaussian white noise. *Ann. Statist.* **23**(3) (1995) 732–739. With K. S. Song.

Independence and determination of probabilities. *Proc. Amer. Math. Soc.* **125**(12) (1997) 3721–3723. With Z. Chen and R. A. Vitale.

Versions of normality and some weak forms of the axiom of choice. *MLQ Math. Log. Q.* **44**(3) (1998) 367–382. With P. Howard, K. Keremedis and J. E. Rubin.

Disjoint unions of topological spaces and choice. *MLQ Math. Log. Q.* **44**(4) (1998) 493–508. With P. Howard, K. Keremedis and J. E. Rubin.

Correlation in a Bayesian framework. *Canadian J. Statist.* **28**(4) (2000) 675–687. With A. DasGupta, G. Casella, M. Delampady, C. Genest and W. Strawderman.

A new approach to default priors. *Canadian J. Statist.* **29**(3) (2001) 437–450. With G. Casella, A. DasGupta, M. Delampady and W. Strawderman.

Bayes estimates as expanders in one and two dimensions. *J. Statist. Plann. Inference* **113**(1) (2003) 1–14. With A. DasGupta.

Estimation of the Binomial parameters when both parameters are unknown. *Special Issue of Journal of Statistical Planning and Inference in Honor of Herman Chernoff* (in Press). With A. DasGupta.

A contemporary review of infinitely divisible distributions and processes. *Sankhyā A* **64**, Part 3(ii) (2002) 763–819. With A. Bose and A. DasGupta.

Books

Equivalents of the Axiom of Choice. North Holland, Amsterdam, 1963. With J. E. Rubin.

Equivalents of the Axiom of Choice II. North Holland, Amsterdam, 1985. With J. E. Rubin.

Technical Reports

Limit theorems for transformed variables. *MSU Research Memorandum* **85** 1961.

Construction of random variables with specified distributions. *MSU Research Memorandum* **88** 1961.

Maximum likelihood estimates of matrices of finite rank. *MSU Research Memorandum* **119** 1964.

Calculation of upper tail percentiles for the Chi-square distribution. Technical Report No. 102, Stanford University, 1964. With J. Zidek.

Approximation to the distribution function of sums of independent Chi random variables. Technical Report No. 106, Stanford University, 1965. With J. Zidek.

A waiting time distribution arising from the coupon collector's problem. Technical Report No. 107, Stanford University, 1965. With J. Zidek.

Distribution of a sum of waiting times in coupon collection. Technical Report No. 109, Stanford University, 1965. With G. Chase.

First passage times and other analysis of continuous parameter Markov chains. Technical Report No. 111, Stanford University, 1965. With I. MacNeill.

A stochastic model for the evolution of cloud cover-estimation of parameters and goodness of fit based on Boston data. *MSU Research Memorandum* **132** 1965. With Martin Fox.

An iterative process of high order. *MSU Research Memorandum* **170** 1966.

Some fast methods of generating random variables with preassigned distributions. I. General acceptance–rejection procedures. Technical Report #466, 1976, Purdue University, Department of Statistics.

The effect of dead time on the physical generation of random digits. Technical Report #467, 1976, Purdue University, Department of Statistics.

A new 'definition' of the integral. Technical Report #468, 1976, Purdue University, Department of Statistics.

Mellin transforms from Fourier Transforms. Technical Report #81-14, Purdue University, Department of Statistics.

Drawing a random sample from a density selected at random. *I.M.S. Bulletin* **12** (1983) 214. With Jeesen Chen.

Fallacies of classical statistics. Technical Report #86-31, Purdue University, Department of Statistics.

An efficient method of generating infinite-precision exponential random variables. Technical Report #86-39, Purdue University, Department of Statistics.

The effect of non-normalization on the risks of the density estimators. Technical Report #86-50, Purdue University, Department of Statistics. With Jeesen Chen.

A new approach to integration. Technical Report #86-56, Purdue University, Department of Statistics.

Regular extensions of measures. Technical Report #86-57, Purdue University, Department of Statistics.

Generating non-uniform random variables: Infinite precision procedures and computational complexity. Technical Report #86-58, Purdue University, Department of Statistics.

A purely probabilistic method for finding deterministic sums. Technical Report #95-06 (1995). With A. DasGupta.

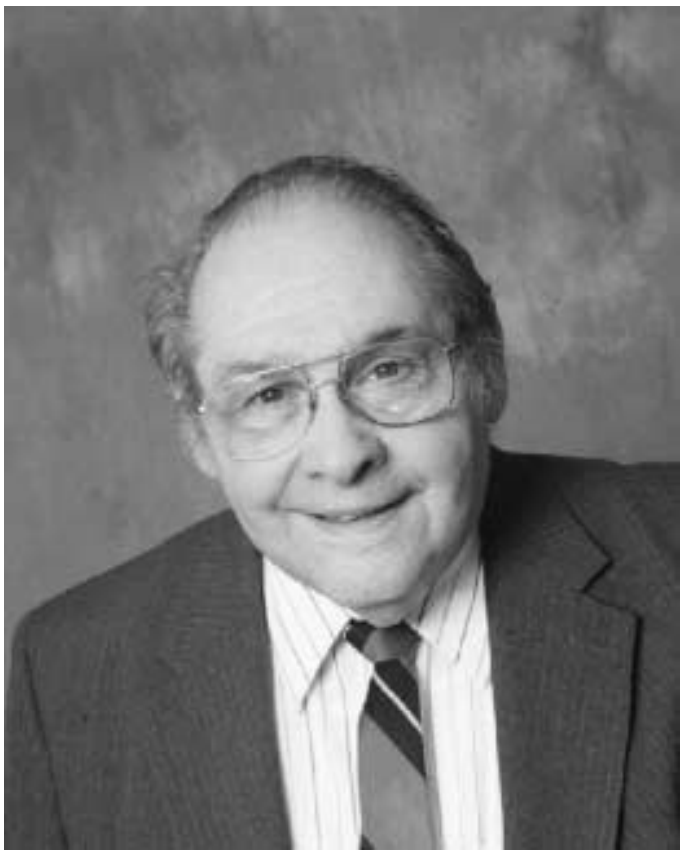
Ph.D. Students of Herman Rubin

Stanford University
Om Aggarwal
Charles Boll
Craig Magwire
Lincoln Moses
A. R. Roy
Oscar Wesler

University of Oregon
Jack Borsting
R. P. Pakshirajan

Michigan State University
Thurman Brown
Donald Fisk
Gerald Funk
John McCloskey
B. L. S. Prakasa Rao

Purdue University
Jeesen Chen
Philip Cohen
John Dirkse
J. K. Ghorai
S. R. Paranjape



Herman Rubin

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 1–4
© Institute of Mathematical Statistics, 2004

Some reminiscences of my friendship with Herman Rubin

Herman Chernoff[1]

Harvard University

I first met Herman Rubin in 1947 while I was writing a dissertation in absentia at Columbia University and he was a Fellow at the Institute of Advanced Study at Princeton. I had recently been appointed as a research assistant at the Cowles Commission for Research in Economics which was then located at the University of Chicago. Herman had completed or almost completed his dissertation at the University of Chicago, and we were to be colleagues at the Cowles Commission from June 1948 to September 1949.

While I was at Columbia, I was supposed to investigate the possibility of inverting large matrices by computer, because the method used by the Cowles Commission for estimating the parameters of the Economy, by maximizing a function using a gradient method, involved the inversion of matrices. I worked at the Watson Laboratories which were located then near Columbia and had use of a "Relay Calculator" which could be programmed (with plug boards) to multiply matrices. With the use of the Relay calculator and a card sorter and lots of fancy footwork, it was possible to do the job. At that time the engineers at Watson were beginning to build the electronic computer which was to become one of the bases for the future development of the IBM computers to follow. But I did not have access to that machine. However I did have access to Herman Rubin who came around to kibbitz, and to do some of the fancy footwork. At one point the sorter decided to put the cards with the digit 4 into the box for the digit 7. We counterattacked by instructing the 7 to go into the reject box. That scheme worked for a while, but the sorter replied by putting the 3 into the reject box. I think that we ended up doing some of the card sorting by hand.

At Cowles we had adjacent offices which was not exactly a blessing because Herman had a bad habit. He would come in to the office about 7 AM, pound his calculator (electric and not electronic) for an hour and then prove a few theorems for an hour, and then was ready to discuss important matters with me when I came to work. These important matters were usually how to handle certain bridge hands. Whatever I suggested was usually wrong. That did not bother me as much as the time I had to spend on bridge, a game that I never properly mastered.

I had a few friends in the Mathematics Department at the University. One of them, who had become a long term fixture, related to me how he had thought he was very smart (IQ about 180) when he was an undergraduate, until this little high school kid showed up, and obviously was more capable than most of the graduate students. Needless to say that that enfant terrible was our Herman Rubin.

While we were at Cowles we coauthored a paper, the main object of which was to show that even when not all of the standard conditions were satisfied, large sample theory demonstrated that we could still have confidence in our conclusions. I must admit that my contributions to this effort were only to translate

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA. e-mail: chernoff@stat.harvard.edu

Herman's work into comprhensible English, and to insist on the admittedly improper use of the word "blithely" to indicate that we could proceed as though all was well.

On another occasion L.J. Savage announced that he had resolved Wald's dilemma in using the Minimax principle, by claiming that what Wald had really meant was "Minimax Regret". In illustrating this principle in a class I was teaching, I discovered that not only could Minimax lead to poor choices, but Minimax Regret violated the principle of "Independence of Irrelevant Alternatives", a principle that had recently been enunciated in Arrow's outstanding result on the Welfare Function. When I confronted Savage with this problem, he first denied that it was serious, but after some discussion, indicated that maybe we should follow up on recent work by De Finetti proposing the Bayesian approach.

In fact I laid out the axioms that I felt should be satisfied by an objective method of rational decision making. The current terminology is "coherent". My results were sort of negative and later published in Econometrica after I let them simmer for a few years. The only thing that almost made sense, was that if we neglected one of the axioms, then the rational way to proceed is to treat each unknown state of nature as equally likely. This was an unsatisfactory result for those hoping for an objective way to do inductive inference. In the meantime both Savage and Rubin pursued the Bayesian approach. Savage later became the high priest of the Bayesian revolution. But no one seemed to notice, that two days after the discussion with Savage, Rubin wrote a discussion paper deriving the Bayesian solution. What was special about this paper, was that by omitting unnecessary verbiage, it was about three pages long, and was, unlike most of Herman's writing, eminently readable. Unfortunately, a copy of this paper which I treasured for many years, disappeared in my files, and as far as I know, no copy of it exists today.

I recall going to a seminar in the Mathematics Department, where I confess that I did not understand the lecture. At the end someone asked the speaker whether his results could be generalized to another case. The speaker said that he had thought about it, but was not clear about how to proceed. Herman spoke up, indicating that it was perfectly clear, and explained exactly how the generalization would go. This was one of many examples where it was apparent that Herman could instantly absorb results that were presented to him, and even see further nontrivial consequences of these results. I envied this clarity of thought because my own thinking processes tend to be much more confused and usually some time is needed for me to get things straightened out.

In 1949, Rubin and Arrow left the Cowles Commission to go to Stanford. Rubin joined the new Statistics Department organized by Albert Bowker with the help of Abraham Girshick. Arrow was joint in Statistics and Economics. I went to the University of Illinois, and was invited to visit Stanford for a semester two years later. I found the department to be an exciting place to be, partly because of the distinguished visitors which included David Blackwell and partly because of the presence of ONR funding for applied and theoretical programs. Herman was teaching courses in measure theory and topology, because the Mathematics Department was busy with other topics and he felt that Statistics students should at least have those basics.

While I was there, Girshick once was teasing Herman about the fact that the news indicated that an African American had just received his Ph.D. at age 18, and Herman had not gotten his degree until he was 19. Herman, taking this teasing seriously, complained that he had spent a year in the Army.

That semester, two topics that arose from the ONR project gave rise to two papers that I wrote and of which I was very proud. They pointed to a direction in optimal experimental design on which I spent much time later. Part of one of these papers involved finding asymptotic upper and lower bounds on the probability that the mean of a sample of independent identically distributed random variables would exceed a certain constant. This paper represented the first application of large deviation theory to a statistical problem. Cramer had derived a much more elegant result in 1938, of which I had been ignorant. My result, involving the infimum of a moment generating function, was less elegant and less general than the Cramer result, but did not require a special condition that Cramer required. Also, my proof could be described as crudely beating the problem to death. Herman claimed that he could get a lower bound much easier. I challenged him, and he produced a short Chebyshev Inequality type proof, which was so trivial that I did not trouble to cite his contribution.

What a mistake! It seems that Shannon had incorrectly applied the Central Limit theorem to the far tails of the distribution in one of his papers on Information theory. When his error was pointed out, he discovered the lower bound of Rubin in my paper and rescued his results. As a result I have gained great fame in electrical engineering circles for the Chernoff bound which was really due to Herman. One consequence of the simplicity of the proof was that no one ever bothered to read the original paper of which I was very proud. For years they referred to Rubin's bound as the Chernov bound, not even spelling my name correctly. I once had the pleasure of writing to a friend who sent me a copy of a paper improving on the Chernov bound, that I was happy that my name was not associated with such a crummy bound. For many years, electrical engineers have come to me and told me that I saved their lives, because they were able to describe the bound on their preliminary doctoral exams. Fortunately for me, my lasting fame, if any, will depend, not on Rubin's bound, but on Chernoff faces.

As I was preparing to return to the University of Illinois to finish off my year in 1952, my wife and I had a long discussion with Herman in which he mentioned that he had certain requirements for marriage. Evidently his proposed wife would have to be a paragon of virtues, beautiful, brilliant, and Jewish. When we returned to Stanford five months later, Herman had discovered this paragon and she was willing and they were already married.

For a few years after I came to Stanford, Rubin and I had neighboring offices at Sequoia Hall. Frequently when I came across a problem that seemed to be one that must have been treated in the literature, I would approach Herman and ask him about it. It was not unusual for him to say that it was not yet in the literature, but that he had already solved that problem. He would then reach into the depths of the mountain of paper on his desk, and pull out his solution. Often I would come to him with a problem on which I was working, and suggest an approach that I might use. His invariable response was "That is the worst way to attack that problem." This response frightened off many students and colleagues, but I found that if I persisted in asking why it was the worst way, he would sometimes explain why and sometimes admit that maybe it was a sensible approach. It required a certain amount of stubbornness, which not everyone had, to confront Herman. But I found that, because Herman was my neighbor, I was often saved from following false trails, often shown what was known, and often encouraged to pursue profitable directions that seemed problematic.

The Japanese have a title of National Treasure which they assign to outstanding artists and scholars. In my opinion, Herman Rubin, the eternal enfant terrible

of Statistics, has served as an American National Treasure by his willingness to counsel those not too frightened to hear "That is the worst way". As I recently became an octogenarian, I realize that Herman is no longer the 20 year old I once knew, but I have no doubt that he is still intellectually a slightly matured 20 year old who has contributed mightily to Statistics and from whom we can expect more.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 5–20
© Institute of Mathematical Statistics, 2004

Evaluating improper priors and the recurrence of symmetric Markov chains: An overview

Morris L. Eaton[1]

University of Minnesota

Abstract: Given a parametric statistical model, an improper prior distribution can often be used to induce a proper posterior distribution (an inference). This inference can then be used to solve decision problems once an action space and loss have been specified. One way to evaluate the inference is to ask for which estimation problems does the above formal Bayes method produce admissible estimators. The relationship of this problem to the recurrence of an associated symmetric Markov chain is reviewed.

Appreciation

Near the end of my graduate study at Stanford, Carl Morris and I had a conversation which lead us to ask whether or not the usual χ^2-test for a point null hypothesis in a multinomial setting was in fact a proper Bayes test. After a few months of struggle, we eventually reduced the problem to one involving La Place transforms. At this point it was clear we needed help, and even clearer whose assistance we should seek – namely Herman Rubin. Herman's stature as a researcher, problem solver and font of mathematical knowledge was well known to the Stanford students.

Within a few days of having the problem described to him, Herman had sketched an elegant solution minus a few "obvious" details that Carl and I were able to supply in the next month or so. This eventually led to an Eaton–Morris–Rubin publication in the *Journal of Applied Probability*. During this collaboration, I was struck with Herman's willingness to share his considerable gifts with two fledgling researchers. In the succeeding years it has become clear to me that this is an essential part of his many contributions to our discipline. Thank you Herman.

1. Introduction

This expository paper is concerned primarily with some techniques for trying to evaluate the formal Bayes method of solving decision problems. Given a parametric model and an improper prior distribution, the method has two basic steps:

1. Compute the formal posterior distribution (proper) for the parameter given the data (assuming this exists)

2. Use the formal posterior to solve the "no data" version of the decision problem.

This two step process produces a decision rule whose properties, both desirable and undesirable, can be used in the assessment of the posterior distribution and hence

[1]School of Statistics, University of Minnesota, 224 Church Street S. E., Minneapolis, MN 55455, USA. e-mail: eaton@stat.umn.edu

Keywords and phrases: formal Bayes rules, admissibility, Markov chains.
AMS 2000 subject classifications: 62A01, 62C15, 62F15.

the improper prior. Typically, when frequentist measures of assessment are proposed, they often include some discussion of admissibility (or almost admissibility) for the formal Bayes rules obtained from the posterior. However, there is a delicate balance that arises immediately. If only a few decision problems are considered in the assessment, then the evidence may not be very convincing that the posterior is suitable since admissibility is, by itself, a rather weak optimality property. On the other hand, even in simple situations with appealing improper prior distributions, it is certainly possible that there are interesting decision problems where formal Bayes solutions are inadmissible (for example, see Blackwell (1951), Eaton (1992, Example 7.1), and Smith (1994)).

One approach to the above problem that has yielded some interesting and useful results is based on estimation problems with quadratic loss. In this case, formal Bayes decision rules are just the posterior means of the functions to be estimated and risk functions are expected mean squared error. Conditions for admissibility, obtained from the Blyth–Stein method (see Blyth (1951) and Stein (1955)), involve what is often called the integrated risk difference (IRD). In the case of quadratic loss estimation, various techniques such as integration by parts or non-obvious applications of the Cauchy–Schwarz inequality applied to the IRD, sometimes yield expressions appropriate for establishing admissibility (for example, see Karlin (1958), Stein (1959), Zidek (1970), and Brown and Hwang (1982)). These might be described as "direct analytic techniques."

In the past thirty years or so, two rather different connections have been discovered that relate quadratic loss estimation problems to certain types of "recurrence problems." The first of these appeared in Brown (1971) who applied the Blyth–Stein method to the problem of establishing the admissibility of an estimator of the mean vector of a p-dimensional normal distribution with covariance equal to the identity matrix. The loss function under consideration was the usual sum of squared errors. In attempting to verify the Blyth–Stein condition for a given estimator δ, Brown showed that there corresponds a "natural" diffusion process, although this connection is far from obvious. However, the heuristics in Section 1 of Brown's paper provide a great deal of insight into the argument. A basic result in Brown (1971) is that the estimator δ is admissible iff the associated diffusion is recurrent. This result depends on some regularity conditions on the risk function of δ, but holds in full generality when the risk function of δ is bounded. The arguments in Brown's paper depend to some extent on the underlying multivariate normal sampling model. Srinivasan (1981) contains material related to Brown (1971). The basic approach in Brown has been extended to the Poisson case in Johnstone (1984, 1986) where the diffusion is replaced by a birth and death process. A common feature of the normal and Poisson problems is that the associated continuous time stochastic process whose recurrence implies admissibility, are defined on the sample space (as opposed to the parameter space) of the estimation problem. In addition the inference problems under consideration are the estimation of the "natural" parameters of the model. Brown (1979) describes some general methods for establishing admissibility of estimators. These methods are based on the ideas underlying the admissibility–recurrence connection described above.

Formal Bayes methods are the focus of this paper. Since the posterior distribution is the basic inferential object in Bayesian analysis, it seems rather natural that evaluative criteria will involve this distribution in both proper and improper prior contexts. As in Brown (1971), just why "recurrence problems" arise in this context is far from clear. Briefly, the connection results from using admissibility in quadratic loss estimation problems to assess the suitability of the posterior distri-

bution. In particular, if the posterior distribution of θ given the data x is $Q(d\theta|x)$ (depending, of course, on a model and an improper prior), then the formal Bayes estimator of any bounded function of θ, say $\phi(\theta)$, is the posterior mean of $\phi(\theta)$, say

$$\hat{\phi}(x) = \int \phi(\theta) Q(d\theta|x).$$

It was argued in Eaton (1982, 1992) that the "admissibility" of $\hat{\phi}$ for all bounded ϕ constituted plausible evidence that the formal posterior might be suitable for making inferences about θ. To connect the admissibility of $\hat{\phi}$ to recurrence, first observe that when $\phi_A(\theta) = I_A(\theta)$ is an indicator function of a subset A of the parameter space, then the formal Bayes estimator

$$\hat{\phi}_A(x) = Q(A|x)$$

is the posterior probability of A. If η denotes the "true value of the model parameter" from which X was sampled, then the expected value (under the model) of the estimator $Q(A|X)$ is

$$R(A|\eta) = \mathcal{E}_\eta Q(A|X). \tag{1.1}$$

Next, observe that R in (1.1) is a transition function defined on the parameter space Θ of the problem. Thus, R induces a discrete time Markov chain whose state space is Θ. The remainder of this paper is devoted to a discussion of the following result.

Theorem 1.1. *If the Markov chain on Θ defined by R in* (1.1) *is "recurrent," then $\hat{\phi}$ is "admissible" for each bounded measurable ϕ when the loss is quadratic.*

Because Θ is allowed to be rather general, the recurrence of the Markov chain has to be defined rather carefully – this is the reason for the quotes on recurrent. As in Brown (1971), what connects the decision theoretic aspects of the problem to the Markov chain is the Blyth–Stein technique – and this yields what is often called "almost admissibility." Thus, the quotes on admissibility.

The main goal of this paper is to explain why Theorem 1.1 is correct by examining the argument used to prove the result. The starting point of the argument is that the Blyth–Stein condition that involves the IRD provides a sufficient condition for admissibility. Because this condition is somewhat hard to verify directly, it is often the case that a simpler condition is provided via an application of the Cauchy–Schwarz Inequality. In the development here, this path leads rather naturally to a mathematical object called a Dirichlet form. Now, the connection between the resulting Dirichlet form, the associated chain with the transition function R in Theorem 1.1, and the recurrence of the chain is fairly easy to explain.

In brief, this paper is organized as follows. In Section 2, the Blyth–Stein condition is described and the basic inequality that leads to the associated Dirichlet form is presented. In Section 3 the background material (mainly from the Appendix in Eaton (1992)) that relates the Markov chain to the Dirichlet form is described. The conclusion of Theorem 1.1 is immediate once the connections above are established.

The application of Theorem 1.1 in particular examples is typically not easy – primarily because establishing the recurrence of a non-trivial Markov chain is not easy. Examples related to the Pitman estimator of a translation parameter are discussed in Section 4. The fact that the Chung-Fuchs (1951) Theorem is used here supports the contention that interesting examples are not routine applications of general theory. Also in Section 4, a recent result of Lai (1996) concerning the multivariate normal translation model is described.

A detailed proof of Theorem 3.2 is given in an appendix to this paper. The conclusion of Theorem 3.2 is hinted at in Eaton (1992), but a rigorous proof is rather more involved than I originally believed it would be. Thus the careful proof here.

Although the Markov chain of interest here has the parameter space as its state space, some interesting related work of Hobert and Robert (1999) use a related chain on the sample space in some examples where the two spaces are both subsets of the real line.

2. The Blyth–Stein condition

Here are some basic assumptions that are to hold throughout this paper. The sample space $\mathcal{X}$ and the parameter space Θ are both Polish spaces with their respective σ-algebras of Borel sets. All functions are assumed to be appropriately measurable. The statistical model for $X \in \mathcal{X}$ is $\{P(\cdot|\theta)|\theta \in \Theta\}$ and the improper prior distribution ν is assumed to be σ-finite on the Borel sets of Θ. The *marginal measure* M on $\mathcal{X}$ is defined by

$$M(B) = \int_{\Theta} P(B|\theta)\nu(d\theta). \tag{2.1}$$

Because $\nu(\Theta) = +\infty$ it is clear that $M(\mathcal{X}) = +\infty$. However, in some interesting examples, the measure M is not σ-finite and this prevents the existence of a formal posterior distribution [For example, look at $\mathcal{X} = \{0, 1, \cdots, m\}$, the model is Binomial (m, θ) and $\nu(d\theta) = [\theta(1-\theta)]^{-1}d\theta$ on (0,1). No formal posterior exists here]. In all that follows the measure M is assumed to be σ-finite. In this case, there exists a proper conditional distribution $Q(d\theta|x)$ for θ given $X = x$ which satisfies

$$P(dx|\theta)\nu(d\theta) = Q(d\theta|x)M(dx). \tag{2.2}$$

Equation (2.2) means that the two joint measures on $\mathcal{X} \times \Theta$ agree. Further, $Q(\cdot|x)$ is unique almost everywhere M. For more discussion of this, see Johnson (1991).

Given the formal posterior, $Q(\cdot|x)$, the formal Bayes estimator for any bounded function $\phi(\theta)$ when the loss is quadratic is the posterior mean

$$\hat{\phi}(x) = \int \phi(\theta)Q(d\theta|x). \tag{2.3}$$

The risk function of this estimator is

$$R(\hat{\phi}, \theta) = \mathcal{E}_\theta\big[\hat{\phi}(X) - \phi(\theta)\big]^2 \tag{2.4}$$

where $\mathcal{E}_\theta$ denotes expectation with respect to the model. Because ϕ is bounded, $\hat{\phi}$ exists and $R(\hat{\phi}, \theta)$ is a bounded function of θ. The bounded assumption on ϕ simplifies the discussion enormously and allows one to focus on the essentials of the admissibility-recurrence connection. For a version of this material that is general enough to handle the estimation of unbounded ϕ's, see Eaton (2001).

The appropriate notion of "admissibility" for our discussion here is captured in the following definition due to C. Stein.

Definition 2.1. The estimator $\hat{\phi}$ is *almost-ν-admissible* if for any other estimator $t(X)$ that satisfies

$$R(t, \theta) \le R(\hat{\phi}, \theta) \quad \text{for all } \theta, \tag{2.5}$$

the set

$$B = \big\{\theta|R(t, \theta) < R(\hat{\phi}, \theta)\big\} \tag{2.6}$$

has ν-measure zero.

Definition 2.2. The formal posterior $Q(\cdot|x)$ is *strongly admissible* if the estimator $\hat{\phi}$ is almost-ν-admissible for every bounded function ϕ.

The notion of strong admissibility is intended to capture a robustness property of the formal Bayes method across problems – at least for quadratic loss estimation problems when ϕ is bounded. The soft argument is that $Q(\cdot|x)$ cannot be too badly behaved if $\hat{\phi}$ is almost-ν-admissible for all bounded ϕ.

To describe a convenient version of the Blyth–Stein conditions for almost-ν-admissibility, consider a bounded function $g \geq 0$ defined on Θ and satisfying

$$0 < \int g(\theta)\nu(d\theta) \equiv c < +\infty. \tag{2.7}$$

Now $\nu_g(d\theta) = g(\theta)\nu(d\theta)$ is a finite measure on Θ so we can write

$$P(dx|\theta)\nu_g(d\theta) = \tilde{Q}_g(d\theta|x)M_g(dx) \tag{2.8}$$

where M_g is the marginal measure defined by

$$M_g(dx) = \int P(dx|\theta)\nu_g(d\theta). \tag{2.9}$$

Of course, $\tilde{Q}_g(d\theta|x)$ is a version of the conditional distribution of θ given $X = x$ when the proper prior distribution of θ is $c^{-1}\nu_g$. Setting

$$\hat{g}(x) = \int g(\theta)Q(d\theta|x), \tag{2.10}$$

it is not hard to show that

$$M_g(dx) = \hat{g}(x)M(dx). \tag{2.11}$$

Since the set $\{x|\hat{g}(x) = 0\}$ has M_g-measure zero, it follows that a version of $\tilde{Q}_g(d\theta|x)$ is

$$Q_g(d\theta|x) = \begin{cases} \dfrac{g(\theta)}{\hat{g}(x)}Q(d\theta|x), & \text{if} \quad \hat{g}(x) > 0, \\ Q(d\theta|x), & \text{if} \quad \hat{g}(x) = 0. \end{cases} \tag{2.12}$$

In all that follows, (2.12) is used as the conditional distribution of θ given $X = x$ when the prior distribution is ν_g.

Now, the Bayes estimator for $\phi(\theta)$, given the posterior (2.12), is

$$\hat{\phi}_g(x) = \int \phi(\theta)Q_g(d\theta|x) \tag{2.13}$$

whose risk function is

$$R(\hat{\phi}_g, \theta) = \mathcal{E}_\theta\big[\hat{\phi}_g(X) - \phi(\theta)\big]^2. \tag{2.14}$$

The so called *Integrated Risk Difference*,

$$IRD(g) = \int \big[R(\hat{\phi}, \theta) - R(\hat{\phi}_g, \theta)\big]g(\theta)\nu(d\theta) \tag{2.15}$$

plays a key role in the Blyth–Stein condition for the almost-ν-admissibility of $\hat{\phi}$. To describe this condition, consider a measurable set $C \subseteq \Theta$ with $0 < \nu(C) < +\infty$ and let

$$U(C) = \left\{ g \geq 0 \;\middle|\; \begin{array}{l} g \text{ is bounded}, \quad g(\theta) \geq 1 \quad \text{for } \theta \in C, \\ \text{and } \displaystyle\int g(\theta)\nu(d\theta) < +\infty \end{array} \right\} \tag{2.16}$$

.

Theorem 2.1 (Blyth–Stein). *Assume*

$$\begin{cases} \text{There is a sequence of sets } C_i \subseteq C_{i+1} \subseteq \Theta, \quad i = 1, \cdots \quad \text{with} \\ 0 < \nu(C_i) < +\infty \quad \text{and} \quad C_i \nearrow \Theta \quad \text{so that} \\ \inf_{g \in U(C_i)} IRD(g) = 0 \quad \text{for } i = 1, 2, \cdots \end{cases} \tag{2.17}$$

Then $\hat{\phi}$ is almost-ν-admissible.

The proof of this well known result is not repeated here. The usual interpretation of Theorem 2.1 is that when $\hat{\phi}$ is "close enough to a proper Bayes rule $\hat{\phi}_g$" then $\hat{\phi}$ is almost-ν-admissible, but the notion of closeness is at best rather vague.

A possible first step in trying to apply Theorem 2.1 is to find a tractable (and fairly sharp) upper bound for $IRD(g)$ in (2.15). Here is the key inequality that allows one to see eventually why "recurrence" implies strong-admissibility.

Theorem 2.2. *For a real valued measurable function h defined on Θ, let*

$$\Delta(h) = \iiint (h(\theta) - h(\eta))^2 Q(d\theta|x) Q(d\eta|x) M(dx). \tag{2.18}$$

Then for each bounded function ϕ, there is constant K_ϕ so that

$$IRD(g) \le K_\phi \Delta\left(\sqrt{g}\right), \tag{2.19}$$

for all bounded non-negative g satisfying $\int g(\theta)\nu(d\theta) < +\infty$.

Proof. A direct proof of (2.19) using the Cauchy–Schwarz Inequality follows. First, let $A = \{x | \hat{g}(x) > 0\}$ and recall that A^c has M_g measure zero. Thus,

$$\begin{aligned} IRD(g) &= \int_{\mathcal{X}} \int_\Theta \left[\left(\hat{\phi}(x) - \phi(\theta)\right)^2 - \left(\hat{\phi}_g(x) - \phi(\theta)\right)^2 \right] P(dx|\theta) g(\theta) \nu(d\theta) \\ &= \int_{\mathcal{X}} \int_\Theta \left[\left(\hat{\phi}(x) - \phi(\theta)\right)^2 - \left(\hat{\phi}_g(x) - \phi(\theta)\right)^2 \right] Q_g(d\theta|x) M_g(dx) \\ &= \int_A \left(\hat{\phi}(x) - \hat{\phi}_g(x)\right)^2 \hat{g}(x) M(dx) \\ &= \int_A \left[\int_\Theta \phi(\theta) \left(1 - \frac{g(\theta)}{\hat{g}(x)}\right) Q(d\theta|x) \right]^2 \hat{g}(x) M(dx) \\ &= \int_A \frac{1}{\hat{g}(x)} \left[\int_\Theta \phi(\theta) \left(g(\theta) - \hat{g}(x)\right) Q(d\theta|x) \right]^2 M(dx). \end{aligned} \tag{2.20}$$

A bit of algebra shows that for each x,

$$\begin{aligned} &\int_\Theta \phi(\theta) \left(g(\theta) - \hat{g}(x)\right) Q(d\theta|x) \\ &\quad = \frac{1}{2} \iint \left(\phi(\theta) - \phi(\eta)\right) \left(g(\theta) - g(\eta)\right) Q(d\theta|x) Q(d\eta|x). \end{aligned}$$

Using the non-negativity of g and the Cauchy–Schwarz inequality we have

$$\begin{aligned} &\left| \iint \left(\phi(\theta) - \phi(\eta)\right) \left(g(\theta) - g(\eta)\right) Q(d\theta|x) Q(d\eta|x) \right| \\ &\quad \le W(x) \cdot \left[\iint \left(\sqrt{g(\theta)} - \sqrt{g(\eta)}\right)^2 Q(d\theta|x) Q(d\eta|x) \right]^{\frac{1}{2}} \end{aligned}$$

where

$$W^2(x) = \iint (\phi(\theta) - \phi(\eta))^2 \left(\sqrt{g(\theta)} + \sqrt{g(\eta)}\right)^2 Q(d\theta|x)Q(d\eta|x).$$

Since ϕ is bounded, say $|\phi(\theta)| \leq c_0$, and since $(\sqrt{g(\theta)} + \sqrt{g(\eta)})^2 \leq 2(g(\theta) + g(\eta))$, we have

$$W^2(x) \leq 4c_0^2 \hat{g}(x).$$

Substituting these bounds into the final expression in (2.20) yields

$$\begin{aligned} IRD(g) &\leq 4c_0^2 \int_A \iint \left(\sqrt{g(\theta)} - \sqrt{g(\eta)}\right)^2 Q(d\theta|x)Q(d\eta|x)M(dx) \\ &\leq 4c_0^2 \Delta\left(\sqrt{g}\right). \end{aligned}$$

Setting $K_\phi = 4c_0^2$ yields the result. □

Combining Theorem 2.1 and Theorem 2.2 gives the main result of this section.

Theorem 2.3. *Assume*

$$\begin{cases} \text{There is a sequence of increasing sets } C_i \subseteq \Theta, \quad i = 1, 2, \ldots \\ \text{with } 0 < \nu(C_i) < +\infty \quad \text{and} \quad C_i \nearrow \Theta \quad \text{so that} \\ \inf\limits_{g \in U(C_i)} \Delta\left(\sqrt{g}\right) = 0, \text{ for each } i. \end{cases} \tag{2.21}$$

Then $Q(d\theta|x)$ is strongly admissible.

Proof. When (2.21) holds, inequality (2.19) shows that (2.17) holds for each bounded measurable ϕ. Then $Q(d\theta|x)$ is strongly admissible. □

It should be noted that the assumption (2.21) does not involve ϕ (as opposed to assumption (2.17)). Thus the conditions for strong admissibility involve the behavior of Δ. It is exactly the functional Δ that provides the connection between (2.21) and the "recurrence" of the Markov chain with transition function R in (1.1).

To put the material of the next section in perspective, it is now useful to isolate some of the essential features of the decision theory problem described above – namely, under what conditions on the given model $P(dx|\theta)$ and the improper prior $\nu(d\theta)$ with the formal posterior $Q(d\theta|x)$ be strongly admissible? A basic ingredient in our discussion will be the transition function

$$R(d\theta|\eta) = \int Q(d\theta|x)P(dx|\eta) \tag{2.22}$$

introduced in Section 1. A fundamental property of R is its symmetry with respect to ν – that is, the measure on $\Theta \times \Theta$ defined by

$$s(d\theta, d\eta) = R(d\theta|\eta)\nu(d\eta). \tag{2.23}$$

is a symmetric measure in the sense that

$$\begin{aligned} s(A \times B) &= \iint I_A(\theta) I_B(\eta) R(d\theta|\eta)\nu(d\eta) \\ &= s(B \times A) \end{aligned} \tag{2.24}$$

for Borel subsets A and B of Θ. This is easily established from the definition of R. It is this symmetry that drives the theory of the next section and allows us to connect the behavior of Δ, namely

$$\Delta(h) = \iint (h(\theta) - h(\eta))^2 \, R(d\theta|\eta)\nu(d\eta), \tag{2.25}$$

to the "recurrence" of the Markov chain defined by R. The expression (2.25) for Δ follows from (2.18) and the disintegration formula (2.2).

Also, note that ν is a stationary measure for R – that is,

$$\int R(A|\eta)\nu(d\eta) = \nu(A) \tag{2.26}$$

for all Borel sets A. This is an easy consequence of the symmetry of s in (2.23).

The discussion in the next section begins with an abstraction of the above observations. Much of the discussion is based on the Appendix in Eaton (1992).

Here is the standard Pitman example that gives a concrete non-trivial example of what the above formulation yields.

Example 2.1. Consider $X_1, \ldots, X_n$ that are independent and identically distributed random vectors in R^p with a density $f(x - \theta)$ (with respect to Lebesgue measure). Thus $\Theta = R^p$ and the model for $X = (X_1, \ldots, X_n)$ is

$$P(dx|\theta) = \prod_{i=1}^{n} f(x_i - \theta)dx_i$$

on the sample space $\mathcal{X} = R^{pn}$. With dx as Lebesgue measure on $\mathcal{X}$, the density of $P(dx|\theta)$ with respect to dx is

$$p(x|\theta) = \prod_{i=1}^{n} f(x_i - \theta).$$

Next take $\nu(d\theta) = d\theta$ on $\Theta = R^p$ and assume, for simplicity, that

$$m(x) = \int_{R^p} p(x|\theta)d\theta$$

is in $(0, \infty)$ for all x. Then a version of "$Q(d\theta|x)$" is

$$Q(d\theta|x) = \frac{p(x|\theta)}{m(x)} \, d\theta.$$

Thus the transition function R is given by

$$R(d\theta|\eta) = \left(\int_{\mathcal{X}} \frac{p(x|\theta)p(x|\eta)}{m(x)} \, dx \right) d\theta.$$

Therefore,

$$R(d\theta|\eta) = r(\theta|\eta)d\theta$$

where the density $r(\cdot|\eta)$ is

$$r(\theta|\eta) = \int_{\mathcal{X}} \frac{p(x|\theta) \; p(x|\eta)}{m(x)} \, dx.$$

Now, it is easy to show that for each vector $u \in R^p$,

$$r(\theta + u|\eta + u) = r(\theta|\eta)$$

so that r is only a function of $\theta - \eta$, say

$$t(\theta - \eta) = r(\theta - \eta|0).$$

Further routine calculations give

$$\begin{cases} t(u) = t(-u) & \text{for } u \in R^p \\ \int t(u)du = 1 \end{cases}$$

In summary then, for the translation model with $d\theta$ as the improper prior distribution, the induced transition function is

$$R(d\theta|\eta) = t(\theta - \eta)d\theta$$

and t is a symmetric density function on R^p. We will return to this example later.

3. Symmetric Markov chains

Here, a brief sketch of symmetric Markov chain theory, recurrence and Dirichlet forms is given. The purpose of this section is two-fold – first to explain the relationship between recurrence and the Dirichlet form and second to relate this to the strong admissibility result of Theorem 2.3.

Let Y be a Polish space with the Borel σ-algebra $\mathcal{B}$ and consider a Markov Kernel $R(dy|z)$ on $\mathcal{B} \times Y$. Also let λ be a non-zero σ-finite measure on $\mathcal{B}$.

Definition 3.1. The kernel $R(dy|z)$ is λ-*symmetric* if the measure

$$\alpha(dy, dz) = R(dy|z)\lambda(dz) \tag{3.1}$$

is a symmetric measure on $\mathcal{B} \times \mathcal{B}$.

Typically, R is called symmetric without reference to λ since λ is fixed in most discussions. As the construction in Section 2 shows, interesting examples of symmetric kernels abound in statistical decision theory. In all that follows, it is assumed that R is λ-symmetric. Note that the assumption of σ-finiteness for λ is important.

Given a λ-symmetric R, consider a real valued measurable function h and let

$$\Delta(h) = \iint (h(y) - h(z))^2 R(dy|z)\lambda(dz). \tag{3.2}$$

The quadratic form Δ (or sometimes $\frac{1}{2}\Delta$) is often called a Dirichlet form. Such forms are intimately connected with continuous time Markov Process Theory (see Fukushima et al (1994)) and also have played a role in some work on Markov chains (for example, see Diaconis and Strook (1991)). A routine calculation using the symmetry of R shows that

$$\Delta(h) \leq 4 \int h^2(y)\lambda(dy) \tag{3.3}$$

so Δ is finite for $h \in L_2(\lambda)$, the space of λ-square integrable functions.

Now, given $R(dy|z)$, there is a Markov chain with state space Y and transition function $R(dy|z)$. More precisely, consider the path space $\mathcal{W} = Y^\infty = Y \times Y \times \cdots$ with the usual product σ-algebra. Given the initial value w_0, there is a Markov chain $W = (w_0, W_1, W_2, \ldots)$ so that $R(dw_{i+1}|w_i)$ is the conditional distribution of W_{i+1} given $W_i = w_i$, for $i = 0, 1, 2, \ldots$. The unique probability measure on path space that is consistent with this *Markov specification*, is denoted by $S(\cdot|w_0)$.

Because the space Y is rather general, the definition of recurrence has to be selected with some care. The reader should note that neither irreducibility nor periodicity occur in the discussion that follows (see Meyn and Tweedie (1993) for a discussion of such things in the general state space case). Let $C \subseteq Y$ satisfy $0 < \lambda(C) < +\infty$. Such measurable sets are called λ-*proper*. Define the random variable T_C on W as follows:

$$T_C = \begin{cases} +\infty & \text{if } W_i \notin C \quad \text{for } i = 1, 2, \ldots \\ 1 & \text{if } W_1 \in C \\ n & \text{if } W_n \in C \quad \text{for some } n \geq 2 \quad \text{and} \\ & \quad W_i \notin C \quad \text{for } i = 1, \ldots, n-1 \end{cases} \tag{3.4}$$

Then T_C ignores the starting value of the chain and records the first hitting time of C for times greater than 0. The set

$$B_C = \{T_C < +\infty\} \tag{3.5}$$

is the event where the chain hits C at some time after time 0.

Definition 3.2. A λ-proper set $C \subseteq Y$ is called *locally -λ-recurrent* if the set

$$B_0 = \{w_0 \in C | S(B_C|w_0) < 1\}$$

has λ-measure zero.

Definition 3.3. A λ-proper set $C \subseteq Y$ is called λ-*recurrent* if the set

$$B_1 = \{w_0 | S(B_C|w_0) < 1\}$$

has λ-measure zero.

In other words, C is locally-ν-recurrent if whenever the chain starts in C, it returns to C w.p.1, except for a set of starting values of λ-measure zero. It is this notion of recurrence that is most relevant for admissibility considerations. Of course, C is λ-recurrent if the chain hits C no matter where it starts, except for a set of starting values of λ-measure zero. This second notion is closer to traditional ideas related to recurrence.

To describe the connection between the Dirichlet form Δ and local-λ-recurrence, consider

$$V(C) = \left\{ h \in L_2(\lambda) \;\middle|\; \begin{array}{ll} h \geq 0, & h(y) \geq 1 \quad \text{for } y \in C, \\ h \text{ is bounded} & \end{array} \right\}. \tag{3.6}$$

Note that $U(C)$ in (2.16) and $V(C)$ are in one-to-one correspondence via the relation $h(y) = \sqrt{g(y)},\ y \in Y$.

Theorem 3.1. *For a λ-proper set C,*

$$\inf_{h \in V(C)} \Delta(h) = 2 \int_C (1 - S(B_C|w))\lambda(dw). \tag{3.7}$$

A proof of this basic result can be found in Appendix 2 of Eaton (1992). From (3.7), it is immediately obvious that C is a locally-λ-recurrent set iff the inf over $V(C)$ of the Dirichlet form Δ is zero.

Definition 3.4. The Markov chain $W = (W_0, W_1, W_2, \ldots)$ is *locally-λ-recurrent* if each λ-proper set C is locally-λ-recurrent.

In applications, it is useful to have some conditions that imply local-λ-recurrence since the verification that every λ-proper C is locally-λ-recurrent can be onerous. To this end, we have

Theorem 3.2. *The following are equivalent:*

(i) *The Markov chain $W = (W_0, W_1, W_2, \ldots)$ is locally-λ-recurrent*

(ii) *There exists an increasing sequence of λ-proper sets C_i, $i = 1, \ldots$ such that $C_i \longrightarrow Y$ and each C_i is locally-λ-recurrent*

Proof. Obviously (i) implies (ii). The converse is proved in the appendix. □

In a variety of decision theory problems, it is often sufficient to find *one* set B_0 that is "recurrent" in order to establish "admissibility." For an example of the "one-set phenomenon," see Brown and Hwang (1982). In the current Markov chain context, here is a "one-set" condition that implies local-λ-recurrence for the chain W.

Theorem 3.3. *Suppose there exists a λ-proper set B_0 that is λ-recurrent (see Definition* 3.3*). Then the Markov chain W is locally-ν-recurrent.*

Proof. Since λ is σ-finite, there is a sequence of increasing λ-proper sets $B_i, i = 1, 2, \ldots$ such that $B_i \longrightarrow Y$. Let $C_i = B_i \cup B_0$, $i = 1, 2, \ldots$ so the sets C_i are λ-proper, are increasing, and $C_i \longrightarrow Y$. The first claim is that each C_i is locally-λ-recurrent. To see this, let N be the λ-null set where $S(T_{B_0} < +\infty | w) < 1$. Then for $w \notin N$, the chain hits B_0 w.p.1 after time 0 when $W_0 = w$. Thus, for $w \notin N$, the chain hits $B_0 \cup B_i$ w.p.1 after time 0 when $W_0 = w$. Therefore the set $C_i = B_0 \cup B_i$ is locally-λ-recurrent. By Theorem 3.2, W is locally-λ-recurrent. □

The application of the above results to the strong-admissibility problem is straightforward. In the context of Section 2, consider a given model $P(dx|\theta)$ and a σ-finite improper prior distribution $\nu(d\theta)$ so that the marginal measure M in (2.1) is σ-finite. This allows us to define the transition $R(d\theta|\eta)$ in (2.22) that is ν-symmetric. Therefore the above theory applies to the Markov chain $W = (W_0, W_1, W_2, \ldots)$ on Θ^∞ defined by $R(d\theta|\eta)$. Here is the main result that establishes "Theorem 1.1" stated in the introductory section of this paper.

Theorem 3.4. *Suppose the Markov chain W with state space Θ and transition function $R(d\theta|\eta)$ is locally-ν-recurrent. Then the posterior distribution $Q(d\theta|x)$ defined in* (2.2) *is strongly-admissible.*

Proof. Because W is locally-ν-recurrent, the infimum in (3.7) is zero for each ν-proper set C. This implies that condition (2.21) in Theorem 2.3 holds. Thus, $Q(d\theta|x)$ is strongly admissible. □

Of course Theorem 3.2 makes it a bit easier to show W is locally-ν-recurrent, while Theorem 3.3 provides an extremely useful sufficient condition for this property of W. An application is given in the next section.

4. Examples

Here we focus on two related examples. The first is based on the Pitman model introduced in Example 2.1. In this case, the induced Markov chain is a random walk on the parameter space, and as is well known, under rather mild moment conditions for dimensions $p = 1$ and $p = 2$, the random walks are recurrent. But for $p \geq 3$, there are no recurrent random walks on R^p that have densities with respect to Lebesque measure. Of course this parallels what decision theory yields for admissibility of the Pitman estimator of a mean vector – admissibility for $p = 1$ and $p = 2$ (under mild moment conditions) and inadmissibility in many examples when $p \geq 3$. The results here do not concern estimation of a mean vector, but rather involve the strong admissibility of the posterior, and again the dimension phenomenon prevails.

The second example is from the thesis of Lai (1996) and concerns the p-dimensional normal distribution with an unknown mean vector and the identity covariance matrix. In essence, Lai's results provide information regarding a class of improper priors that yield strong admissibility when the parameter space is R^p. Even in the case of the normal distribution there are many open questions concerning the "inappropriateness" of the posterior when the improper prior is $d\theta$ on R^p, $p \geq 3$.

Example 4.1 (continued). As shown in Section 2, the induced transition function on the parameter space $\Theta = R^p$ is given by

$$R(d\theta|\eta) = t(\theta - \eta)\, d\theta \tag{4.1}$$

where the density function t is defined in Example 2.1. The Markov chain induced by R is just a random walk on R^p. When $p = 1$, the results of Chung and Fuchs (1951) apply directly. In particular, if $p = 1$ and

$$\int_{-\infty}^{\infty} |u| t(u)\, du < +\infty, \tag{4.2}$$

then the Markov chain is recurrent and so the posterior distribution in this case is strongly admissible. A sufficient condition for (4.2) to hold is that the original density function f in Example 2.1 has a finite mean (see Eaton (1992) for details).

When $p = 2$, a Chung–Fuchs-like argument also applies (see Revuz (1984)). In particular, if

$$\int_{R^2} \|u\|^2 t(u)\, du < +\infty, \tag{4.3}$$

then the Markov chain on R^2 is recurrent so strong admissibility obtains. Again, it is not too hard to show that the existence of second moments under f in Example 2.1 imply that (4.3) holds. These results for $p = 1, 2$ are suggested by the work of Stein (1959) and James and Stein (1961).

For $p \geq 3$, the Markov chain obtained from R in (4.1) can never be recurrent (see Guivarc'h, Keane, and Roynette (1977)) suggesting that the posterior distribution obtained from the improper prior $d\theta$ on $\Theta = R^p$ is suspect. At present, the question of "inadmissibility" of the posterior when $p \geq 3$ remains largely open. This ends our discussion of Example 2.1.

Example 4.2. The material in this example is based on the work of Lai (1996). Suppose X is a p-dimensional random vector with a normal distribution $N_p(\theta, I_p)$.

Here $\theta \in \Theta = R^p$ is unknown and the covariance matrix of X is the $p \times p$ identity. Consider an improper prior distribution of the form

$$\nu(d\theta) = (a + \|\theta\|^2)^\alpha \, d\theta \tag{4.4}$$

where the constant a is positive and α is a real number. In this setting Lai proved the following.

Theorem 4.1 (Lai (1996)). *If $\alpha \in (-\infty, -\frac{p}{2} + 1]$, then the posterior distribution for θ is strongly admissible.*

The above follows from the more general Theorem 5.3.3 in Lai (1996), but well illustrates the use of the Markov chain techniques. Lai's argument consists of proving that for the range of α indicated, the induced Markov chain on Θ is locally-ν-recurrent so strong admissibility obtains. In fact, the Markov chain techniques developed by Lai to handle this example include extensions of some recurrence criteria of Lamperti (1960) and an application of Theorem 3.3 given above. Although the class of priors in (4.4) is quite small, the extension of Theorem 4.1 to other improper priors can be verified via Remark 3.3 in Eaton (1992). In this particular example, Remark 3.3, coupled with Theorem 4.1, implies the following.

Theorem 4.2. *Consider a prior distribution ν of the form* (4.4) *with $\alpha \in (-\infty, -\frac{p}{2} + 1]$ and let $g(\theta)$ satisfy*

$$c \le g(\theta) \le \frac{1}{c} \quad \textit{for all } \theta$$

for some $c > 0$. Then the Markov chain induced by the prior distribution

$$\nu_g(d\theta) = g(\theta)\nu(d\theta) \tag{4.5}$$

is locally-ν-recurrent and the induced posterior distribution is strongly admissible.

For applications of Lai's ideas to the multivariate Poisson case, we refer the reader to Lai's thesis. This completes Example 4.1.

Appendix

Here we provide a proof of Theorem 3.2. To this end, consider a measurable subset $C \subseteq Y$ that is λ-proper and let

$$H(C) = \inf_{h \in V(C)} \Delta(h). \tag{A.1}$$

Also, let

$$V^*(C) = \{h | h \in V(C), \quad h(y) \in [0,1] \quad \text{for } y \in C^c\}.$$

The results in Appendix 2 of Eaton (1992) show that

$$H(C) = \inf_{h \in V^*(C)} \Delta(h). \tag{A.2}$$

Lemma 1.1. *Consider measurable subsets A and B of Y that are both λ-proper. If $A \subseteq B$, then*

$$H^{\frac{1}{2}}(A) \le H^{\frac{1}{2}}(B) \le H^{\frac{1}{2}}(A) + 2^{\frac{1}{2}} \lambda^{\frac{1}{2}}(B \cap A^c). \tag{A.3}$$

Proof. Since $V(A) \supseteq V(B)$, $H(A) \leq H(B)$ so the left hand inequality in (A.3) is obvious. For the right hand inequality, first note that $\Delta^{\frac{1}{2}}$ is a subadditive function defined on $L_2(\lambda)$– that is,

$$\Delta^{\frac{1}{2}}(h_1 + h_2) \leq \Delta^{\frac{1}{2}}(h_1) + \Delta^{\frac{1}{2}}(h_2). \tag{A.4}$$

A proof of (A.4) is given below. With $h_3 = h_1 + h_2$, (A.4) yields

$$\Delta^{\frac{1}{2}}(h_3) \leq \Delta^{\frac{1}{2}}(h_1) + \Delta^{\frac{1}{2}}(h_3 - h_1), \tag{A.5}$$

for h_1 and h_3 in $L_2(\lambda)$. Now consider $h \in V^*(A)$ and write

$$\tilde{h}(y) = h(y) + g(y)$$

where

$$g(y) = (1 - h(y))I_{B\cap A^c}(y).$$

Then $\tilde{h} \in V^*(B)$ and (A.5) implies that

$$\Delta^{\frac{1}{2}}(\tilde{h}) \leq \Delta^{\frac{1}{2}}(h) + \Delta^{\frac{1}{2}}(g).$$

Thus,

$$H^{\frac{1}{2}}(B) \leq \Delta^{\frac{1}{2}}(h) + \Delta^{\frac{1}{2}}(g). \tag{A.6}$$

Because $g(y) \in [0, 1]$,

$$\begin{aligned}
\Delta(g) &= \iint (g(y) - g(z))^2 R(dy|z)\lambda(dz) \\
&= 2\left[\int g^2(y)\lambda(dy) - \iint g(y)g(z)R(dy|z)\lambda(dz)\right] \\
&\leq 2\int_{B\cap A^c} g^2(y)\lambda(dy) \leq 2\lambda\big(B \cap A^C\big).
\end{aligned}$$

Substituting this into (A.6) yields

$$H^{\frac{1}{2}}(B) \leq \Delta^{\frac{1}{2}}(h) + 2^{\frac{1}{2}}\lambda^{\frac{1}{2}}\big(B \cap A^c\big). \tag{A.7}$$

Since (A.7) holds for any $h \in V^*(A)$, the right hand inequality in (A.3) holds. This completes the proof. □

The proof of (A.4) follows. For h_1 and h_2 in $L_2(\lambda)$, consider the symmetric bilinear form

$$< h_1, h_2 >= \int h_1(y)h_2(y)\lambda(dy) - \iint h_1(y)h_2(z)R(dy|z)\lambda(dz).$$

That $< \cdot, \cdot >$ is non-negative definite is a consequence of the symmetry of R and the Cauchy–Schwarz inequality:

$$\begin{aligned}
&\left(\iint h_1(y)h_1(z)R(dy|z)\lambda(dz)\right)^2 \\
&\quad \leq \left(\iint h_1^2(y)R(dy|z)\lambda(dz)\right) \cdot \left(\iint h_1^2(z)R(dy|z)\lambda(dz)\right) \\
&\quad = \left(\int h_1^2(y)\lambda(dy)\right)^2.
\end{aligned}$$

Thus $\|h\| =< h, h >^{\frac{1}{2}}$ is a semi-norm on $L_2(\lambda)$ and so is subadditive. Since $\Delta(h) = 2\|h\|^2$, inequality (A.4) holds.

The proof of Theorem 3.2 is now easy. Let C be any λ-proper set so $\lambda(C) < +\infty$ and let

$$E_i = C_i \cap C, \quad i = 1, 2, \ldots.$$

Since C_i is locally-λ-recurrent, $H(C_i) = 0$ so $H(E_i) = 0$ by Lemma A.1. Since $E_i \nearrow C$ and $\lambda(C) < +\infty$, we have

$$\lim_{i \longrightarrow \infty} \lambda(E_i) \longrightarrow \lambda(C)$$

and

$$\lim_{i \longrightarrow \infty} \lambda(C \cap E_i^c) \longrightarrow 0.$$

Applying (A.3) yields

$$H^{\frac{1}{2}}(C) \leq H^{\frac{1}{2}}(E_i) + 2^{\frac{1}{2}} \lambda(C \cap E_i^c).$$

The right hand side of this inequality converges to zero as $i \longrightarrow \infty$. Hence $H(C) = 0$. Since C was an arbitrary λ-proper set, the chain W is locally-ν-recurrent.

Acknowledgment

Many thanks to Jim Hobert, Tiefeng Jiang and Galin Jones for their valuable comments. Also special thanks to Anirban Das Gupta for his efforts on this Festschrift for Herman Rubin and his many comments on this contribution.

References

[1] Blackwell, D. (1951). On the translation parameter problem for discrete variables. *Ann. Math. Statist.* **22**, 105–110. MR43418

[2] Blyth, C. R. (1951). On minimax statistical procedures and their admissibility. *Ann. Math. Statist.* **22**, 22–42. MR39966

[3] Brown, L. (1971). Admissible estimators, recurrent diffusions, and insolvable boundary value problems. *Ann. Math. Statist.* **42**, 855–904. MR286209

[4] Brown, L. (1979). A heuristic method for determining admissibility of estimators – with applications. *Ann. Statist.* **7**, 960–994. MR536501

[5] Brown, L. D. and Hwang, J. T. (1982). A unified admissibility proof. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1**, 205–230. New York: Academic Press. MR705290

[6] Chung, K. L. and Fuchs, W. H. (1951). On the distribution of values of sums of random variables. *Mem. Amer. Math. Soc.* **6**, 1–12. MR40610

[7] Diaconis, P. and Strook, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Am. Appl. Probab.* **1**, 36–61. MR1097463

[8] Eaton, M. L. (1982). A method for evaluating improper prior distributions. In *Statistics Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1**, 329–352. New York:Academic Press.

[9] Eaton, M. L. (1992). A statistical diptych: Admissible inferences- Recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147–1179. MR1186245

[10] Eaton, M. L. (2001). Markov chain conditions for admissibility with quadratic loss. In *State of the Art in Statistics and Probability*, a Festschrift for Willem van Zwet edited by M. de Gunst, C. Klaassen, and A. van der Vaart, pp. 223–243. Volume **36** in the IMS Lecture Note Series, Beachwood, Ohio. MR1836563

[11] Fukushima, M., Oshima, Y., Takeda, M. (1994). *Dirichlet Forms and Symmetric Markov Processes*. New York: de Gruyter. MR1303354

[12] Guivarc'h, Y., Keane, M. and Roynette, B. (1977). *March aleatoires sur les Groupes de Lie. Lecture Notes in Math* **624**, New York: Springer. MR517359

[13] Hobert, J. and Robert, C. (1999). Eaton's Markov chain, its conjugate partner, and P-admissibility. *Ann. Statist.* **27**, 361–373. MR1701115

[14] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 361–380. Berkeley: University of California Press. MR133191

[15] Johnson, B. (1991). On the admissibility of improper Bayes inferences in fair Bayes decision problems. Ph.D. Thesis, University of Minnesota.

[16] Johnstone, I. (1984). Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Statist.* **12**, 1173–1198. MR760682

[17] Johnstone, I. (1986). Admissible estimation, Dirichlet principles, and recurrence of birth-death chains on Z_+^p. *Probab. Theory Related Fields* **71**, 231–269. MR816705

[18] Karlin, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Math. Statist.* **29**, 406–436. MR124101

[19] Lai, W. (1996). Admissibility and the recurrence of Markov chains with applications. Ph.D. Thesis, University of Minnesota.

[20] Lamperti, J. (1960). Criteria for the recurrence or transience of stochastic processes I. *Journal of Math. Anal. Appl.* **1**, 314–330. MR126872

[21] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. New York: Springer-Verlag. MR1287609

[22] Revuz, D. (1984). *Markov Chains, 2nd Edition*. North Holland, Amsterdam. MR415773

[23] Smith, J. Q. (1994). The inadmissibility of certain Bayes decision rules under vague priors in location problems. *J. R. S. S. - B.* **56**, 543–548. MR1278226

[24] Srinivasan, C. (1981). Admissible generalized Bayes estimators and exterior boundary value problems. *Sankhya Ser. A.* **43**, 1–25. MR656265

[25] Stein, C. (1955). A necessary and sufficient condition for admissibility. *Ann. Math. Statist.* **76**, 518–522. MR70929

[26] Stein, C. (1959). The admissibility of Pitman's estimator of a single location parameter. *Ann. Math. Statist.,* **30**, 970–979. MR109392

[27] Zidek, J. V. (1970). Sufficient conditions for admissibility under squared error loss of formal Bayes estimators. *Ann. Math. Statist.* **41**, 446–456. MR261742

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 21–44
© Institute of Mathematical Statistics, 2004

Estimation in restricted parameter spaces: A review

Eric Marchand*[1] and William E. Strawderman[2]

University of New Brunswick

Rutgers University

Abstract: In this review of estimation problems in restricted parameter spaces, we focus through a series of illustrations on a number of methods that have proven to be successful. These methods relate to the decision-theoretic aspects of admissibility and minimaxity, as well as to the determination of dominating estimators of inadmissible procedures obtained for instance from the criteria of unbiasedness, maximum likelihood, or minimum risk equivariance. Finally, we accompany the presentation of these methods with various related historical developments.

1. Introduction

Herman Rubin has contributed in deep and original ways to statistical theory and philosophy. He has selflessly shared his keen intuition into and extensive knowledge of mathematics and statistics with many of the researchers represented in this volume. The statistical community has been vastly enriched by his contributions through his own research and through his influence, direct and indirect, on the research and thinking of others. We are pleased to join in this celebration in honor of Professor Rubin.

This review paper is concerned with estimation of a parameter or vector parameter θ, when θ is restricted to lie in some (proper) subset of the "usual" parameter space. The approach is decision theoretic. Hence, we will not be concerned with hypothesis testing problems, or with algorithmic problems of calculating maximum likelihood estimators. Excellent and extensive sources of information on these aspects of restricted inference are given by Robertson, Wright and Dykstra (1988), Akkerboom (1990), and Barlow, Bartholomew, Bremner and Brunk (1972). We will not focus either on the important topic of interval estimation. Along with the recent review paper by Mandelkern (2002), here is a selection of interesting work concerning methods for confidence intervals, for either interval bounded, lower bounded, or order restricted parameters: Zeytinoglu and Mintz (1984, 1988), Stark (1992), Hwang and Peddada (1994), Drees (1999), Kamboreva and Mintz (1999), Iliopoulos and Kourouklis (2000), and Zhang and Woodroofe (2003).

We will focus mostly on point estimation and we will particularly emphasize finding estimators which dominate classical estimators such as the Maximum Likelihood or UMVU estimator in the unrestricted problem. Issues of minimaxity and admissibility will also naturally arise and be of interest. Suppose, for example, that the problem is a location parameter problem and that the restricted (and of course

*Research supported by NSERC of Canada.

[1]Department of Mathematics and Statistics, University of New Brunswick, P.O. Box 4400, Fredericton, N.B. Canada, E3B 5A3. e-mail: marchand@math.unb.ca

[2]Department of Statistics, Rutgers University, 501 Hill Center, Busch Campus, Piscataway, NJ 08855, USA. e-mail: straw@rutgers.edu

Keywords and phrases: restricted parameter spaces, minimaxity, admissibility, dominating procedures, maximum likelihood estimation.

AMS 2000 subject classifications: 62-02, 62C99, 62C15, 62C20, 62C10, 62F10, 62F30.

the original space) is non-compact. In this case it often happens that these classical estimators are minimax in both the original problem and the restricted problem. If the restriction is to a convex subset, projection of the classical procedure onto the space will typically produce an improved minimax procedure, but the resulting procedure will usually not be admissible because of violation of technical smoothness requirements. In these cases there is a natural interest in finding minimax generalized Bayes estimators. The original result in this setting is that of Katz (1961) who showed (among other things) for the normal location problem with the mean restricted to be non-negative, that the generalized Bayes estimator with respect to the uniform prior (under quadratic loss) is minimax and admissible and dominates the usual (unrestricted ML or UMVU) estimator. Much of what follows has Katz's result as an examplar. A great deal of the material in sections 2, 3 and 5 is focussed on extending aspects of Katz's result.

If, in the above normal example, the restricted space is a compact interval, then the projection of the usual estimator still dominates the unrestricted MLE but cannot be minimax for quadratic loss because it is not Bayes. In this case Casella and Strawderman (1981) and Zinzius (1981) showed that the unique minimax estimator of the mean θ for a restriction of the form $\theta \in [-m, m]$ is the Bayes estimator corresponding to a 2 point prior on $\{-m, m\}$ for m sufficiently small. The material in section 6 deals with this result, and the large related literature that has followed.

In many problems, as in the previous paragraph, Bayes or Generalized Bayes estimators are known to form a complete class. When loss is quadratic and the prior (and hence typically the posterior) distribution is not degenerate at a point, the Bayes estimator cannot take values on the boundary of the parameter space. There are many results in the literature that use this phenomenon to determine inadmissibility of certain estimators that take values on (or near) the boundary. Moors (1985) developed a useful technique which has been employed by a number of authors in proving inadmissibility and finding improved estimators. We investigate this technique and the related literature in section 4.

An interesting and important issue to which we will not devote much effort is the amount of (relative or absolute) improvement in risk obtained by using procedures which take the restrictions on the parameter space into account. In certain situations the improvement is substantial. For example, if we know in a normal problem that the variance of the sample mean is 1 and that the population mean θ is positive, then risk, at $\theta = 0$, of the (restricted) MLE is 0.5, so there is a 50% savings in risk (at $\theta = 0$). Interestingly, at $\theta = 0$, the risk of the Bayes estimator corresponding to the uniform prior on $[0, \infty)$ is equal to the risk of the MLE so there is no savings in risk at $\theta = 0$. There is, however, noticeable improvement some distance from $\theta = 0$. An interesting open problem is to find admissible minimax estimators in this setting which do not have the same risk at $\theta = 0$ as the unrestricted MLE, and, in particular, to find an admissible minimax estimator dominating the restricted MLE.

We will concern ourselves primarily with methods that have proven to be successful in such problems, and somewhat less so with cataloguing the vast collection of results that have appeared in the literature. In particular, we will concentrate on the following methods.

In Section 2, we describe a recent result of Hartigan (2003). He shows, if $X \sim N_p(\theta, I_p)$, loss is $L(\theta, d) = \|d - \theta\|^2$, and $\theta \in C$, where C is any convex set (with non empty interior), then the Bayes estimator with respect to the uniform prior distribution on C dominates the (unrestricted MRE, UMVU, unrestricted ML) estimator $\delta_0(X) = X$. Hartigan's result adds a great deal to what was already

known and provides a clever new technique for demonstrating domination.

In Section 3, we study the Integral Expression of Risk Difference (IERD) method introduced by Kubokawa (1994a). The method is quite general as regards to loss function and underlying distribution. It has proven useful in unrestricted as well as restricted parameter spaces. In particular, one of its first uses was to produce an estimator dominating the James-Stein estimator of a multivariate normal mean under squared error loss.

In Section 4, following a discussion on questions of admissibility concerning estimators that take values on the boundary of a restricted parameter space, we investigate a technique of Moors (1985) which is useful in constructing improvements to such estimators under squared error loss.

Section 5 deals with estimating parameters in the presence of additional information. For example, suppose X_1 and X_2 are multivariate normal variates with unknown means θ_1 and θ_2, and known covariance matrices $\sigma_1^2 I$ and $\sigma_2^2 I$. We wish to estimate θ_1 with squared error loss $\|\delta - \theta_1\|^2$ when we know for example that $\theta_1 - \theta_2 \in A$ for some set A. We illustrate the application of a rotation technique, used perhaps first by Blumenthal and Cohen (1968a), as well as Cohen and Sackrowitz (1970), which, loosely described, permits to subdivide the estimation problem into parts that can be separately handled.

Section 6 deals with minimaxity, and particularly those results related to Casella and Strawderman (1981) and Zinzius (1981) establishing minimaxity of Bayes estimators relative to 2 point priors on the boundary of a sufficiently small one dimensional parameter space of the form $[a, b]$.

2. Hartigan's result

Let $X \sim N_p(\theta, I_p)$, $\theta \in C$ where C is an arbitrary convex set in $\Re^p$ with an open interior. For estimating θ under squared error loss, Hartigan (2003) recently proved the striking result that the (Generalized) Bayes estimator relative to the uniform prior distribution on C dominates the usual (unrestricted) MRE estimator X. It seems quite fitting to begin our review of methods useful in restricted parameter spaces by discussing this, the newest of available techniques. Below, ∇ and ∇^2 will denote respectively the gradient and Laplacian operators.

Theorem 1 (Hartigan, 2003). *For $X \sim N_p(\theta, I_p)$, $\theta \in C$ with C being a convex subset of $\Re^p$ with a non-empty interior, the Bayes estimator $\delta_U(X)$ with respect to a uniform prior on C dominates $\delta_0(X) = X$ under squared error loss $\|d - \theta\|^2$.*

Proof. Writing

$$\delta_U(X) = X + \frac{\nabla_X m(X)}{m(X)} \quad \text{with} \quad m(X) = (2\pi)^{-p/2} \int_C e^{-\frac{1}{2}|X-\nu|^2}\, d\nu,$$

we have following Stein (1981),

$$\begin{aligned}
&R\big(\theta, \delta_U(X)\big) - R\big(\theta, \delta_0(X)\big) \\
&\quad = E_\theta \left[\|X + \frac{\nabla_X m(X)}{m(X)} - \theta\|^2 - \|X - \theta\|^2 \right] \\
&\quad = E_\theta \left[\frac{\|\nabla_X m(X))\|^2}{m^2(X)} + 2(X-\theta)' \frac{\nabla_X m(X)}{m(X)} \right] \\
&\quad = E_\theta \left[\frac{\|\nabla_X m(X)\|^2}{m^2(X)} + \operatorname{div} \left(\frac{\nabla_X m(X)}{m(X)} \right) + \frac{(X-\theta)' \nabla_X m(X)}{m(X)} \right]
\end{aligned}$$

$$= E_\theta \left[\frac{\|\nabla_X m(X)\|^2}{m^2(X)} + \frac{m(X)\nabla_X^2 m(X) - \|\nabla_X m(X)\|^2}{m^2(X)} + \frac{(X-\theta)'\nabla_X m(X)}{m(X)} \right]$$
$$= E_\theta \left[\frac{1}{m(X)} H(X,\theta) \right],$$

where $H(x,\theta) = \nabla_x^2 m(x) + (x-\theta)'\nabla_x m(x)$.

It suffices to show tht $H(x,\theta) \le 0$ for all $x \in \Re^p$ and $\theta \in C$. Now, observe that $\nabla_x(e^{-\frac{1}{2}\|x-\nu\|^2}) = -\nabla_\nu(e^{-\frac{1}{2}\|x-\nu\|^2})$ and $\nabla_x^2(e^{-\frac{1}{2}\|x-\nu\|^2}) = \nabla_\nu^2(e^{-\frac{1}{2}\|x-\nu\|^2})$ so that

$$\begin{aligned}
(2\pi)^{p/2} H(x,\theta) &= \nabla_x^2 \int_C e^{-\frac{1}{2}\|x-\nu\|^2} d\nu + (x-\theta)'\nabla_x \int_C e^{-\frac{1}{2}\|x-\nu\|^2} d\nu \\
&= \int_C \nabla_\nu^2 \big(e^{-\frac{1}{2}\|x-\nu\|^2}\big)\, d\nu - (x-\theta)' \int_C \nabla_\nu \big(e^{-\frac{1}{2}\|x-\nu\|^2}\big)\, d\nu \\
&= \int_C \nabla_\nu' \big\{ \big(\nabla_\nu \big(e^{-\frac{1}{2}\|x-\nu\|^2}\big) - (x-\theta) e^{-\frac{1}{2}\|x-\nu\|^2}\big)\big\}\, d\nu \\
&= \int_C \mathrm{div}_\nu \big[(\theta-\nu) e^{-\frac{1}{2}\|x-\nu\|^2}\big]\, d\nu. \qquad (1)
\end{aligned}$$

By the Divergence theorem, this last expression gives us

$$(2\pi)^{p/2} H(x,\theta) = \int_{\partial C} \eta(\nu)'(\theta-\nu) e^{-\frac{1}{2}\|x-\nu\|^2}\, d\sigma(\nu), \qquad (2)$$

where $\eta(\nu)$ is the unit outward normal to C at ν on ∂C, and $d\sigma(\nu)$ is the surface area Lebesgue measure on ∂C (for $p = 1$, see Example 1). Finally, since C is convex, the angle between the directions $\eta(\nu)$ and $\theta - \nu$ for a boundary point ν is obtuse, and we thus have $\eta(\nu)'(\theta-\nu) \le 0$, for $\theta \in C$, $\nu \in \partial C$, yielding the result. □

Remark 1.

(a) If θ belongs to the interior C° of C; (as in part (a) of Example 1); notice that $\eta(v)'(\theta - v) < 0$ a.e. $\sigma(v)$, which implies $H(x,\theta) < 0$, for $\theta \in C^\circ$ and $x \in \Re^p$, and consequently $R(\theta, \delta_U(X)) < R(\theta, \delta_0(X))$ for $\theta \in C^\circ$.

(b) On the other hand, if C is a pointed cone at θ_0; (as in part (b) of Example 1); then $\eta(\nu)'(\theta_0 - \nu) = 0$ for all $\nu \in \partial C$ which implies $R(\theta_0, \delta_U(X)) = R(\theta_0, \delta_0(X))$.

As we describe below, Theorem 1 has previously been established for various specific parameter spaces C. However, Hartigan's result offers not only a unified and elegant proof, but also gives many non-trivial extensions with respect to the parameter space C. We pursue with the instructive illustration of a univariate restricted normal mean.

Example 1.

(a) For $X \sim N(\theta, 1)$ with $\theta \in C = [a,b]$, we have by (1),

$$\begin{aligned}
(2\pi)^{\frac{1}{2}} H(x,\theta) &= \int_a^b \frac{d}{d\nu}(\theta-\nu) e^{-\frac{1}{2}(x-\nu)^2}\, d\nu \\
&= \Big[(\theta-\nu)\, e^{-\frac{1}{2}(x-v)^2}\Big]_a^b \\
&= (\theta-b)\, e^{-\frac{1}{2}(x-b)^2} - (\theta-a)\, e^{-\frac{1}{2}(x-a)^2} \\
&< 0, \qquad \text{for all } \theta \in [a,b].
\end{aligned}$$

This tells us that $R(\theta, \delta_U(X)) < R(\theta, \delta_0(X))$ for all $\theta \in C = [a, b]$.

(b) For $X \sim N(\theta, 1)$ with $\theta \in C = [a, \infty)$ (or $C = (-\infty, a]$), it is easy to see that the development in part (a) remains valid with the exception that $H(x, a) = 0$ for all $x \in \Re$, which tells us that $R(\theta, \delta_U(X)) \leq R(\theta, \delta_0(X))$ for $\theta \in C$ with equality iff $\theta = a$.

The dominance result for the bounded normal mean in Example 1(a) was established by MacGibbon, Gatsonis and Strawderman (1987), in a different fashion, by means of Stein's unbiased estimate of the difference in risks, and sign change arguments following Karlin (1957). The dominance result for the lower bounded normal mean in Example 1(b) was established by Katz (1961), where he also showed that $\delta_U(X)$ is a minimax and admissible estimator of θ.[1] Notice that these results by themselves lead to extensions of the parameter spaces C where $\delta_U(X)$ dominates $\delta_0(X)$, for instance to hyperrectangles of the form $C = \{\theta \in \Re^p : \theta_i \in [a_i, b_i];\ i = 1, \ldots, p\}$, and to intersection of half-spaces since such problems can be expressed as "products" of one-dimensional problems.

Balls and cones in $\Re^p$ are two particularly interesting classes of convex sets for which Hartigan's result gives new and useful information. It is known that for balls of sufficiently small radius, (see e.g., Marchand and Perron, 2001, and Section 4.3 below), the uniform prior leads to dominating procedures (of the mle), but Hartigan's result implies that the uniform Bayes procedures always dominate $\delta_0(X) = X$. Also, for certain types of cones such as intersections of half spaces, Katz's result implies domination over X as previously mentioned. However, Hartigan's result applies to all cones, and, again, increases greatly the catalog of problems where the uniform Bayes procedure dominates X under squared error loss.

Now, Hartigan's result, as described above, although very general with respect to the choice of the convex parameter space C, is nevertheless limited to: (i) normal models, (ii) squared error loss, (iii) the uniform prior as leading to a dominating Bayes estimator; and extensions in these three directions are certainly of interest. Extensions to general univariate location families and general location invariant losses are discussed in Section 3.2. Finally, it is worth pointing out that in the context of Theorem 1, the maximum likelihood estimator $\delta_{\text{mle}}(X)$, which is the projection of X onto C, also dominates $\delta_0(X) = X$. Hence, dominating estimators of $\delta_0(X)$ can be generated by convex linear combinations of $\delta_U(X)$ and $\delta_{\text{mle}}(X)$. Thus the inadmissibility itself is obvious but the technique and the generality are very original and new.

3. Kubokawa's method

Kubokawa (1994a) introduced a powerful method, based on an integral expression of risk difference (IERD), to give a unified treatment of point and interval estimation of the powers of a scale parameter, including the particular case of the estimation of a normal variance. He also applied his method for deriving a class of estimators improving on the James-Stein estimator of a multivariate mean. As reviewed by Kubokawa (1998,1999), many other applications followed such as in: estimation of variance components, estimation of non-centrality parameters, linear calibration, estimation of the ratio of scale parameters, estimation of location and scale parameters under order restrictions, and estimation of restricted location

[1] Although the result is correct, the proof given by Katz has an error (see for instance van Eeden, 1995).

and scale parameters. As well, a particular strength resides in the flexibility of the method in handling various loss functions.

3.1. Example

Here is an illustration of Kubokawa's IERD method for an observable X generated from a location family density $f_\theta(x) = f_0(x-\theta)$, with known f_0, where $E_\theta[X] = \theta$, and $E_\theta[X^2] < \infty$. For estimating θ, with squared error loss $(d-\theta)^2$, under the constraint $\theta \geq a$ (hereafter, we will take $a = 0$ without loss of generality), we show that the Generalized Bayes estimator $\delta_U(X)$ with respect to the uniform prior $\pi(\theta) = 1_{(0,\infty)}(\theta)$ dominates the MRE estimator $\delta_0(X) = X$. As a preliminary to the following dominance result, observe that $\delta_U(X) = X + h_U(X)$, where

$$h_U(y) = \frac{\int_0^\infty (\theta - y) f_0(y-\theta)\, d\theta}{\int_0^\infty f_0(y-\theta)\, d\theta} = \frac{-\int_{-\infty}^y u f_0(u)\, du}{\int_{-\infty}^y f_0(u)\, du} = -E_0[X|X \leq y];$$

and that h_U is clearly continuous, nonincreasing, with $h_U(y) \geq -\lim_{y\to\infty} E_0[X|X \leq y] = -E_0[X] = 0$.

Theorem 2. *For the restricted parameter space $\theta \in \Theta = [0,\infty)$, and under squared error loss:*

(a) *Estimators $\delta_h(X) = \delta_0(X) + h(X)$ with absolutely continuous, non-negative, nonincreasing h, dominate $\delta_0(X) = X$ whenever $h(x) \leq h_U(x)$ (and $\delta_h \neq \delta_0$);*

(b) *The Generalized Bayes estimator $\delta_U(X)$ dominates the MRE estimator $\delta_0(X)$.*

Proof. First, part (b) follows from part (a) and the above mentioned properties of h_U. Observing that the properties of h and h_U imply $\lim_{y\to\infty} h(y) = 0$, and following Kubokawa (1994a), we have

$$\begin{aligned}(x-\theta)^2 - (x+h(x)-\theta)^2 &= (x+h(y)-\theta)^2\,|_{y=x}^\infty. \\ &= \int_x^\infty \frac{\partial}{\partial y}(x+h(y)-\theta)^2\, dy \\ &= 2\int_x^\infty h'(y)(x+h(y)-\theta)\, dy,\end{aligned}$$

so that

$$\begin{aligned}\Delta_h(\theta) &= E_\theta[(X-\theta)^2 - (X+h(X)-\theta)^2] \\ &= 2\int_{-\infty}^\infty \left\{\int_x^\infty h'(y)(x+h(y)-\theta)\, dy\right\} f_0(x-\theta)\, dx \\ &= 2\int_{-\infty}^\infty h'(y)\left\{\int_{-\infty}^y (x+h(y)-\theta) f_0(x-\theta)\, dx\right\} dy.\end{aligned}$$

Now, since $h'(y) \leq 0$ (h' exists a.e.), it suffices in order to prove that $\Delta_h(\theta) \geq 0$; $\theta \geq 0$; to show that

$$G_h(y,\theta) = \int_{-\infty}^y (x+h(y)-\theta) f_0(x-\theta)\, dx \leq 0$$

for all y, and $\theta \geq 0$. But, this is equivalent to

$$\begin{aligned}
&\frac{\int_{-\infty}^{y}(x+h(y)-\theta)f_0(x-\theta)\,dx}{\int_{-\infty}^{y}f_0(x-\theta)\,dx} \leq 0 \\
&\Leftrightarrow \frac{\int_{-\infty}^{y-\theta}(u+h(y))f_0(u)\,du}{\int_{-\infty}^{y-\theta}f_0(u)\,du} \leq 0 \\
&\Leftrightarrow h(y) \leq -E_0[X|X \leq y-\theta]; \quad \text{for all } y, \quad \text{and} \quad \theta \geq 0; \\
&\Leftrightarrow h(y) \leq \inf_{\theta \geq 0}\{-E_0[X|X \leq y-\theta]\}; \quad \text{for all } y; \\
&\Leftrightarrow h(y) \leq -E_0[X|X \leq y] = h_U(y); \quad \text{for all } y;
\end{aligned}$$

given that $E_0[X|X \leq z]$ is nondecreasing in z. This establishes part (a), and completes the proof of the theorem. □

Remark 2. In Theorem 2, it is worth pointing out, and it follows immediately that $G_h(y,\theta) \leq 0$, for all y, with equality iff $h = h_U$ and $\theta = 0$, which indicates that, for the dominating estimators of Theorem 2, $R(\theta, \delta_h(X)) \leq R(\theta, \delta_0(X))$ with equality iff $h = h_U$ and $\theta = 0$. As a consequence, $\delta_U(X)$ fails to dominate any of these other dominating estimators $\delta_h(X)$, and this includes the case of the truncation of $\delta_0(X)$ onto $[0,\infty)$, $\delta^+(X) = \max(0, \delta_0(X))$ (also see Section 4.4 for a discussion on a normal model $\delta^+(X)$).

3.2. Some related results to Theorem 2

For general location family densities $f_0(x-\theta)$, and invariant loss $L(\theta,d) = \rho(d-\theta)$ with strictly convex ρ, Farrell (1964) established: (i) part (b) of Theorem 2, and (ii) the minimaxity of $\delta_U(X)$, and (iii) the admissibility of $\delta_U(X)$ for squared error loss ρ. Using Kubokawa's method, Marchand and Strawderman (2003,a) establish extensions of Theorem 2 (and of Farrell's result (i)) to strictly bowl-shaped losses ρ. They also show, for quite general (f_0, ρ), that the constant risk of the MRE estimator $\delta_0(X)$ matches the minimax risk. This implies that dominating estimators of $\delta_0(X)$, such as those in extensions of Theorem 2, which include $\delta_U(X)$ and $\delta^+(X)$, are necessarily minimax for the restricted parameter space $\Theta = [0,\infty)$. Marchand and Strawderman (2003,a,b) give similar developments for scale families, and for cases where the restriction on θ is to an interval $[a,b]$. Related work for various models and losses includes Jozani, Nematollahi and Shafie (2002), van Eeden (2000, 1995), Parsian and Sanjari (1997), Gajek and Kaluszka (1995), Berry (1993), and Gupta and Rohatgi (1980), and many of the references contained in these papers.

Finally, as previously mentioned, Kubokawa's method has been applied to a wide range of problems, but, in particular for problems with ordered scale or location parameters (also see Remark 4), results and proofs similar to Theorem 2 have been established by Kubokawa and Saleh (1994), Kubokawa (1994b), and Ilioupoulos (2000).

4. Estimators that take values on the boundary of the parameter space

Theoretical difficulties that arise in situations when estimating procedures take values on, or close to the boundary of constrained parameter spaces are well documented. For instance, Sacks (1963), and Brown (1971), show for estimating under squared error loss a lower bounded normal mean θ with known variance, that the maximum likelihood estimator is an inadmissible estimator of θ. More recently,

difficulties such as those encountered with interval estimates have been addressed in Mandelkern (2002). In this section, we briefly expand on questions of admissibility and on searches for improved procedures, but we mostly focus on a method put forth by Moors (1985) which is useful in providing explicit improvements of estimators that take values on, or close to the boundary of a restricted parameter space.

4.1. Questions of admissibility

Here is a simple example which illustrates why, in many cases, estimators that take values on the boundary of the parameter space are inadmissible under squared error loss. Take $X \sim N_p(\theta, I_p)$ where θ is restricted to a ball $\Theta(m) = \{\theta \in \Re^p : \|\theta\| \leq m\}$. Complete class results indicate that admissible estimators are necessarily Bayes for some prior π (supported on $\Theta(m)$, or a subset of $\Theta(m)$). Now observe that prior and posterior pairs $(\pi, \pi|x)$ must be supported on the same set, and that a Bayes estimator takes values $\delta_\pi(x) = E(\theta|x)$ on the interior of the convex $\Theta(m)$, as long as $\pi|x$, and hence π, is not degenerate at some boundary point θ_0 of $\Theta(m)$. The conclusion is that non-degenerate estimators $\delta(X)$ which take values on the boundary of $\Theta(m)$ (i.e., $\mu\{x : \delta(x) \in \partial(\Theta(m)\} > 0$, with μ as Lebesgue measure); which includes for instance the MLE; are inadmissible under squared error loss. In a series of papers, Charras and van Eeden (1991a, 1991b, 1992, 1994) formalize the above argument for more general models, and also provide various results concerning the admissibility and Bayesianity under squared error loss of boundary estimators in convex parameter spaces. Useful sources of general complete class results, that apply for bounded parameter spaces, are the books of Berger (1985), and Brown (1986).

Remark 3. As an example where the prior and posterior do not always have the same support, and where the above argument does not apply, take $X \sim Bi(n, \theta)$ with $\theta \in [0, m]$. Moreover, consider the MLE which takes values on the boundary of $[0, m]$. It is well known that the MLE is admissible (under squared error loss) for $m = 1$. Interestingly, again for squared error loss, Charras and van Eeden (1991a) establish the admissibility of the MLE for cases where $m \leq 2/n$, while Funo (1991) establishes its inadmissibility for cases where $m < 1$ and $m > 2/n$. Interestingly and in contrast to squared error loss, Bayes estimators under absolute-value loss may well take values on the boundary of the parameter space. For instance, Iwasa and Moritani (1997) show, for a normal mean bounded an interval $[a, b]$ (known standard deviation), that the MLE is a proper Bayes (and admissible) estimator under absolute-value loss.

The method of Moors, described in detail in Moors (1985), and further illustrated by Moors (1981) and Moors and van Houwelingen (1987), permits the construction of improved estimators under squared error loss of invariant estimators that take values on, or too close to the boundary of closed and convex parameter spaces. We next pursue with an illustration of this method.

4.2. The method of Moors

Illustrating Moors' method, suppose an observable X is generated from a location family density $f(x - \theta)$ with known positive and symmetric f. For estimating $\theta \in \Theta = [-m, m]$ with squared error loss, consider invariant estimators (with respect to sign changes) which are of the form

$$\delta_g(X) = g(|X|)\frac{X}{|X|}.$$

The objective is to specify dominating estimators of $\delta_g(X)$, for cases where $\delta_g(X)$ takes values on or near the boundary $\{-m, m\}$ (i.e., $|m - g(x)|$ is "small" for some x).

Decompose the risk of $\delta_g(X)$ by conditioning on $|X|$ (i.e., the maximal invariant) to obtain (below, the notation $E_\theta^{|X|}$ represents the expectation with respect to $|X|$)

$$\begin{aligned} R(\theta, \delta_g(X)) &= E_\theta^{|X|}\left\{E_\theta\left[\left(g(|X|)\,\frac{X}{|X|} - \theta\right)^2 \middle| |X|\right]\right\} \\ &= E_\theta^{|X|}\left\{\theta^2 + g^2(|X|) - 2E_\theta\left[\frac{\theta X}{|X|} g(|X|)\middle| |X|\right]\right\} \\ &= E_\theta^{|X|}\left\{\theta^2 + g^2(|X|) - 2g(|X|)A_{|X|}(\theta)\right\}, \end{aligned}$$

where

$$A_{|X|}(\theta) = \theta E_\theta\left[\frac{X}{|X|}\middle| |X|\right] = \theta\left\{\frac{f(|X|-\theta) - f(|X|+\theta)}{f(|X|-\theta) + f(|X|+\theta)}\right\},$$

(as in (6) below) by symmetry of f. Now, rewrite the risk as

$$R(\theta, \delta_g(X)) = E_\theta^{|X|}\left[\theta^2 - A_{|X|}^2(\theta)\right] + E_\theta^{|X|}\left[\left(g(|X|) - A_{|X|}(\theta)\right)^2\right], \tag{3}$$

to isolate with the second term the role of g, and to reflect the fact that the performance of the estimator $\delta_g(X)$, for $\theta \in [-m, m]$, is measured by the average distance $\left(g(|X|) - A_{|X|}(\theta)\right)^2$ under $f(x-\theta)$. Continue by defining the $A_{|x|}$ as the convex hull of the set $\{A_{|x|}(\theta) : -m \le \theta \le m\}$. Coupled with the prior representation (3) of $R(\theta, \delta_g(X))$, we can now state the following result.

Theorem 3. *Suppose $\delta_g(X)$ is an estimator such that $\mu\{x : g(|x|) \notin A_{|x|}\} > 0$, then the estimator $\delta_{g_0}(X)$ with $g_0(|x|)$ being the projection of $g(|x|)$ onto $A_{|x|}$ dominates $\delta_g(X)$, with squared error loss under f, for $\theta \in \Theta = [-m, m]$.*

Example 2. (Normal Case) Consider a normal model f with variance 1. We obtain $A_{|x|}(\theta) = \theta\tanh(\theta|x|)$, and $A_{|x|} = [0, m\tanh(m|x|)]$, since $A_{|x|}(\theta)$ is increasing in $|\theta|$. Consider an estimator $\delta_g(X)$ such that $\mu\{x : g(|x|) > m\tanh(m|x|)\} > 0$. Theorem 3 tells us that $\delta_{g_0}(X)$, with $g_0(|X|) = \min(m\tanh(m|X|), g(|X|))$, dominates $\delta_g(X)$.

Here are some additional observations related to the previous example (but also applicable to the general case of this section).

(i) The set $A_{|x|} = [0, m\tanh(m|x|)]$ can be interpreted as yielding a complete class of invariant estimators with the upper envelope corresponding to the Bayes estimator $\delta_{BU}(X)$ associated with the uniform prior on $\{-m, m\}$.

(ii) In Example 2, the dominating estimator $\delta_{g_0}(X)$ of Theorem 3 will be given by the Bayes estimator $\delta_{BU}(X)$ if and only if $m\tanh(m|x|) \le g(|x|)$, for all x. In particular, if $\delta_g(X) = \delta_{\text{mle}}(X)$, with $g(|X|) = \min(|X|, m)$, it is easy to verify that $\delta_{g_0}(X) = \delta_{BU}(X)$ iff $m \le 1$.

(iii) It is easy to see that improved estimators $\delta_{g'}(X)$ of $\delta_g(X)$ can be constructed by projecting $g(|x|)$ a little bit further onto the interior of $A_{|x|}$, namely by selecting g' such that $\frac{1}{2}\left[g'(|x|) + g(|x|)\right] \le g_0(|x|)$ whenever $g(|x|) > g_0(|x|)$.

4.3. Some related work

Interestingly, the dominance result in (iii) of the normal model MLE was previously established, in a diferent manner, by Casella and Strawderman (1981) (see also Section 6). As well, other dominating estimators here were provided numerically by Kempthorne (1988).

For the multivariate version of Example 2: $X \sim N_p(\theta, I_p)$; $(p \geq 1)$; with $\|\theta\| \leq m$, Marchand and Perron (2001) give dominating estimators of $\delta_{\rm mle}(X)$ under squared error loss $\|d-\theta\|^2$. Namely, using a similar risk decomposition as above, including argument (ii), they show that $\delta_{BU}(X)$ (Bayes with respect to a boundary uniform prior) dominates $\delta_{\rm mle}(X)$ whenever $m \leq \sqrt{p}$. By pursuing with additional risk decompositions, they obtain various other dominance results. In particular, it is shown that, for sufficiently small radius m, $\delta_{\rm mle}(X)$ is dominated by all Bayesian estimators associated with orthogonally invariant priors (which includes the uniform Bayes estimator δ_U). Finally, Marchand and Perron (2003) give extensions and robustness results involving δ_{BU} to spherically symmetric models, and Perron (2003) gives a similar treatment for the model $X \sim Bi(n, \theta)$ with $|\theta - \frac{1}{2}| \leq m$.

4.4. Additional topics and the case of a positive normal mean

Other methods have proven useful in assessing the performance of boundary estimators in constrained parameter spaces, as well as providing improvements. As an example, for the model $X_i \sim \text{Bin}(n_i, \theta_i)$; $i = 1, \ldots, k$; with $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$, Sackrowitz and Strawderman (1974) investigated the admissibility (for various loss functions) of the MLE of $(\theta_1, \ldots, \theta_k)$, while Sackrowitz (1982) provided improvements (under sum of squared error losses) to the MLE in the cases above where it is inadmissible. Further examples consist of a series of papers by Shao and Strawderman (1994,1996a,1996b) where, in various models, improvements under squared error loss to truncated estimators are obtained. Further related historical developments are given in the review paper of van Eeden (1996).

We conclude this section by expanding upon the case of a positive (or lower-bounded) normal mean θ, for $X \sim N(\theta, 1), \theta \geq 0$. While a plausible and natural estimator is given by the MLE $max(0, X)$, its efficiency requires examination perhaps because it discards part of the sufficient statistic X (i.e., the MLE gives a constant estimate on the region $X \leq 0$). Moreover, as previously mentioned, the MLE has long been known to be inadmissible (e.g., Sacks, 1963) under squared error loss. Despite the age of this finding, it was not until the paper of Shao and Strawderman (1996a) that explicit improvements were obtained (under squared error loss), and there still remains the open question of finding admissible improvements. As well, Katz's (1961) uniform Generalized Bayes estimator remains (to our knowledge) the only known minimax and admissible estimator of θ (under squared error loss).

5. Estimating parameters with additional information

In this section, we present a class of interesting estimation problems which can be transformed to capitalize on standard solutions for estimation problems in constrained parameter spaces. The key technical aspect of subdividing the estimation problem into distinct pieces that can be handled separately is perhaps due to the early work of Blumenthal, Cohen and Sackrowitz. As well, these types of problems have been addressed in some recent work of Constance van Eeden and Jim Zidek.

Suppose X_j; $j = 1, 2$; are independently distributed as $N_p(\theta_j, \sigma_j^2 I_p)$, with $p \geq 1$ and known σ_1^2, σ_2^2. Consider estimating θ_1 under squared error loss $L(\theta_1, d) =$

$\|d-\theta_1\|^2$ with the prior information $\theta_1-\theta_2\in A$; A being a proper subset of $\Re^p$. For instance, with order restrictions of the form $\theta_{1,i}\geq\theta_{2,i}, i=1,\ldots p$, we would have $A=(\Re^+)^p$. Heuristics suggest that the independently distributed X_2 can be used in conjunction with the information $\theta_1-\theta_2\in A$ to construct estimators that improve upon the unrestricted MLE (and UMVU estimator) $\delta_0(X_1,X_2)=X_1$. For instance, suppose $\frac{\sigma_2^2}{\sigma_1^2}\approx 0$, and that A is convex. Then, arguably, estimators of θ_1 should shrink towards $A+x_2=\{\theta_1:\theta_1-x_2\in A\}$. The recognition of this possibility (for $p=1$ and $A=(0,\infty)$) goes back at least as far as Blumenthal and Cohen (1968a), or Cohen and Sackrowitz (1970); and is further discussed in some detail by van Eeden and Zidek (2003).

Following the rotation technique used by Blumenthal and Cohen (1968a), Cohen and Sackrowitz (1970), van Eeden and Zidek (2001,2003) among others, we illustrate in this section how one can exploit the information $\theta_1-\theta_2\in A$; for instance to improve on the unrestricted MLE $\delta_0(X_1,X_2)=X_1$. It will be convenient to define $\mathcal{C}_1$ as the following subclass of estimators of θ_1:

Definition 1.

$$\mathcal{C}_1=\Big\{\delta_\phi:\delta_\phi(X_1,X_2)=Y_2+\phi(Y_1),\\ \text{with}\quad Y_1=\frac{X_1-X_2}{1+\tau},\ Y_2=\frac{\tau X_1+X_2}{1+\tau},\quad\text{and }\tau=\sigma_2^2/\sigma_1^2\Big\}.$$

Note that the above defined Y_1 and Y_2 are independently normally distributed (with $E[Y_1]=\mu_1=\frac{\theta_1-\theta_2}{1+\tau}$, $E[Y_2]=\mu_2=\frac{\tau\theta_1+\theta_2}{1+\tau}$, $Cov(Y_1)=\frac{\sigma_1^2}{1+\tau}I_p$, and $Cov(Y_2)=\frac{\tau\sigma_1^2}{1+\tau}I_p$). Given this independence, the risk function of δ_ϕ (for $\theta=(\theta_1,\theta_2)$) becomes

$$\begin{aligned}R\big(\theta,\delta_\phi(X_1,X_2)\big)&=E_\theta\big[\|Y_2+\phi(Y_1)-\theta_1\|^2\big]\\&=E_\theta\left[\left\|\left(Y_2-\frac{\tau\theta_1+\theta_2}{1+\tau}\right)+\left(\phi(Y_1)-\frac{\theta_1-\theta_2}{1+\tau}\right)\right\|^2\right]\\&=E_\theta\big[\|Y_2-\mu_2\|^2\big]+E_\theta\big[\|\phi(Y_1)-\mu_1\|^2\big].\end{aligned}$$

Therefore, the performance of $\delta_\phi(X_1,X_2)$ as an estimator of θ_1 is measured solely by the performance of $\phi(Y_1)$ as an estimator of μ_1 under the model $Y_1\sim N_p(\mu_1,\frac{\sigma_1^2}{1+\tau}I_p)$, with the restriction $\mu_1\in C=\{y:(1+\tau)y\in A\}$. In particular, one gets the following dominance result.

Proposition 1. *For estimating θ_1 under squared error loss, with $\theta_1-\theta_2\in A$, the estimator $\delta_{\phi_1}(X_1,X_2)$ will dominate $\delta_{\phi_0}(X_1,X_2)$ if and only if*

$$E_{\mu_1}\big[\|\phi_1(Y_1)-\mu_1\|^2\big]\leq E_{\mu_1}\big[\|\phi_0(Y_1)-\mu_1\|^2\big],$$

for $\mu_1\in C$ (with strict inequality for some μ_1).

We pursue with some applications of Proposition 1, which we accompany with various comments and historical references.

(A) *Case where $\delta_{\phi_0}(X_1,X_2)=X_1$ (i.e., the unrestricted mle of θ_1), and where A is convex with a non empty interior.*

This estimator arises as a member of $\mathcal{C}_1$ for $\phi_0(Y_1)=Y_1$. Hartigan's result (Theorem 1) applies to $\phi_0(Y_1)$ (since A convex implies C convex), and tells

us that the Bayes estimator $\phi_{U_C}(Y_1)$ of μ_1 with respect to a uniform prior on C dominates $\phi_0(Y_1)$ (under squared-error loss). Hence, Proposition 1 applies with $\phi_1 = \phi_{U_C}$, producing the following dominating estimator of $\delta_0(X_1, X_2)$:

$$\delta_{\phi_{U_C}}(X_1, X_2) = Y_2 + \phi_{U_C}(Y_1). \tag{4}$$

For $p = 1$ and $A = [-m, m]$, the dominance of δ_{ϕ_0} by the estimator given in (4) was established by van Eeden and Zidek (2001), while for $p = 1$ and $A = [0, \infty)$ (or $A = (-\infty, 0]$), this dominance result was established by Kubokawa and Saleh (1994). In both cases, Kubokawa's IERD method, as presented in Section 3, was utilized to produce a class of dominating estimators which includes $\delta_{\phi_{U_C}}(X_1, X_2)$. As was the case in Section 2, these previously known dominance results yield extensions to sets A which are hyperrectangles or intersection of half-spaces, but Hartigan's result yields a much more general result.

Remark 4. Here are some additional notes on previous results related to the case $p = 1$ and $A = [0, \infty)$. Kubokawa and Saleh (1994) also provide various extensions to other distributions with monotone likelihood ratio and to strict bowl-shaped losses, while van Eeden and Zidek (2003) introduce an estimator obtained from a weighted likelihood perspective and discuss its performance in comparison to several others including $\delta_{\phi_{U_C}}(X_1, X_2)$. The admissibility and minimaxity of $\delta_{\phi_{U_C}}(X_1, X_2)$ (under squared error loss) were established by Cohen and Sackrowitz (1970). Further research concerning this problem, and the related problem of estimating jointly θ_1 and θ_2, has appeared in Blumenthal and Cohen (1968b), Brewster and Zidek (1974), and Kumar and Sharma (1988) among many others. There is equally a substantial body of work concerning estimating a parameter θ_1 (e.g., location, scale, discrete family) under various kinds of order restrictions involving k parameters $\theta_1, \ldots, \theta_k$ (other than work referred to elsewhere in this paper, see for instance van Eeden and Zidek, 2001, 2003 for additional annotated references).

Another dominating estimator of $\delta_{\phi_0}(X_1, X_2) = X_1$, which may be seen as a consequence of Proposition 1, is given by $\delta_{\phi_1}(X_1, X_2) = Y_2 + \phi_{\text{mle}}(Y_1)$, where $\phi_{\text{mle}}(Y_1)$ is the mle of μ_1, $\mu_1 \in C$. This is so because, as remarked upon in Section 2, $\phi_1(Y_1) = \phi_{\text{mle}}(Y_1)$ dominates under squared error loss, as an estimator of $\mu_1; \mu_1 \in C$; $\phi_0(Y_1) = Y_1$. Observe further that the maximum likelihood estimator $\delta_{\text{mle}}(X_1, X_2)$ of θ_1 for the parameter space $\Theta = \{(\theta_1, \theta_2) : \theta_1 - \theta_2 \in A, \tau\theta_1 + \theta_2 \in \Re^p\}$ is indeed given by: $\delta_{\text{mle}}(X_1, X_2) = \hat{(\mu_2)}_{\text{mle}} + \hat{(\mu_1)}_{\text{mle}} = Y_2 + \phi_{\text{mle}}(Y_1)$, given the independence and normality of Y_1 and Y_2, and the fact that Y_2 is the MLE of μ_2 ($\mu_2 \in \Re^p$).

Our next two applications of Proposition 1 deal with the estimator $\delta_{\text{mle}}(X_1, X_2)$.

(B) *Case where A is a ball and $\delta_{\phi_0}(X_1, X_2) = \delta_{mle}(X_1, X_2)$.*

For the case where $\theta_1 - \theta_2 \in A$, with A being a p-dimensional ball of radius m centered at 0, the estimator $\delta_{\text{mle}}(X_1, X_2)$ arises as a member of $\mathcal{C}_1$ for $\phi_0(Y_1) = \phi_{\text{mle}}(Y_1) = (\|Y_1\| \wedge \frac{m}{1+\tau})\frac{Y_1}{\|Y_1\|}$. By virtue of Proposition 1, it follows that dominating estimators $\phi_*(Y_1)$ of $\phi_{\text{mle}}(Y_1)$ (for the ball with $\|\mu_1\| \leq m/(1+\tau)$), such as those given by Marchand and Perron (2001) (see Section 4.3 above), yield dominating estimators $\delta_{\phi_*}(X_1, X_2) = \frac{\tau X_1 + X_2}{1+\tau} + \phi_*(\frac{X_1 - X_2}{1+\tau})$ of $\delta_{\text{mle}}(X_1, X_2)$.

(C) *Case where* $A = [0, \infty)$*, and* $\delta_{\phi_0}(X_1, X_2) = \delta_{mle}(X_1, X_2)$. This is similar to (B), and dominating estimators can be constructed by using Shao and Strawderman's (1996) dominating estimators of the MLE of a positive normal mean (see van Eeden and Zidek, 2001, Theorem 4.3)

Observe that results in (B) (for $p = 1$) and (C) lead to further applications of Proposition 1 for sets A which are hyperrectangles or intersection of half-spaces. We conclude by pointing out that the approach used in this section may well lead to new directions in future research. For instance, the methods used above could be used to specify dominating estimators for the case $p \geq 3$, (of $\delta_{\phi_0}(X_1, X_2) = Y_2 + \phi_0(Y_1)$) of the form $\phi_2(Y_2) + \phi_1(Y_1)$ where, not only is $\phi_1(Y_1)$ a dominating estimator of $\phi_0(Y_1)$ for $\mu_1 \in C$, but for $p \geq 3$, $\phi_2(Y_2)$ is a Stein-type estimator of μ_2 which dominates Y_2.

6. Minimax estimation

This section presents an overview of minimax estimation in compact parameter spaces, with a focus on the case of an interval constraint of the type $\theta \in [a, b]$ and analytical results giving conditions for which the minimax estimator is a Bayesian estimator with respect to a boundary prior on $\{a, b\}$. Historical elements are first described in Section 6.1, a somewhat novel expository example is presented in Section 6.2., and we further describe complementary results in Section 6.3.

6.1. Two point least favourable priors

With the criteria of minimaxity playing a vital role in the development of statistical theory and practice; as reviewed in Brown (1994) or Strawderman (2000) for instance; the results of Casella and Strawderman (1981), as well as those of Zinzius (1981) most certainly inspired a lot of further work. These results presented analytically obtained minimax estimators, under squared error loss, of a normal model mean θ, with known variance, when θ is known to be restricted to a small enough interval. More precisely, Casella and Strawderman showed, for $X \sim N(\theta, 1)$ with $\theta \in \Theta = [-m, m]$; (there is no loss of generality in assuming the variance to be 1, and the interval to be symmetric about 0); that the uniform boundary Bayes estimator $\delta_{BU}(x) = m \tanh(mx)$ is unique minimax iff $m \leq m_0 \approx 1.0567$. Furthermore, they also investigated three-point priors supported on $\{-m, 0, m\}$, and obtained sufficient conditions for such a prior to be least favourable. It is worth mentioning that these results give immediately minimax multivariate extensions to rectangular constraints where $X_i \sim N(\theta_i, 1); i = 1, \ldots, p$; with $|\theta_i| \leq m_i \leq m_0$, under losses $\sum_{i=1}^{p} \omega_i (d_i - \theta_i)^2$, (with arbitrary positive weights ω_i), since the least favourable prior for estimating $(\theta_1, \ldots, \theta_p)$ is obtained, in such a case, as the product of the least favourable priors for estimating $\theta_1, \ldots, \theta_p$ individually. Now, the above minimaxity results were obtained by using the following well-known criteria for minimaxity (e.g., Berger, 1985, Section 5.3, or Lehmann and Casella, 1998, Section 5.1).

Lemma 1. *If* δ_π *is a Bayes estimator with respect to a prior distribution* π*, and* $S_\pi = \{\theta \in \Theta : \sup_\theta \{R(\theta, \delta_\pi)\} = R(\theta, \delta_\pi)\}$*, then* δ_π *is minimax whenever* $P_\pi(\theta \in S_\pi) = 1$.

Casella and Strawderman's work capitalized on Karlin's (1957) sign change arguments for implementing Lemma 1 while, in contrast, the sufficient conditions

obtained by Zinzius concerning the minimaxity of $\delta_{BU}(X)$ were established using the "convexity technique" as stated as part (b) of the following Corollary to Lemma 1. Part (a), introduced here as an alternative condition, will be used later in this section.

Corollary 1. *If δ_π is a Bayes estimator with respect to a two-point prior on $\{a, b\}$ such that $R(a, \delta_\pi) = R(b, \delta_\pi)$, then δ_π is minimax for the parameter space $\Theta = [a, b]$ whenever, as a function of $\theta \in [a, b]$,*

(a) $\frac{\partial}{\partial\theta} R(\theta, \delta_\pi)$ *has at most one sign change from $-$ to $+$; or*

(b) $R(\theta, \delta_\pi)$ *is convex.*

Although the convexity technique applied to the bounded normal mean problem gives only a lower bound for m_0; (Bader and Bischoff (2003) report that the best known bound using convexity is $\frac{\sqrt{2}}{2}$, as given by Bischoff and Fieger (1992)); it has proven very useful for investigating least favourable boundary supported priors for other models and loss functions. In particular, DasGupta (1985) used subharmonicity to establish, for small enough compact and convex parameter spaces under squared error loss, the inevitability of a boundary supported least favourable prior for a general class of univariate and multivariate models. As well, the work of Bader and Bischoff (2003), Boratyńska (2001), van Eeden and Zidek (1999), and Eichenauer-Hermann and Fieger (1992), among others, establish this same inevitability with some generality with respect to the loss and/or the model. Curiously, as shown by Eichenauer-Hermann and Ickstadt (1992), and Bischoff and Fieger (1993), there need not exist a boundary least favourable prior for convex, but not strictly convex, losses. Indeed, their results both include the important case of a normal mean restricted to an interval and estimated with absolute-value loss, where no two-point least favourable prior exists.

6.2. Two-point least favourable priors in symmetric location families

We present here a new development for location family densities (with respect to Lebesgue measure on $\Re^1$) of the form

$$f_\theta(x) = e^{-h(x-\theta)}, \quad \text{with convex and symmetric } h. \tag{5}$$

These assumptions on h imply that such densities f_θ are unimodal, symmetric about θ, and possess monotone increasing likelihood ratio in X. For estimating θ with squared error loss under the restriction $\theta \in [-m, m]$, our objective here is to present a simple illustration of the inevitability of a boundary supported least favourable prior for small enough m, i.e., $m \leq m_0(h)$. Namely, we give for densities in (5) with concave $h'(x)$ for $x \geq 0$ (this implies convex $h'(x)$ for $x \leq 0$) a simple lower bound for $m_0(h)$. We pursue with two preliminary lemmas; the latter one giving simple and general conditions for which a wide subclass of symmetric estimators (i.e., equivariant under sign changes) $\delta(X)$ of θ have increasing risk $R(\theta, \delta(X))$ in $|\theta|$ under squared error loss.

Lemma 2. *If g is a bounded and almost everywhere differentiable function, then under* (5)*:*

$$\frac{d}{d\theta} E_\theta\, [\, g(X)] = E_\theta\, [\, g'(X)].$$

Proof. First, interchange derivative and integral to obtain $\frac{d}{d\theta} E_\theta[\, g(X)\,] = E_\theta\, [g(X)\, h'(X - \theta)]$. Then, integrating by parts yields the result. □

Lemma 3. *For models in* (5)*, and estimators* $\delta(X)$ *with the properties:* (a) $\delta(x) = -\delta(-x)$*;* (b) $\delta'(x) \geq 0$*; and* (c) $\delta'(x)$ *decreasing in* $|x|$*; for all* $x \in \Re$*; either one of the following conditions is sufficient for* $R(\theta, \delta(X))$ *to be increasing in* $|\theta|$*;* $|\theta| \leq m$*:*

(i) $E_\theta[\delta(X)] \leq \theta(1 - E_\theta[\delta'(X)])$*, for* $0 \leq \theta \leq m$*;*

(ii) $E_\theta[\delta(X)] \leq \theta(1 - \delta'(0))$*, for* $0 \leq \theta \leq m$*;*

(iii) $\delta'(0) \leq \frac{1}{2}$.

Proof. It will suffice to work with the condition $\frac{\partial}{\partial\theta} R(\theta, \delta(X)) \geq 0$, $0 \leq \theta \leq m$, since $R(\theta, \delta(X))$ is an even function of θ, given property (a) and the symmetry of h. Differentiating directly the risk and using Lemma 2, we obtain

$$\frac{1}{2}\frac{\partial}{\partial\theta} R\big(\theta, \delta(X)\big) = \theta - E_\theta[\delta(X)] - \theta E_\theta\big[\delta'(X)\big] + E_\theta\big[\delta(X)\delta'(X)\big].$$

With properties (a) and (b), the function $\delta(x)\delta'(x)$ changes signs once (at $x = 0$) from $-$ to $+$, and, thus, sign change properties under h imply that $E_\theta[\delta(X)\delta'(X)\,]$ changes signs at most once from $-$ to $+$ as a function of θ. Since $E_0[\delta(X)\delta'(X)] = 0$ by symmetry of $\delta(x)\delta'(x)$ and h, we must have $E_\theta[\delta(X)\delta'(X)] \geq 0$ for $\theta \geq 0$. It then follows that

$$\frac{1}{2}\frac{\partial}{\partial\theta} R\big(\theta, \delta(X)\big) \geq \theta - E_\theta\big[\delta(X)\big] - \theta E_\theta\big[\delta'(X)\big];$$

and this yields directly sufficient condition (i). Now, property (c) tells us that $\delta'(x) \leq \delta'(0)$, and this indicates that condition (ii) implies (i), hence its sufficiency. Finally, condition (iii) along with Lemma 2 and the properties of $\delta(X)$ implies (ii) since $\frac{\partial}{\partial\theta} E_\theta[\delta(X) + \theta(\delta'(0) - 1)] = E_\theta[\delta'(X) + (\delta'(0) - 1)] \leq E_\theta[2\delta'(0) - 1] \leq 0$, and $E_\theta[\delta(X) + \theta(\delta'(0) - 1)]|_{\theta=0} = 0$.

We pursue by applying Lemma 3 to the case of the boundary uniform Bayes estimator $\delta_{BU}(X)$ to obtain, by virtue of Corollary 1, part (a), a minimaxity result for $\delta_{BU}(X)$. □

Corollary 2. *For models in* (5)*,* $\delta_{BU}(X)$ *is a unique minimax estimator of* θ (*under squared error loss*) *for the parameter space* $[-m, m]$ *when either one of the following situations arises:*

(A) *Condition* (i) *of Lemma* 3 *holds;*

(B) $h'(x)$ *is concave for* $x \geq 0$*, and condition* (ii) *of Lemma* 3 *holds;*

(C) $h'(x)$ *is concave for* $x \geq 0$*, and* $m \leq m^*(h)$ *where* $m^*(h)$ *is the solution in* m *of the equation* $mh'(m) = \frac{1}{2}$.

Proof. We apply Corollary 1, part (a), and Lemma 3. To do so, we need to investigate properties (b) and (c) of Lemma 3 for the estimator $\delta_{BU}(X)$ (property (a) is necessarily satisfied since the uniform boundary prior is symmetric). Under model (5), the Bayes estimator $\delta_{BU}(X)$ and its derivative (with respect to X) may be expressed as:

$$\delta_{BU}(x) = \frac{me^{-h(x-m)} - me^{-h(x+m)}}{e^{-h(x-m)} + e^{-h(x+m)}} = m\tanh\left(\frac{h(x+m) - h(x-m)}{2}\right); \qquad (6)$$

and

$$\delta'_{BU}(x) = \{m^2 - \delta_{BU}(x)^2\}\left\{\frac{h'(x+m) - h'(x-m)}{2m}\right\}. \tag{7}$$

Observe that $|\delta_{BU}(x)| \leq m$, and $h'(x+m) \geq h'(x-m)$ by the convexity of h, so that $\delta'_{BU}(x) \geq 0$ given (7). This establishes property (b) of Lemma 3, and part (A). Now, $m^2 - \delta^2_{BU}(x)$ is decreasing in $|x|$, and so is $h'(x+m) - h'(x-m)$ given, namely, the concavity of $h'(x)$ for $x \geq 0$. This tells us that $\delta_{BU}(X)$ verifies property (c) of Lemma 3, and (B) follows. Hence, condition (iii) of Lemma 3 applies becoming equivalent to $mh'(m) \leq \frac{1}{2}$, as $\delta'(0) = mh'(m)$ by (7). Finally, the result follows by the fact that $mh'(m)$ is a continuous and increasing of function of m, $m > 0$. □

Remark 5. As the outcome of the above argument, combining both sign change arguments and convexity considerations, containing other elements which may be independent interest, part (C) of Corollary 2 gives a simple sufficient condition for the minimaxity of δ_{BU}, and is applicable to a wide class of models in (5). Namely, for Exponential Power families where, in (5), $h(y) = \alpha|y|^\beta$ with $\alpha > 0$ and $1 \leq \beta \leq 2$, part (C) of Corollary 2 applies, and tells us that $\delta_{BU}(X)$ (which may be derived from (6)) is unique minimax whenever $m \leq m^*(h) = (\frac{1}{2\alpha\beta})^{1/\beta}$. In particular for double-exponential cases, (i.e., $\beta = 1$), we obtain $m^*(h) = \frac{1}{2\alpha}$; and for the standard normal case, (i.e., $(\alpha, \beta) = (\frac{1}{2}, 2)$, we obtain $m^*(h) = \frac{\sqrt{2}}{2}$. Observe that the normal case $m^*(h)$ matches the one given by Bischoff and Fieger (1992); and that, as expected with the various lower bounds used for the derivative of the risk, it falls somewhat below Casella and Strawderman's necessary and sufficient cutoff point of $m_0 \approx 1.05674$.

6.3. Some additional results and comments

The problem considered in Section 6.2, was studied also by Eichenauer-Hermann and Ickstadt (1992), who obtained similar results using a convexity argument for the models in (5) with $L^p, p > 1$ loss. Additional work concerning least favourable boundary priors for various models can be found in: Moors (1985), Berry (1989), Eichenauer (1986), Chen and Eichenauer (1988), Eichenauer-Hermann and Fieger (1989), Bischoff (1992), Bischoff and Fieger (1992), Berry (1993), Johnstone and MacGibbon (1992), Bischoff, Fieger and Wurlfert (1995), Bischoff, Fieger, and Ochtrop (1995), Marchand and MacGibbon (2000), and Wan, Zou and Lee (2000).

Facilitated by results guaranteeing the existence of a least favourable prior supported on a finite number of points (e.g., Ghosh, 1964), the dual problem of searching numerically for a least favourable prior π, as presented in Lemma 1, is very much the standard approach for minimax estimation problems in compact parameter spaces. Algorithms to capitalize on this have been presented by Nelson (1965), and Kempthorne (1987), and have been implemented by Marchand and MacGibbon (2000), for a restricted binomial probability parameter, MacGibbon, Gourdin, Jaumard, and Kempthorne (2000) for a restricted Poisson parameter, among others. Other algorithms have been investigated by Gourdin, Jaumard, and MacGibbon (1994).

Analytical and numerical results concerning the related criteria of Gamma-Minimaxity in constrained parameter spaces have been addressed by Vidakovic and DasGupta (1996), Vidavovic (1993), Lehn and Rummel (1987), Eichenauer and Lehn (1989), Bischoff (1992), Bischoff and Fieger (1992), Bischoff, Fieger and Wurlfert (1995), and Wan et al. (2000).

For spherical bounds of the form $\|\theta\| \leq m$, Berry (1990) generalized Casella and Strawderman's minimaxity of δ_{BU} result for multivariate normal models $X \sim N_p(\theta, I_p)$. He showed with sign change arguments that δ_{BU} is unique minimax for $m \leq m_0(p)$, giving defining equations for $m_0(p)$. Recently, Marchand and Perron (2002) showed that $m_0(p) \geq \sqrt{p}$, and that $m_0(p)/\sqrt{p} \approx 1.15096$ for large p. For larger m, least favourable distributions are mixtures of a finite number of uniform distributions on spheres (see Robert, 2001, page 73, and the given references), but the number, position and mixture weights of these spheres require numerical evaluation.

Early and significant contributions to the study of minimax estimation of a normal mean restricted to an interval or a ball of radius m, were given by Bickel (1981) and Levit (1980). These contributions consisted of approximations to the minimax risk and least favourable prior for large m under squared error loss. In particular, Bickel showed that, as $m \to \infty$, the least favourable distributions rescaled to $[-1, 1]$ converge weakly to a distribution with density $\cos^2(\pi x/2)$, and that the minimax risks behave like $1 - \frac{\pi^2}{8m^2} + o(m^{-2})$. Extensions and further interpretations of these results were given by Melkman and Ritov (1987), Gajek and Kaluszka (1994), and Delampady and others (2001). There is also a substantial literature on the comparative efficiency of mimimax procedures and affine linear minimax estimators for various models, restricted parameter spaces, and loss functions. A small sample of such work includes Pinsker (1980), Ibragimov and Hasminskii (1984), Donoho, Liu and MacGibbon (1990), and Johnstone and MacGibbon (1992,1993).

Acknowledgements

The authors are grateful to John Hartigan for showing us his result (Theorem 1), and permitting us to make use of it in this review paper. The authors also benefited from discussions with François Perron, helpful in particular for the development presented in Section 6.2. Finally, the authors are thankful to Anirban DasGupta for a diligent reading of a first draft of this review paper, as well as constructive comments and suggestions.

References

[1] Akkerboom, J. C. (1990). *Testing problems with linear or angular inequality constraints*, Lecture Notes in Statistics, **62**, Springer-Verlag, New York. MR1044097

[2] Bader, G. and Bischoff, W. (2003). Old and new aspects of minimax estimation of a bounded parameter, *Mathematical Statistics and Applications, Festschrift for Constance van Eeden*. IMS Lecture Notes-Monograph Series, **42**, pp. 15–30.

[3] Barlow, R. E., Bartholomew, J. M., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions, The Theory and Application of Isotonic Regression*. Wiley, New York. MR326887

[4] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 2nd edition. MR804611

[5] Berry, C. (1989). Bayes minimax estimation of a Bernouilli p in a restricted parameter space, *Communications in Statistics: Theory and Methods*, **18**, 4607–4616. MR1046729

[6] Berry, C. (1990). Minimax estimation of a bounded normal mean vector, *Journal of Multivariate Analysis*, **35**, 130–139. MR1084946

[7] Berry, C. (1993). Minimax estimation of a restricted exponential location parameter, *Statistics and Decisions*, **11**, 307–316. MR1261841

[8] Bickel, P. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted, *Annals of Statistics*, **9**, 1301–1309. MR630112

[9] Bischoff, W. (1992). Minimax estimation and Γ-minimax estimation for functions of a scale parameter family under L_p loss, *Statistics and Decisions*, **29**, 45–61. MR1165702

[10] Bischoff, W. and Fieger, W. (1992). Minimax estimators and Γ-minimax estimators for a bounded normal mean under the loss $\ell_p(\theta, d) = |\theta - d|^p$, *Metrika*, **39**, 185–197. MR1173577

[11] Bischoff, W. and Fieger, W. (1993). On least favourable two-point priors and minimax estimators under absolute error loss, *Metrika*, **40**, 283–298. MR1242740

[12] Bischoff, W., Fieger, W., and Ochtorp, S. (1995). Minimax estimation for the bounded mean of a bivariate normal distribution, *Metrika*, **42**, 379–394. MR1366381

[13] Bischoff, W., Fieger, W., and Wulfert, S. (1995). Minimax and Γ-minimax estimation of a bounded normal mean under linex loss, *Statistics and Decisions*, **13**, 287–298. MR1345373

[14] Blumenthal, S. and Cohen, A. (1968a). Estimation of the larger translation parameter, *Annals of Mathematical Statistics*, **39**, 502–516. MR223006

[15] Blumenthal, S. and Cohen, A. (1968b). Estimation of two ordered translation parameters, *Annals of Mathematical Statistics*, **39**, 517–530. MR223007

[16] Boratyńska, A. (2001). Two point priors and minimax estimation of a bounded parameter, *Unpublished manuscript.*

[17] Brewster, J. F. and Zidek, J. K. (1974). Improving on equivariant estimators, *Annals of Statistics*, **2**, 21–38. MR381098

[18] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems, *Annals of Mathematical Statistics*, **42**, 855-903. MR286209

[19] Brown, L. D. (1994). Minimaxity, more or less, in *Statistical Decision Theory and Related Topics, V*, editors S. S. Gupta and J. O. Berger. Springer, New York, pp. 1–18. MR1286291

[20] Brown, L. D. (1996). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* IMS Lecture Notes-Monograph Series, **9**.

[21] Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean, *Annals of Statistics*, **9**, 870–878. MR619290

[22] Charras, A. and van Eeden, C. (1991,a). Bayes and admissibility properties of estimators in truncated parameter spaces, *Canadian Journal of Statistics*, **19**, 121–134. MR1128402

[23] Charras, A. and van Eeden, C. (1991,b). Limit of Bayes estimators in convex truncated parameter spaces, Statistics & Probability Letters, **11**, 479–483. MR1116740

[24] Charras, A. and van Eeden, C. (1992). Bayes properties of estimators of location parameters in truncated parameter spaces, *Statistics and Decisions*, **10**, 81–86. MR1165705

[25] Charras, A. and van Eeden, C. (1994). Inadmissibility for squared loss when the parameter to be estimated is restricted to the interval $[a, \infty)$, *Statistics and Decisions*, **12**, 257–266. MR1309666

[26] Chen, L. and Eichenauer, J. (1988). Two point priors and Γ-minimax estimation in families of uniform distributions, *Statistical Papers*, *29*, 45–57. MR947507

[27] Cohen, A. and Sackrowitz, H. B. (1970). Estimation of the last mean of a monotone sequence, *Annals of Mathematical Statistics*, **41**, 2021–2034. MR270483

[28] DasGupta, A. (1985). Bayes minimax estimation in multiparameter families when the parameter space is restricted to a bounded convex set, *Sankhyā: Series A*, **47**, 326–332. MR863726

[29] Delampady, M., DasGupta, A., Casella, G., Rubin, H., and Strawderman, W. E. (2001). A new approach to default priors and robust Bayes methodology. *Canadian Journal of Statistics*, **29**, 437–450. MR1872645

[30] Drees, H. (1999). On fixed-length confidence intervals for a bounded normal mean, *Statistics & Probability Letters*, **44**, 399–404. MR1721440

[31] Donoho, D. L., Liu, R. C., and MacGibbon, B. (1990). Minimax estimation over hyperrectangles with implications, *Annals of Statistics*, **18**, 1416–1437. MR1062717

[32] Eichenauer, J. (1986). Least favourable two-point priors in estimating the bounded location parameter of a noncentral exponential distribution, *Statistics and Decisions*, **4**, 389–392. MR876878

[33] Eichenauer-Herrmann, J. and Fieger, W. (1989). Minimax estimation in scale parameter families when the parameter interval is bounded, *Statistics and Decisions*, **7**, 363–376. MR1038495

[34] Eichenauer-Hermann, J. and Lehn, J. (1989). Gamma-minimax estimators for a bounded normal mean under squared error loss, *Statistics and Decisions*, **7**, 37–62. MR1000926

[35] Eichenauer-Herrmann, J. and Fieger, W. (1992). Minimax estimation under convex loss when the parameter interval is bounded, *Metrika*, **39**, 27–43. MR1162684

[36] Eichenauer-Hermann, J. and Ickstadt, K. (1992). Minimax estimators for a bounded location parameter, *Metrika*, **39**, 227–237. MR1173580

[37] Farrell, R. H. (1964). Estimators of a location parameter in the absolutely continuous case, *Annals of Mathematical Statistics*, **35**, 949–998. MR171359

[38] Funo, E. (1991). Inadmissibility results of the MLE for the multinomial problem when the parameter space is restricted or truncated, *Communications in Statistics: Theory and Methods*, **20**, 2863–2880. MR1141983

[39] Gatsonis, C., MacGibbon, B., and Strawderman, W. E. (1987). On the estimation of a restricted normal mean, *Statistics & Probability Letters*, **6**, 21–30. MR907255

[40] Gajek, L., Kaluszka, M. (1994). Lower bounds for the asymptotic Bayes risk in the scale model (with applications to the second order minimax estimation), *Annals of Statistics*, **22**, 1831–1839. MR1329170

[41] Gajek, L., Kaluszka, M. (1995). Nonexponential applications of a global Cramèr-Rao inequality, *Statistics*, **26**, 111–122. MR1318207

[42] Ghosh, M. N. (1964). Uniform approximation of minimax point estimates, *Annals of Mathematical Statistics*, **35**, 1031–1047. MR164418

[43] Gourdin, E., Jaumard, B., and MacGibbon, B. (1994). Global optimization decomposition methods for bounded parameter minimax risk evaluation, *Siam Journal of Scientific Computing*, **15**, 16–35. MR1257152

[44] Gupta, A. K. and Rohatgi, V. K. (1980). On the estimation of a restricted mean, *Journal of Statistical Planning and Inference*, **4**, 369–379. MR596771

[45] Hartigan, J. (2003). Uniform priors on convex sets improve risk, *Statistics & Probability Letters*. To appear.

[46] Hwang, G. and Peddada, S. D. (1994). Confidence intervals subject to order restrictions, *Annals of Statistics*, **22**, 67–93. MR1272076

[47] Ibragimov, I. A. and Hasminskii, R. Z. (1984). Nonparametric estimation of the value of a linear functional in a Gaussian white noise, *Theory of Probability and Its Applications*, **29**, 1–32. MR739497

[48] Iliopoulos, G. (2000). A note on decision theoretic estimation of ordered parameters, *Statistics & Probability Letters*, **50**, 33–38. MR1804623

[49] Iliopoulos, G. and Kourouklis, S. (2000). Interval estimation for the ratio of scale parameters and for ordered scale parameters, *Statistics and Decisions*, **18**, 169–184. MR1782343

[50] Iwasa, M. and Moritani, Y. (1997). A note on the admissibility of the maximum likelihood estimator for a bounded normal mean, *Statistics & Probability Letters*, **32**, 99–105. MR1439501

[51] Johnstone, I. and MacGibbon, B. (1992). Minimax estimation of a constrained Poisson vector, *Annals of Statistics*, **20**, 807–831. MR1165594

[52] Johnstone, I. and MacGibbon, B. (1993). Asymptotically minimax estimation of a constrained Poisson vector via polydisc transforms, *Annales de l'Institut Henri Poincaré- Probabilité et Statistiques*, **2**, 289–319. MR1227421

[53] Jafari Jozani, M., Nematollahi, N., Shafie, K. (2002). An admissible minimax estimator of a bounded scale-parameter in a subclass of the exponential family under scale-invariant squared error loss, *Statistics & Probability Letters*, **60**, 437–444. MR1947183

[54] Kamberova, G. and Mintz, M. (1999). Minimax rules under zero-one loss for a restricted location parameter, *Journal of Statistical Planning and Inference*, **79**, 205–221. MR1700760

[55] Karlin, S. (1957). Pólya type distributions, II, *Annals of Mathematical Statistics*, **28**, 281–309. MR88148

[56] Katz. M. (1961). Admissible and Minimax Estimates of Parameters in Truncated Spaces, *Annals of Mathematical Statistics*, **32**, 136–142. MR119287

[57] Kempthorne, P. J. (1987). Numerical specification of discrete least favorable prior distributions, *Siam Journal of Scientific Computing*, **8**, 171–184. MR879409

[58] Kempthorne, P. J. (1988). Dominating inadmissible procedures using compromise decision theory, *Statistical Decision Theory and Related Topics IV*, editors S. S. Gupta and J. O. Berger. Springer-Verlag, New York, pp. 381–396. MR927113

[59] Kubokawa, T., (1994a). A Unified Approach to Improving Equivariant Estimators, *Annals of Statistics*, **22**, 290–299. MR1272084

[60] Kubokawa, T., (1994b). Double shrinkage estimation of ratio of scale parameters, *Annals of the Institute of Statistical Mathematics*, **46**, 95–119. MR1272751

[61] Kubokawa, T. and Saleh, A. K. MD. E. (1994). Estimation of location and scale parameters under order restrictions. *Journal of Statistical Research*, **28**, 41–51. MR1370413

[62] Kubokawa, T., (1998). The Stein phenomenon in simultaneous estimation: A review. *Applied Statistical Science III* (editors S. E. Ahmed, M. Ahsanullah, and B. K. Sinha), NOVA Science Publishers, New York, pp. 143–173. MR1673649

[63] Kubokawa, T., (1999). Shrinkage and modification techniques in estimation of variance and the related problems: A review. *Communications in Statistics: Theory and Methods*, **28**, 613–650. MR1703774

[64] Kumar S. and Sharma, D. (1988). Simultaneous estimation of ordered parameters. *Communications in Statistics: Theory and Methods*, **17**, 4315–4336. MR981031

[65] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, 2nd edition. MR1639875

[66] Lehn, J. and Rummel, F. (1987). Gamma-minimax estimation of a binomial probability under squared error loss, *Statistics and Decisions*, **5**, 229–249. MR905239

[67] Levit, B.Y.A. (1980). On asymptotic minimax estimation of the second order, *Theory Probab. Appl.*, **25**, 552–568. MR582585

[68] MacGibbon, B., Gourdin, E., Jaumard, B., and Kempthorne, P. J. (2000). Minimax estimation of exponential family means over l_p bodies under quadratic loss, in *CMS Conference Proceedings*, **26**, 319–330. MR1765018

[69] Marchand, É. and MacGibbon, B. (2000). Minimax estimation of a constrained binomial proportion, *Statistics and Decisions*, **18**, 129–167. MR1782342

[70] Marchand, É. and Perron, F. (2001). Improving on the MLE of a bounded normal mean. *Annals of Statistics*, **29**, 1078–1093. MR1869241

[71] Marchand, É. and Perron, F. (2002). On the minimax estimator of a bounded normal mean. *Statistics & Probability Letters*, **58**, 327–333. MR1923455

[72] Marchand, É. and Perron, F. (2003). Improving on the mle of a bounded location parameter for spherical distributions, *Journal of Multivariate Analysis*. To appear. MR1869241

[73] Marchand, É. and Strawderman, W. E. (2003a). On improving on the minimum risk equivariant estimator of a location parameter which is constrained to an interval or a half-interval, *Annals of the Institute of Statistical Mathematics*. To appear.

[74] Marchand, É. and Strawderman, W. E. (2003b). On improving on the minimum risk equivariant estimator of a scale parameter under a lower-bound constraint, *Journal of Statistical Planning and Inference*. To appear.

[75] Mandelkern, M. (2002). Setting Confidence Intervals for Bounded Parameters with discussion, *Statistical Science*, **17**, 149–172. MR1939335

[76] Melkman, A. A. and Ritov, Y. (1987). Minimax estimation of the mean of a general distribution when the parameter space is restricted, **15**, 432–442. MR885749

[77] Moors, J. J. A. (1981). Inadmissibility of linearly invariant estimators in the truncated parameter spaces, *Journal of the American Statistical Association*, **76**, 910–915. MR650903

[78] Moors, J. J. A. (1985). *Estimation in truncated parameter spaces*, Ph.D. thesis, Tilburg University.

[79] Moors, J. J. A. and van Houwelingen, J. C. (1993). Estimation of linear models with inequality restrictions, Statistica Neerlandica, **47**, 185–198. MR1243854

[80] Nelson, W. (1965). Minimax solution of statistical decision problems by iteration, *Annals of Mathematical Statistics*, **37**, 1643–1657. MR198635

[81] Parsian, A., Sanjari Farsipour, N. (1997). Estimation of parameters of exponential distribution in the truncated space using asymmetric loss function. *Statistical Papers*, 38, 423–443. MR1488641

[82] Perron, F. (2003). Improving on the mle of p for a Binomial(n, p) when p is around 1/2, *Mathematical Statistics and Applications, Festschrift for Constance van Eeden*, IMS Lecture Notes-Monograph Series, **42**, pp. 45–64.

[83] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise, *Problems Inform. Transmission*, 120–133.

[84] Robert, C. (2001). *The Bayesian Choice: From decision-theoretic foundations to computational implementation*. Springer-Verlag, New York, 2nd edition. MR1835885

[85] Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York. MR961262

[86] Rukhin, A. L. (1990). Comments on 'Developments in decision-theoretic variance estimation', by Maatta J. M. and Casella, G., *Statistical Science*, **5**, 113–116.

[87] Sacks, J. (1963). Generalized Bayes solutions in estimation problems, *Annals of Mathematical Statistics*, **34**, 751–768. MR150908

[88] Sackrowitz, H. B. and Strawderman, W. (1974). On the admissibility of the MLE for ordered binomial parameters, *Annals of Statistics*, **2**, 822–828. MR368242

[89] Sackrowitz, H. B. (1982). Procedures for improving the MLE for ordered binomial parameters, *Journal of Statistical Planning and Inference*, **6**, 287–296. MR664351

[90] Shao, P. Y.-S. and Strawderman, W. (1994). Improving on truncated estimators, in *Statistical Decision Theory and Related Topics, V*, editors S. S. Gupta and J. O. Berger. Springer, New York, pp. 369–376. MR1286313

[91] Shao, P. Y.-S. and Strawderman, W. (1996a). Improving on the mle of a positive normal mean, *Statistica Sinica*, **6**, 275–287. MR1379061

[92] Shao, P. Y.-S. and Strawderman, W. (1996b). Improving on truncated linear estimates of exponential and gamma scale parameters, *Canadian Journal of Statistics*, **24**, 105–114. MR1394744

[93] Stark, P. (1992). Affine minimax confidence intervals for a bounded normal mean, *Statistics & Probability Letters*, **13**, 39–44. MR1147637

[94] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Annals of Statistics*, **9**, 1135–1151. MR630098

[95] Strawderman, W. (2000). Minimaxity, *Journal of the American Statistical Assocation*, **95**, 1364–1368. MR1825295

[96] van Eeden, C. (1995). Minimax estimation of a lower-bounded scale parameter of a gamma distribution for scale-invariant squared error loss, *Canadian Journal of Statistics*, **23**, 245–256. MR1363590

[97] van Eeden, C. (1996). Estimation in restricted parameter spaces- some history and recent developments, *CWI Quarterly*, **9**, 69–76. MR1420018

[98] van Eeden, C. and Zidek, J. V. (1999). Minimax estimation of a bounded scale parameter for scale-invariant squared-error loss, *Statistics and Decisions*, **17**, 1–30. MR1701333

[99] van Eeden, C. (2000). Minimax estimation of a lower-bounded scale parameter of an F distribution, *Statistics & Probability Letters*, **46**, 283–286.

[100] van Eeden, C. and Zidek, J. V. (2001). Estimating one of two normal means when their difference is bounded, *Statistics & Probability Letters*, **51**, 277–284. MR1822735

[101] van Eeden, C. and Zidek, J. V. (2003). Combining sample information in estimating ordered normal means, *Sankhyā, Series A*, **64**, 588–610. MR1985401

[102] Vidakovic, B. (1993). On the efficiency of affine minimax rules in estimating a bounded multivariate normal mean, *Communications in Statistics: Theory and Methods*, **22**, 655–669. MR1235867

[103] Vidakovic, B. and DasGupta, A. (1996). Efficiency of linear rules for estimating a bounded normal mean, *Sankhyā* , **58**, 81–100. MR1659051

[104] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York. MR36976

[105] Wan, T., Zou, G., and Lee, A. (2000). Minimax and Γ-minimax estimation for a Poisson distribution under linex loss when the parameter space is restricted, *Statistics & Probability Letters*, **50**, 23–32. MR1804622

[106] Zeytinoglu, M. and Mintz, M. (1984). Optimal fixed size confidence procedures for a restricted parameter space, *Annals of Statistics*, **12**, 945–957. MR751284

[107] Zeytinoglu, M. and Mintz, M. (1988) Robust fixed size confidence procedures for a restricted parameter space, *Annals of Statistics*, **16**, 1241–1253. MR959199

[108] Zhang, T. and Woodroofe, M. (2003). Credible confidence sets for restricted parameter spaces, *Journal of Statistical Planning and Inference*, **115**, 479–490. MR1985880

[109] Zinzius, E. (1981). *Minimaxschätzer für den Mittelwert θ einer normalverteilen Zufallsgröß e mit bekannter Varianz bei vorgegebener oberer und unterer Schranke für θ, Matheamatische Operationsforschung Statistik, Series Statistics*, **12**, 551–557. MR639251

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 45–55
© Institute of Mathematical Statistics, 2004

A Rubinesque theory of decision

J. B. Kadane[1], Mark J. Schervish[1], and Teddy Seidenfeld[1]

Carnegie Mellon University

Abstract: We generalize a set of axioms introduced by Rubin (1987) to the case of partial preference. That is, we consider cases in which not all uncertain acts are comparable to each other. We demonstrate some relations between these axioms and a decision theory based on sets of probability/utility pairs. We illustrate by example how comparisons solely between pairs of acts is not sufficient to distinguish between decision makers who base their choices on distinct sets of probability/utility pairs.

1. Introduction

Rubin (1987) presented axioms for rational choice amongst sets of available actions. These axioms generalize those of Von Neumann and Morgenstern (1947) which deal solely with comparisons between pairs of actions. Both of these sets of axioms imply that all actions that are choosable from a given set are equivalent in the sense that the rational agent would be indifferent between choosing amongst them. We weaken the axioms of Rubin (1987) by allowing that the agent might not be able to choose between actions without being indifferent between them.

There are several reasons for allowing noncomparbility (unwillingness to choose without being indifferent) between actions. One simple motivation is a consideration of robustness of decisions to changes in parts of a statistical model. For example, consider an estimation problem with a loss function but several competing models for data and/or parameters. We might be interested in determining which estimators can be rejected in the sense that they do not minimize the expected loss under even a single one of the competing models. The agent may not be indifferent between the estimators that remain without being able to select a best one.

With regard to sets of choices, (Rubin, 1987, p. 49) says "The basic concept is that of a *choice set*. This is a set of actions that will be chosen by decision maker; we do *not* assume the decision maker can select a unique action." Nevertheless, the axioms of Rubin (1987) lead to a unique (up to positive affine transformation) utility that ranks all actions, just as do the axioms of Von Neumann and Morgenstern (1947). The weakening of the axioms that we present here is consistent with a set of utilities combined through a Pareto-style criterion, which we introduce in Section 3.

2. Comparison of axioms

Initially, we consider a nonempty convex collection $\mathcal{A}$ of *acts*. In particular, for every $x_1, x_2 \in \mathcal{A}$ and every $0 < a < 1$, $ax_1 + (1-a)x_2 \in \mathcal{A}$. As such, the set of acts must lie in some part of a space where convex combination makes sense. Typically, we think of acts either as probability distributions over a set $\mathcal{R}$ or as functions from some other set Ω to probability distributions on $\mathcal{R}$. These interpretations make convex combination a very natural operation, but the various axiom systems and the related theorems do not rely on one particular class of interpretations.

[1]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pittsburgh, PA 15213, USA. e-mail: kadane@stat.cmu.edu; mark@stat.cmu.edu; teddy@stat.cmu.edu

Keywords and phrases: partial preference.
AMS 2000 subject classifications: primary 62C05; secondary 62A01.

The classic axioms of Von Neumann and Morgenstern (1947) are the following.

Von Neumann–Morgenstern Axiom 1. There exists a weak order $\preceq$ on $\mathcal{A}$. That is,

- for every $x \in \mathcal{A}$, $x \preceq x$,
- for every $x, y \in \mathcal{A}$, either $x \preceq y$, or $y \preceq x$, or both, and
- for all $x, y, z \in \mathcal{A}$, if $x \preceq y$ and $y \preceq z$, then $x \preceq z$.

In the case in which $x \preceq y$ and $y \preceq x$, then we say $x \sim y$.

Von Neumann–Morgenstern Axiom 2. For all $x, y, z \in \mathcal{A}$, $x \preceq y$ if and only if for all $0 < a \leq 1$ $ax + (1-a)z \preceq ay + (1-a)z$.

Von Neumann–Morgenstern Axiom 2 is the most controversial of the classic axioms. Its appeal stems from the following scenario. Imagine that a coin (independent of everything else in the problem) is flipped with probability a of landing heads. If the coin lands heads, you must choose between x and y, otherwise, you get z. Presumably the choice you would make between x and y would be the same in this setting as it would be if you merely had to choose between x and y without any coin flip. The controversy arises out of the following scenario. The coin flip that determines which of x or z arises from $ax + (1-a)z$ can be different (although with the same probability) from the coin flip that determines which of y or z arises from $ay + (1-a)z$. From a minimax standpoint, the first scenario can lead to a different choice between $ax + (1-a)z$ and $ay + (1-a)z$ than does the second scenario.

Von Neumann–Morgenstern Axiom 3. For all $x, y, z \in \mathcal{A}$, if $x \preceq y \preceq z$, then there exists $0 \leq a \leq 1$ such that $y \sim ax + (1-a)z$.

Von Neumann–Morgenstern Axiom 3 prevents any acts from being worth infinitely more (or infinitesimally less) than other acts. Under these axioms, Von Neumann and Morgenstern (1947) prove that there exists a utility $U : \mathcal{A} \to \mathbb{R}$ satisfying

- for all $x, y \in \mathcal{A}$, $x \preceq y$ if and only if $U(x) \leq U(y)$,
- for all $x, y \in \mathcal{A}$ and $0 < a < 1$, $U(ax + (1-a)y) = aU(x) + (1-a)U(y)$, and
- U is unique up to positive affine transformation.

The axioms of Rubin (1987), which we state next, make use of the convex hull of a set $E \subseteq \mathcal{A}$ which is denoted $H(E)$. Rubin (1987) was particularly concerned with the idea that, when presented with a set E of actions, the agent might insist on randomizing between actions in E rather than selecting an action from E itself. This is why the choice set from E is a subset of $H(E)$.

Rubin Axiom 1. There is a function $C : 2^{\mathcal{A}} \to 2^{\mathcal{A}}$ that satisfies

- for all $E \in 2^{\mathcal{A}}$ $C(E) \subseteq H(E)$, and if E has 1, 2, or 3 elements then $C(E) \neq \emptyset$.

The set C in Rubin Axiom 1 can be thought of as a generalization of the weak order $\preceq$ from Von Neumann–Morgenstern Axiom 1: $x \preceq y$ if and only if $y \in C(\{x, y\})$.

Rubin Axiom 2. For all $T, S \in 2^{\mathcal{A}}$, if $T \subseteq H(S)$ and $H(T) \cap C(S) \neq \emptyset$, then $C(T) = H(T) \cap C(S)$.

Rubin Axiom 2 says that if an act is choosable from a large set, then it remains choosable from any smaller set that contains it.

If $S \subseteq \mathcal{A}$, $x \in \mathcal{A}$, and $0 \leq a \leq 1$, define $aS + (1-a)x = \{ay + (1-a)x : y \in S\}$.

Rubin Axiom 3. For all $S \subseteq \mathcal{A}$ and all $0 < a < 1$, if $C(S) \neq \emptyset$, then $C(aS + (1-a)x) = aC(S) + (1-a)x$.

Rubin Axiom 3 is the obvious analog to Von Neumann–Morgenstern Axiom 2.

Rubin Axiom 4. Let $S \subseteq \mathcal{A}$ and $x \in H(S)$. If, for all $V \subseteq H(S)$, ($x \in V$ and $C(V) \neq \emptyset$) implies $x \in C(V)$, then $x \in C(S)$.

Rubin Axiom 4 says that, if an act is not choosable from S, then it is not choosable from some subset of S.

Rubin Axiom 5. Let $x, y, z \in \mathcal{A}$ be such that $C(\{x, y\}) = \{x\}$ and $C(\{y, z\}) = \{y\}$. Then there exists $0 < a < 1$ such that $\{y, ax+(1-a)z\} \subseteq C(\{y, ax+(1-a)z\})$.

Rubin Axiom 5 is an obvious analog to Von Neumann–Morgenstern Axiom 3. Under these axioms, Rubin (1987) proves that there exists a utility $U : \mathcal{A} \to \mathbb{R}$ satisfying

- for all $E \subseteq \mathcal{A}$, $C(E) = \{x \in H(E) : \text{ for all } y \in E,\ U(x) \geq U(y)\}$,
- for all $x, y \in \mathcal{A}$ and $0 < a < 1$, $U(ax + (1-a)y) = aU(x) + (1-a)U(y)$, and
- U is unique up to positive affine transformation.

It is fairly simple to show that, if such a U exists, then all of Rubin's axioms hold. Hence, his result is that his axioms characterize choice sets that are related to utility functions in the way described by the three bullets above.

In order to allow noncomparability, we need more general axioms than Von Neumann–Morgenstern Axiom 3 and Rubin Axiom 5. To state the more general axioms, we need a topology on the set of actions. For now, assume that the set of acts $\mathcal{A}$ is a metric space with some metric d. When we consider specific examples, we will construct the metric. Let $\mathcal{F}$ be the collection of nonempty closed subsets of $\mathcal{A}$.

We prefer to state our axioms in terms of a rejection function rather than a choice set function.

Definition 1. A *rejection set* R is a function $R : \mathcal{F} \to 2^{\mathcal{A}}$ such that, for all $E \in \mathcal{F}$, $R(E) \subseteq E$, and $R(E) \neq E$.

Axiom 1. If $B \subseteq R(A)$ and if $A \subseteq D$, then $B \subseteq R(D)$.

Axiom 1 is the same as Sen's property α. (see Sen (1977)). It says that adding more options to a set of acts doesn't make the rejected ones become acceptable.

Axiom 2. If B is a subset of $R(A)$ and if D is a subset of B, then $B \backslash D \subseteq R(\overline{A \setminus D})$.

Axiom 2 says that rejected acts remain rejected even if we remove other rejected acts from the option set.

Definition 2. For $A, B \in \mathcal{F}$, say that $A \prec B$ if $A \subseteq R(A \cup B)$.

Lemma 1. *Assume Axiom 1 and Axiom 2. Then $\prec$ is a strict partial order on $\mathcal{F}$.*

Proof. Let $A, B \in \mathcal{F}$. If $A \prec B$, then $B \not\prec A$ because A and B being closed implies that $R(A \cup B) \neq A \cup B$. For transitivity, assume that $A \prec B$ and $B \prec D$ with $A, B, D \in \mathcal{F}$. Then $A \subseteq R(A \cup B) \subseteq R(A \cup B \cup D)$, by Axiom 1. Also Axiom 1 says that $B \subseteq R(B \cup D) \subseteq R(A \cup B \cup D)$. It follows that $A \cup B \subseteq R(A \cup B \cup D)$. Let $E = B \setminus A$. Then

$$A = (A \cup B) \setminus E \subseteq R(\overline{[A \cup B \cup D] \setminus E}) \subseteq R(A \cup D),$$

where the first inclusion is from Axiom 2 and the second is from Axiom 1. □

Our next axiom is similar to Rubin Axiom 3.

Axiom 3. For all $E \in \mathcal{F}$, all $x \in \mathcal{A}$, and all $0 < a \leq 1$, $B = R(E)$ if and only if $aB + (1 - a)x = R(aE + (1 - a)x)$.

For the continuity axiom, we require the concept of a sequence of sets that are all indexed the same way.

Definition 3. Let G be an index set with cardinality less than that of $\mathcal{A}$. Let H be another index set. Let $\mathcal{H} = \{E_h : h \in H\}$ be a collection of subsets of $\mathcal{A}$. We say that the sets in $\mathcal{H}$ are *indexed in common by* G if For each $h \in H$ and each $g \in G$, there exists $x_{h,g} \in \mathcal{A}$ such that $E_h = \{x_{h,g} : g \in G\}$.

Axiom 4. Let G_A and G_B be index sets with cardinalities less than that of $\mathcal{A}$. Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of elements of $\mathcal{F}$ such that each $A_n = \{x_{n,g} : g \in G_A\}$. Also, let $\{B_n\}_{n=1}^{\infty}$ be a sequence of elements of $\mathcal{F}$ such that each $B_n = \{x_{n,g} : g \in G_B\}$. Suppose that for each $g \in G_A \cup G_B$, $x_{n,g} \Rightarrow x_g \in \mathcal{A}$. Let $A = \{x_g : g \in G_A\}$ and $B = \{x_g : g \in G_B\}$. Let N and J be closed subsets of $\mathcal{A}$.

- If $\forall n B_n \prec A_n$ and $\overline{A} \prec N$, then $\overline{B} \prec N$.
- If $\forall n B_n \prec A_n$ and $J \prec \overline{B}$, then $J \prec \overline{A}$.

The reason for wording Axiom 4 with the additional sets N and J is that the acts in $\overline{B}$ and $\overline{A}$ might be noncomparable when compared to each other because the limit process brings B_n and A_n so close together. But the axiom says that the limit of a sequence of rejected options can't jump over something that is better than the limit of choosable options. Similarly, the limit of a sequence of choosable options can't jump below something that is worse than the limit of rejected options.

We state one additional axiom here that is necessary for the generalization that we hope to achieve. Recall that $H(E)$ is the convex hull of the set E.

Axiom 5. For each $E \in \mathcal{F}$ and $B \subseteq E$, if $B \subseteq R(\overline{H(E)})$, then $B \subseteq R(E)$.

Axiom 5 says that if acts are rejected when the closed convex hull of E is available then they must also be rejected when E alone is available. Closing the convex hull of a closed set of acts should not allow us to reject any acts that we couldn't reject before.

3. Pareto Criteria

After proving the existence of the utility, Rubin (1987) considers cases with many utilities indexed by elements of some set Ω. He then says (p. 53) "Two immediate examples come to mind: Ω may be the class of states of nature, or Ω may be the set of all individuals in a population. Suppose we assume that the choice process

given ω is 'reasonable' for each $\omega \in \Omega$, and the overall process is also reasonable." The first of the two examples envisioned by Rubin (1987) is the usual case in which there is uncertainty about unknown events. The second example is the case in which the "overall process" is governed by a social welfare function. Our approach is motivated by an alternative way of thinking about individuals in a population. Instead of a social welfare function that performs just like an individual's utility, we seek a characterization of the agreements amongst the individuals.

Definition 4. Let $\aleph$ be a set. For each $\alpha \in \aleph$, let $R_\alpha : \mathcal{F} \to 2^{\mathcal{A}}$ be a rejection function. The *Pareto rejection function related to* $\{R_\alpha : \alpha \in \aleph\}$ is $R(E) = \bigcap_{\alpha\in\aleph} R_\alpha(E)$ for all $E \in \mathcal{F}$.

In this definition, an act x is *Pareto rejected* by the group $\aleph$ if it is rejected by every member of the group. The complement of the Pareto rejection function might be called the Pareto choice function $C : \mathcal{F} \to 2^{\mathcal{A}}$ defined by $C(E) = [R(E)]^C$. This is the set of acts that fail to be rejected by at least one individual in $\aleph$.

The general example that motivates our work is the following. Let Ω be a finite set of states. For each $\alpha \in \aleph$ let P_α be a probability on Ω. Let acts be functions from Ω to probability measures over some finite set of prizes $\mathcal{R}$. That is, let $\mathcal{P}_\mathcal{R}$ be the set of probability measures over $\mathcal{R}$ so that each act $x \in \mathcal{A}$ is a function $x : \Omega \to \mathcal{P}_\mathcal{R}$ and $x(\omega)(r)$ is the probability of prize r in state ω. For each $\alpha \in \aleph$, let there be a bounded possibly state-dependent utility $U_\alpha(\cdot|\omega)$, appropriately measurable. Define $V_\alpha : \mathcal{A} \to \mathbb{R}$ by

$$V_\alpha(x) = \sum_{\omega\in\Omega} \left[\sum_{r\in\mathcal{R}} U_\alpha(r|\omega)x(\omega)(r) \right] P_\alpha(\omega).$$

Next, define

$$C_\alpha(E) = \big\{x \in E : V_\alpha(x) \geq V_\alpha(y), \text{ for all } y \in E\big\},$$

and $R_\alpha(E) = [C_\alpha(E)]^C$. Hence, C_α is the set of all Bayes rules in the model with utility U_α and probability P_α. Then $R(E) = \bigcap_{\alpha\in\aleph} R_\alpha(E)$ is the set of all acts x such that, for every model in $\aleph$, x fails to be a Bayes rule. We call this rejection function the *Bayes rejection function related to* $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$. Finally, we define the metric on $\mathcal{A}$. All acts are equivalent to points in a bounded subset of a finite-dimensional Euclidean space. If s is the number of states and t is the number of prizes, then act x is equivalent to an $s \times t$ matrix with (i, j) entry equal to the probability of prize j in state i. We will use the usual Euclidean metric as d. It is now easy to see that V_α is a continuous function of x for each α.

Lemma 2. *If $B \prec A$, then, for each $\alpha \in \aleph$, there is $y \in A \setminus B$ such that $V_\alpha(y) > \sup_{z\in B} V_\alpha(z)$.*

Proof. We can think of B as a closed and bounded subset of a finite-dimensional Euclidean space. For each $\alpha \in \aleph$, V_α is continuous, hence there exists $x \in B$ such that $V_\alpha(x) = \sup_{z\in B} V_\alpha(z)$. Since $x \in B$, there exists $y \in A \cup B$ such that $V_\alpha(x) < V(y)$. By the definition of x it is clear that $y \in A \setminus B$. □

Lemma 3. *The Bayes rejection function related to $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$ satisfies Axiom 1.*

Proof. Let $A \in \mathcal{F}$ and $B \subseteq R(A)$ and $A \subseteq D$. If $x \in B$, then for each $\alpha \in \aleph$, there is $y_\alpha \in A$ such that $V_\alpha(x) < V_\alpha(y_\alpha)$. Since $y_\alpha \in D$ for all α, it follows that $x \in R(D)$. □

Lemma 4. *The Bayes rejection function related to* $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$ *satisfies Axiom* 2.

Proof. Let B be a closed subset of $R(A)$ and let $D \subseteq B$. Let $x \in B \setminus D$. Since $x \in B$, for every $\alpha \in \aleph$, there exists $y_\alpha \in A \setminus B$ such that $V_\alpha(x) < V_\alpha(y_\alpha)$ by Lemma 2. Since $y_\alpha \in A \setminus D$ as well, we have $x \in R(\overline{A \setminus D})$. □

Lemma 5. *The Bayes rejection function related to* $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$ *satisfies Axiom* 3.

Proof. The "if" direction is trivial because $a = 1$ is included. For the "only if" direction, let $0 < a \leq 1$, $x \in \mathcal{A}$ and $E \in \mathcal{F}$. First, we show that $R(aE + (1-a)x) \subseteq aR(E) + (1-a)x$. Let $z \in R(aE + (1-a)x)$. Express $z = ay + (1-a)x$, with $y \in E$. For every $\alpha \in \aleph$ there is $z_\alpha = ay_\alpha + (1-a)x$ with $y_\alpha \in E$ and $V_\alpha(z_\alpha) > V_\alpha(z)$. This implies $V_\alpha(y_\alpha) > V_\alpha(y)$ and $y \in R(E)$, so $z \in aR(E) + (1-a)x$. Finally, let $z \in aR(E) + (1-a)x$, and express $z = ay + (1-a)x$, with $y \in R(E)$. For every $\alpha \in \aleph$, there is $y_\alpha \in E$ such that $V_\alpha(y_\alpha) > V_\alpha(y)$ so that $V_\alpha(ay + (1-a)x) > V_\alpha(z)$. It follows that $z \in R(aE + (1-a)x)$. □

Lemma 6. *The Bayes rejection function related to* $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$ *satisfies Axiom* 4.

Proof. Assume $B_n \prec A_n$ for all n. Let $g \in G_B$ and $\alpha \in \aleph$. For each n, there is $h_{n,g} \in G_A$ such that $V_\alpha(x_{n,g}) < V_\alpha(x_{n,h_{n,g}})$. By continuity of V_α, we have

$$V_\alpha(x_g) \leq \liminf_n V_\alpha(x_{n,h_{n,g}}) \leq \sup_{h \in G_A} V_\alpha(x_h) \leq \sup_{x \in \overline{A}} V_\alpha(x).$$

Because V_α is continuous and $\overline{A}$ is a closed and bounded subset of a finite-dimensional Euclidean space, there exists $y \in \overline{A}$ such that $V_\alpha(y) = \sup_{x \in \overline{A}} V_\alpha(x)$. It follows that

$$\sup_{g \in G_B} V_\alpha(x_g) \leq V_\alpha(y). \tag{1}$$

For the first line of Axiom 4, assume that $\overline{A} \prec N$. For each $g \in G_B$ and each $\alpha \in \aleph$, we need to find $z \in \overline{B} \cup N$ such that $V_\alpha(x_g) < V_\alpha(z)$. Let y be as in (1). Because $\overline{A} \prec N$, there is $z \in N \setminus \overline{A} \subseteq \overline{B} \cup N$ such that $V_\alpha(z) > V_\alpha(y) \geq V_\alpha(x_g)$.

For the second line of Axiom 4, assume that $J \prec \overline{B}$. For each $x \in J$ and each $\alpha \in \aleph$, we need to find $y \in J \cup \overline{A}$ such that $V_\alpha(x) < V_\alpha(y)$. Let $x \in J$ and $\alpha \in \aleph$. By Lemma 2 there is $x_g \in \overline{B} \setminus J$ such that $V_\alpha(x) < V_\alpha(x_g)$. Let y be as in (1). Since $y \in \overline{A} \subseteq J \cup \overline{A}$, we are done. □

Lemma 7. *The Bayes rejection function related to* $\{(P_\alpha, U_\alpha) : \alpha \in \aleph\}$ *satisfies Axiom* 5.

Proof. Let $E \in \mathcal{F}$ and $B \subseteq E$. Assume that $B \subseteq R(\overline{H(E)})$. Let $x \in B$. For each $\alpha \in \aleph$, we know that there exists $z_\alpha \in \overline{H(E)}$ such that $V_\alpha(x) < V_\alpha(z_\alpha)$. This z_α is a limit of elements of $H(E)$ and V_α is continuous, hence there is a $w_\alpha \in H(E)$ such that $V_\alpha(x) < V_\alpha(w_\alpha)$. This w_α is a convex combination of elements of E, $w_\alpha = \sum_{i=1}^{\ell} a_i w_{i,\alpha}$ with $w_{i,\alpha} \in E$ and $\sum_{i=1}^{\ell} a_i = 1$ with all $a_i \geq 0$. Since

$$V_\alpha(w_\alpha) = \sum_{i=1}^{\ell} a_i V_\alpha(w_{i,\alpha}) > V_\alpha(x),$$

there must exist i such that $V_\alpha(w_{i,\alpha}) > V_\alpha(x)$. Let $y_\alpha = w_{i,\alpha}$. □

What the preceding results establish is that the Bayes rejection function related to a collection of probability/utility pairs satisfies our axioms. We would like to consider the opposite implication, that is, whether or not every rejection function that satisfies our axioms is the Bayes rejection function related to some collection of probability utility pairs. This consideration will be postponed until another paper.

4. Pairwise choice is not enough

Seidenfeld, Schervish and Kadane (1995) consider the first four axioms that we have introduced in this paper but restricted to the collection of subsets of the form $\{x, y\}$ with $x, y \in \mathcal{A}$. That is, Seidenfeld, Schervish and Kadane (1995) consider choices between pairs of acts only. They go on to prove that, under these axioms, there exists a collection of bounded utilities $\{V_\alpha : \alpha \in \aleph\}$ that agree with all pairwise choices in the following sense: $\{x\} \prec \{y\}$ if and only if $V_\alpha(x) < V_\alpha(y)$ for all $\alpha \in \aleph$. The following example illustrates why Axiom 5 is necessary in the case of choice between more than two acts at a time.

Example 1. Let $\mathcal{A} = \{(a, b) : 0 \leq a, b \leq 1\}$. Define the rejection function R as follows. For $(a, b) \in E$, $(a, b) \in R(E)$ if and only if there exists $(c, d) \in E$ such that, for every $0 \leq p \leq 1$, $ap + b(1-p) < cp + d(1-p)$. It is not difficult to show that this rejection function satisfies our first four axioms. However, there is no set of utility functions for which this rejection function is the Pareto rejection function. Suppose that U were an element of such a set of utility functions. By Axiom 3, $U(a, b)$ would have to equal $aU(1, 0) + bU(0, 1)$, hence

$$U(0.4, 0.4) = 0.4[U(1, 0) + U(0, 1)] < \max\{U(1, 0), U(0, 1)\}.$$

Hence either $U(1, 0) > U(0.4, 0.4)$ or $U(0, 1) > U(0.4, 0.4)$. Now, let $E = \{(0.4, 0.4), (1, 0), (0, 1)\}$, and notice that $R(E) = \emptyset$. But every utility function U would reject $(0.4, 0.4)$ amongst the actions in E.

The rejection function in Example 1 is an example of "Maximality" that was introduced by Walley (1990). The distinction between pairwise choice and larger choice sets goes beyond the situation of Example 1. Schervish, Seidenfeld, Kadane and Levi (2003) look more carefully at the special case of Bayes rejection functions in which all U_α are the same function U and $\{P_\alpha : \alpha \in \aleph\} = \mathcal{P}$, is a convex set of probabilities on Ω. We call this the case of a *cooperative team*. In this case, they give an example that illustrates how different sets $\mathcal{P}$ lead to the same collections of pairwise choices that satisfy the axioms of Seidenfeld, Schervish and Kadane (1995). Hence, pairwise choices are not sufficient for characterizing the corresponding set of probability/utility pairs even in the cases in which such sets of probability/utility pairs are known to exist.

Example 2. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ consist of three states. Let

$$\begin{aligned}
\mathcal{P}_1 &= \big\{(p_1, p_2, p_3) : p_2 < 2p_1 \quad \text{for } p_1 \leq 0.2\big\} \\
&\quad \bigcup \big\{(p_1, p_2, p_3) : p_2 \leq 2p_1 \quad \text{for } 0.2 < p_1 \leq 1/3\big\}, \\
\mathcal{P}_2 &= \big\{(p_1, p_2, p_3) : p_2 < 2p_1 \quad \text{for } p_1 < 0.2\big\} \\
&\quad \bigcup \big\{(p_1, p_2, p_3) : p_2 \leq 2p_1 \quad \text{for } 0.2 \leq p_1 \leq 1/3\big\}.
\end{aligned}$$

The only difference between the two sets is that $(0.2, 0.4, 0.4) \in \mathcal{P}_2 \setminus \mathcal{P}_1$. Let $R_{\mathcal{P}_1}$ and $R_{\mathcal{P}_2}$ be the Bayes rejection functions corresponding to the two sets of probability/utility pairs $\{(p, U) : p \in \mathcal{P}_1\}$ and $\{(p, U) : p \in \mathcal{P}_2\}$. Each act x can

be represented by the vector whose ith coordinate is $x_i = \sum_{r\in R} U(r|\omega_i)x(\omega_i)(r)$ for $i = 1, 2, 3$. In this way, the expected utility for each probability vector p is $V_p(x) = x^\top p$. Consider two arbitrary acts x and y. We have $\{x\} \in R_{\mathcal{P}_j}(\{x, y\})$ if and only if

$$\sum_{i=1}^{3}(y_i - x_i)p_i > 0, \qquad \text{for all } p \in \mathcal{P}_j.$$

This is equivalent to $(y_1 - x_1, y_2 - x_2, y_3 - x_3)$ being a hyperplane that separates $\{0\}$ from $\mathcal{P}_j$ without intersecting $\mathcal{P}_j$. It is easy to check that a hyperplane separates $\mathcal{P}_1$ from $\{0\}$ without intersecting $\mathcal{P}_1$ if and only if it separates $\mathcal{P}_2$ from $\{0\}$ without intersecting $\mathcal{P}_2$. The reason is that all of the points in the symmetric difference $\mathcal{P}_1 \Delta \mathcal{P}_2$ are extreme but not exposed. Hence, all pairwise comparisons derived from $R_{\mathcal{P}_1}$ are identical to those derived from $R_{\mathcal{P}_2}$.

Consider now a set of acts E that contains only the following three acts (each expressed as a vector of its expected payoffs in the three states as were x and y above):

$$\begin{aligned} f_1 &= (0.2, 0.2, 0.2), \\ f_2 &= (1, 0, 0), \\ g &= (-1.8, 1.2, .2). \end{aligned}$$

First, let $p \in \mathcal{P}_1$. Notice that $V_p(f_2)$ is the highest of the three whenever $p_1 \geq 0.2$, $V_p(f_1)$ is the highest whenever $p_1 \leq 0.2$, and $V_p(g)$ is never the highest. So, $R_{\mathcal{P}_1}(E) = \{g\}$. Next, notice that if $p = (0.2, 0.4, 0.4)$, then $V_p(g) = V_p(f_1) = V_p(f_2) = 0.2$, so $R_{\mathcal{P}_2}(E) = \emptyset$. $\Diamond$

Next, we present a theorem which states that the more general framework of rejection functions operating on sets of size larger than 2 can distinguish between different sets $\mathcal{P}$ in the cooperative team case.

For the general case, let U be a single, possibly state-dependent, utility function. For each probability vector p on Ω and each act x, let

$$V_p(x) = \sum_{\omega\in\Omega} \left[\sum_{r\in R} U(r|\omega)x(\omega)(r)\right] p(\omega).$$

Because the inner sum $w_x(\omega) = \sum_{r\in R} U(r|\omega)x(\omega)(r)$ does not depend on p, we can represent each act x by the vector

$$\big(w_x(\omega_1), \ldots, w_x(\omega_s)\big), \tag{2}$$

where s is the number of states. That is, each act might as well be the vector in (2) giving for each state the state-dependent expected utility with respect to its probability distribution over prizes in that state. If we call the vector in (2) by the name x, this makes $V_p(x) = x^\top p$ for every act x and every probability vector p.

For each convex set $\mathcal{P}$ of probability vectors there is a Bayes rejection function defined by

$$R_{\mathcal{P}}(E) = \bigcap_{p\in\mathcal{P}} \big\{x \in E : V_p(x) \geq V_p(y), \text{ for all } y \in E\big\}^C, \tag{3}$$

for all closed sets E of acts. Example 2 shows that there are cases in which $\mathcal{P}_1 \neq \mathcal{P}_2$ but $R_{\mathcal{P}_1}(E) = R_{\mathcal{P}_2}(E)$ for every E that contains exactly two distinct acts. Theorem 1 below states that, so long as $\mathcal{P}_1 \neq \mathcal{P}_2$ there exists a finite set E of acts such that $R_{\mathcal{P}_1}(E) \neq R_{\mathcal{P}_2}(E)$.

Theorem 1. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be distinct convex sets of probabilities over a set Ω with $s \geq 2$ states. Then there is a set E with with at most $s+1$ acts such that $R_{\mathcal{P}_1}(E) \neq R_{\mathcal{P}_2}(E)$.*

The proof of Theorem 1 is given in the appendix.

5. Summary

In this paper, we consider a generalization of Subjective Expected Utility theory in which not all options are comparable by a binary preference relation. We adapt Rubin's (1987) axioms for rational choice functions to permit a decision maker who has a determinate cardinal utility U for outcomes to have a choice function over simple horse-lottery options that does not coincide with a weak ordering of the option space. In calling the decision maker's choice function "rational", we mean that there is a cardinal utility U and a set $\mathcal{P}$ of coherent probabilities that represent the choice function in the following sense: The allowed choices from an option set are exactly those Bayes-admissible options, i.e. those options that maximize expected utility for some probability $P \in \mathcal{P}$.

In Sections 2 and 3 we give axioms that are necessary for a choice function to be rational in this sense. We show that the axioms that we used in Schervish, Seidenfeld, Kadane and Levi (1995) for a theory of coherent strict partial orders are insufficient for this purpose. Specifically, those axioms are for a strict partial order $\prec$ which is given by pairwise comparisons solely. That theory represents the strict partial order $\prec$ by a set of probability/utility pairs according to a Pareto condition, where each probability/utility pair agrees with the strict partial order according expected utility inequalities. Here we show that the choice function that Walley calls "Maximality" obeys those axioms, but fails to have the desired representation in terms of Bayes-admissible options when the option sets (which may fail to be convex) involve three or more options. Therefore, we add a new Axiom 5 that is necessary for a choice function to be rational, and which is not satisfied by Maximality.

In Section 4 we show that, even when a rational choice function is represented by a convex set of coherent probabilities, and when the option set also is convex, nonetheless the choice function cannot always be reduced to pairwise comparisons. We show how to distinguish the choice functions based on any two different convex sets of probabilities using choice problems that go beyond pairwise comparisons.

In continuing work, we seek a set of axioms that characterize all rational choice functions. The axioms that we offer in Section 2 are currently a candidate for that theory.

A. Proof of Theorem 1

First, we present a few lemmas about convex sets that will be useful for the proof.

The following result gives us a way of reexpressing a half-space of a hyperplane as the intersection of the hyperplane with a half-space of a more convenient form. The main point is that the same constant c that defines the original hyperplane H can also be used to define the new half-space.

Lemma 8. *Let $H = \{x \in \mathbb{R}^n : \beta^\top x = c\}$ for some vector β and some scalar $c \neq 0$. Let α be such that $\beta^\top \alpha = 0$ and let d be a scalar. Then, there is a vector γ such that*

$$\{x \in H : \alpha^\top x \geq d\} = \{x \in H : \gamma^\top x \geq c\}.$$

Proof. It is easy to check that the following vector does the job

$$\gamma = \begin{cases} c\alpha/d & \text{if } cd > 0, \\ \alpha + \beta & \text{if } d = 0, \\ -c\alpha/d + 2\beta & \text{if } cd < 0. \end{cases}$$ □

Definition 5. We say that two convex sets $\mathcal{P}_1$ and $\mathcal{P}_2$ *intersect all of the same supporting hyperplanes* if

- they have the same closure, and
- for every supporting hyperplane H, $H \cap \mathcal{P}_1 \neq \emptyset$ if and only if $H \cap \mathcal{P}_2 \neq \emptyset$.

Definition 6. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be convex sets in $\mathbb{R}^n$. For $i = 1, 2$, define $R_{\mathcal{P}_i}$ as in (3). Let E be a subset of $\mathbb{R}^n$. We say that E *distinguishes* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ if $R_{\mathcal{P}_1}(E) \neq R_{\mathcal{P}_2}(E)$.

We break the proof of Theorem 1 into two parts according to whether or not $\mathcal{P}_1$ and $\mathcal{P}_2$ intersect all of the same supporting hyperplanes. The first part deals with cases in which a single pair of acts can distinguish two convex sets.

Lemma 9. *Suppose that two convex sets $\mathcal{P}_1$ and $\mathcal{P}_2$ do not intersect all of the same supporting hyperplanes. Then there is a set E with one constant act and one possibly nonconstant act that distinguishes $\mathcal{P}_1$ and $\mathcal{P}_2$.*

Proof. First, consider the case in which $\mathcal{P}_1$ and $\mathcal{P}_2$ don't have the same closure. Without loss of generality, let $p_0 \in \mathcal{P}_2 \cap \overline{\mathcal{P}_1}^C$. Let $x \in \mathbb{R}^n$ and c be such that $x^\top p > c$ for all $p \in \overline{\mathcal{P}_1}$ and $x^\top p_0 < c$. Let E consist of the two acts x and the constant $\boldsymbol{c} = (c, \ldots, c)$. Clearly, $\{\boldsymbol{c}\} = R_{\mathcal{P}_1}(E)$ while $\boldsymbol{c} \notin R_{\mathcal{P}_2}(E)$.

Next, consider the case in which $\mathcal{P}_1$ and $\mathcal{P}_2$ have the same closure. Without loss of generality, let $\{p : x^\top p = c\}$ be a supporting hyperplane that intersects $\mathcal{P}_2$ but not $\mathcal{P}_1$ so that $x^\top p > c$ for all $p \in \mathcal{P}_1$. Let $E = \{\boldsymbol{c}, x\}$. Then $\{\boldsymbol{c}\} = R_{\mathcal{P}_1}(E)$ while $\boldsymbol{c} \notin R_{\mathcal{P}_2}(E)$. □

The following result handles the case in which pairwise choice is not sufficient to distinguish two sets. The proof can be summarized as follows. Start with two distinct convex sets of probabilities that intersect all of the same supporting hyperplanes. Find a supporting hyperplane that they intersect in different ways and use this as the first gamble in the set E in such a way that all probabilities in the hyperplane give the gamble the same expected value (say c) and the rest of both convex sets give the gamble smaller expected value. Put the constant $\boldsymbol{c}$ into E as well. Now, the only probabilities that keep the first gamble out of the rejection set are in the hyperplane. We now add further gambles to E in a sequence such that the next one has expected value greater than c except on a boundary of one less dimension than the previous one. By so doing, we reduce the set of probabilities that keep the first gamble out of the rejection set by decreasing its dimension by one each time. Eventually, we get the set of such probabilities to a zero-dimensional set (a single point) that lies in one of the two original convex sets but not the other.

Lemma 10. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be distinct convex sets of probabilities in $\mathbb{R}^s$ ($s \geq 2$) that intersect all of the same supporting hyperplanes. Then there is a set E with at most $s + 1$ gambles that distinguishes $\mathcal{P}_1$ and $\mathcal{P}_2$.*

Proof. Clearly the difference between $\mathcal{P}_1$ and $\mathcal{P}_2$ is all on the common boundary. Hence, there is some supporting hyperplane that intersects both sets but in different ways. Let such a hyperplane be $H_1 = \{p : x_1^\top p = c\}$ such that for all $p \in \mathcal{P}_i$, $x_1^\top p \leq c$ for $i = 1, 2$. Let $\mathcal{P}_{i,1} = \mathcal{P}_i \cap H_1$ for $i = 1, 2$. Let the first two gambles in E be x_1 and $\boldsymbol{c}$. (If $c = 0$, add a constant to every coordinate of x_1 and replace c by that constant.) The remainder of the proof proceeds through at most $s - 1$ additional steps of the type to follow where one new gamble gets added to E at each step. Initialize $j = 1$.

By construction, $\mathcal{P}_{1,j}$ and $\mathcal{P}_{2,j}$ are distinct convex sets that lie in an $s - j$ dimensional hyperplane. If these sets intersect all of the same supporting hyperplanes (case 1), then find a supporting subhyperplane H'_{j+1} of H_j that intersects $\mathcal{P}_{1,j}$ and $\mathcal{P}_{2,j}$ in different ways. If the sets $\mathcal{P}_{1,j}$ and $\mathcal{P}_{2,j}$ don't intersect all of the same supporting hyperplanes (case 2), use Lemma 9 to find a subhyperplane H'_{j+1} of H_j that distinguishes them. In either case, use Lemma 8 to extend H'_{j+1} to $H_{j+1} = \{p : x_j^\top p = c\}$ such that $x_j^\top p \geq c$ for all $p \in \mathcal{P}_{i,j}$ for both $i = 1, 2$. Include x_j in E. Define $\mathcal{P}_{i,j+1} = \mathcal{P}_{i,j} \cap H_{j+1}$ for $i = 1, 2$.

If case 2 holds in the previous paragraph, skip to the next paragraph. If case 1 holds in the previous paragraph, then increment j to $j+1$ and repeat the construction in the previous paragraph. Continue in this way either until case 2 holds or we arrive at $j = s - 1$ with one-dimensional sets $\mathcal{P}_{1,s-1}$ and $\mathcal{P}_{2,s-1}$, which then must be bounded line segments. They differ by at least one of them containing a point that the other does not contain. Without loss of generality, suppose that $\mathcal{P}_{2,s-1}$ contains a point p_0 that is not in $\mathcal{P}_{1,s-1}$. Create one last vector x_s so that $x_s^T p_0 = c$ and $x_s^\top p > c$ for all $p \in \mathcal{P}_{1,s-1}$.

Every gamble $x \in E$ satisfies $x^\top p_0 = c$, while for every $p \in \mathcal{P}_1$, there is $k \geq 2$ such that $x_k^\top p > c$. It now follows that $x_1 \in R_{\mathcal{P}_1}(E)$ but $x_1 \notin R_{\mathcal{P}_2}(E)$. □

References

Rubin, H. (1987). A weak system of axioms for "rational" behavior and the non-separability of utility from prior. *Statistics & Decisions*, **5**, 47–58.

Schervish, M. J., Seidenfeld, T., Kadane, J. B., and Levi, I. (2003). Extentions of expected utility theory and some limitations of pairwise comparisons. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications* (J.-M. Bernard, T. Seidenfeld and M. Zaffalon, eds.) 496–510. Carlelton Scientific.

Seidenfeld, T., Schervish, M. J., and Kadane, J. (1995). A representation of partially ordered preferences. *Annals of Statistics*, **23**, 2168–2217. MR886877

Sen, A. (1977). Social choice theory: A re-examination. *Econometrica*, **45**, 53–89. MR456612

Von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior* (2nd ed.) Princeton University Press, Princeton. MR21298

Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London. MR1145491

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 56–61
© Institute of Mathematical Statistics, 2004

On the distribution of the greatest common divisor

Persi Diaconis[1] **and Paul Erdös**[1]

Stanford University

Abstract: For two integers chosen independently at random from $\{1, 2, \ldots, x\}$, we give expansions for the distribution and the moments of their greatest common divisor and the least common multiple, with explicit error rates. The expansion involves Riemann's zeta function. Application to a statistical question is briefly discussed.

1. Introduction and statement of main results

Let M and N be random intergers chosen uniformly and independently from $\{1, 2, \ldots, x\}$. Throughout (M, N) will denote the greatest common divisor and $[M, N]$ the least common multiple. Cesàro (1885) studied the moments of (M, N) and $[M, N]$. Theorems 1 and 2 extend his work by providing explicit error terms. The distribution of (M, N) and $[M, N]$ is given by:

Theorem 1.

$$P_x\big\{[M,N] \le tx^2 \quad \textit{and} \quad (M,N) = k\big\}$$
$$= \frac{6}{\pi^2}\ \frac{1}{k^2}\big\{kt(1-\log kt)\big\} + O_{k,t}\left(\frac{\log x}{x}\right) \tag{1.1}$$

$$P_x\big\{(M,N) = k\big\} = \frac{6}{\pi^2}\frac{1}{k^2} + O\left(\frac{\log(\frac{x}{k})}{xk}\right) \tag{1.2}$$

$$P_x\big\{[M,N] \le tx^2\big\} = 1 + \frac{6}{\pi^2}\cdot\sum_{j=1}^{[1/t]}\big\{jt(1-\log jt) - 1\big\} + O_t\left(\frac{\log x}{x}\right). \tag{1.3}$$

Where $[x]$ denotes the greatest integer less than or equal to x. Christopher (1956) gave a weaker form of (1.2).

(1.2) easily yields an estimate for the expected value of (M, N):

$$E_x\big\{(M,N)\big\} = \frac{1}{x^2}\sum_{i,j\le x}(i,j) = \sum_{k\le x} k\ P_x\big\{(M,N)=k\big\} = \frac{6}{\pi^2}\log x + O(1).$$

(1.2) does not lead to an estimate for higher moments of (M, N). Similarly the form of (1.3) makes direct computation of moments of $[M, N]$ unwieldy. Using elementary arguments we will show:

Theorem 2.

$$E_x\big\{(M,N)\big\} = \frac{6}{\pi^2}\log x + C + O\left(\frac{\log x}{\sqrt{x}}\right) \tag{1.4}$$

[1]Department of Statistics, Stanford University, Stanford, 94305-4065, CA USA. e-mail: diaconis@math.stanford.edu

Keywords and phrases: Euler constant, gcd, inversion, lcm, moment, random, zeta function.
AMS 2000 subject classifications: 11N37, 11A25, 60E05.

where C is an explicitly calculated constant.

$$\text{for} \qquad k \geq 2,\ E_x\big\{(M,N)^k\big\} = \frac{x^{k-1}}{k+1}\left\{\frac{2\zeta(k)}{\zeta(k+1)} - 1\right\} + O\big(x^{k-2}\log x\big). \tag{1.5}$$

where $\zeta(z)$ is Riemann's zeta function,

$$\text{for} \qquad k \geq 1, E_x\big\{[M,N]^k\big\} = \frac{\zeta(k+2)}{\zeta(2)(k+1)^2}x^{2k} + O\big(x^{2k-1}\log x\big). \tag{1.6}$$

Section two of this paper contains proofs while section three contains remarks, further references and an application to the statistical problem of reconstructing the sample size given a table of rounded percentages.

2. Proofs of main theorems

Throughout we use the elementary estimate

$$\Phi(x) = \sum_{1\leq k\leq x}\varphi(k) = \frac{3}{\pi^2}x^2 + R(x) \tag{2.1}$$

where $R(x) = O(x\log x)$.

See, for example, Hardy and Wright (1960) Theorem 330. Since $\#\ \{m,n \leq x : (m,n) = 1\} = 2\Phi(x) + O(1)$ and $(m,n) = k$ if and only if $k|m$, $k|n$ and $(\frac{m}{k},\frac{n}{k}) = 1$, we see that $\#\ \{m,n \leq x : (m,n) = k\} = 2\Phi(\frac{x}{k}) + O(1)$. This proves (1.2). To prove (1.1) and (1.3) we need a preparatory lemma.

Lemma 1. *If $F_x(t) = \#\{m,n \leq x : mn \leq tx^2$ and $(m,n) = 1\}$, then*

$$F_x(t) = \frac{6}{\pi^2}\ t(1-\log t)x^2 + O_t(x\log x).$$

Proof. Consider the number of lattice points in the region $R_x(t) = \{m,n \leq x : mn \leq tx^2\}$. It is easy to see that there are $t(1-\log t)x^2 + O_t(x) = N_x(t)$ such points. Also, the pair $\langle m,n\rangle \in R_x(t)$ and $(m,n) = k$ if and only if $\langle\frac{m}{k},\frac{n}{k}\rangle \in R_{x/k}(t)$ and $(\frac{m}{k},\frac{n}{k}) = 1$. Thus $N_x(t) = \sum_{1\leq d\leq x}F_{x/d}(t)$. The standard inversion formula says

$$F_x(t) = \sum_{1\leq d\leq x}\mu(d)N_{x/d}(t) = \frac{6}{\pi^2}\ t(1-\log t)x^2 + O_t(x\log x).$$

Lemma 1 immediately implies that the product of 2 random integers is independent of their greatest common divisor:

Corollary 1.

$$P_x\big\{MN \leq tx^2 | (M,N) = k\big\} = t(1-\log t) + O_{t,k}\left(\frac{\log x}{x}\right).$$

To prove (1) note that

$$\begin{aligned} P_x\big\{[M,N] &\leq tx^2 \quad \text{and} \quad (M,N) = k\big\} \\ &= P_x\big\{[M,N] \leq tx^2 | (M,N) = k\big\}\cdot P_x\big\{(M,N) = k\big\} \\ &= P_x\left\{MN \leq \frac{t}{k}\ x^2 | (M,N) = k\right\}\cdot P_x\big\{(M,N) = k\big\}. \end{aligned}$$

Use of (1.2) and Corollary 1 completes the proof of (1.1). To prove (1.3) note that

$$\begin{aligned} P_x\{[M,N] \le tx^2\} &= P_x\left\{(M,N) > \left[\frac{1}{t}\right]\right\} \\ &\quad + \sum_{k=1}^{[1/t]} P_x\{[M,N] \le tx^2 | (M,N) = k\} \cdot P_x\{(M,N) = k\}. \end{aligned}$$

Using (1.2) and Corollary 1 as before completes the proof of Theorem 1.

To prove Theorem 2, write, for $k \ge 1$,

$$\begin{aligned} \sum_{m,n\le x} (m,n)^k &= 2\sum_{1\le m\le x} \sum_{1\le n\le m} (m,n)^k - \sum_{1\le i\le x} i^k \\ &= 2\sum_{1\le m\le x} f_k(m) - \frac{x^{k+1}}{k+1} + O(x^k) \end{aligned} \tag{2.2}$$

where $f_k(m) = \sum_{d|m} d^k \varphi(\frac{n}{d})$. Dirichlet's Hyperbole argument (see, e.g., Saffari (1970)) yields for any t,

$$\sum_{1\le m\le x} f_k(m) = \sum_{1\le i\le t} i^k \Phi\left(\frac{x}{i}\right) + \sum_{1\le i\le x/t} \varphi(i) I_k\left(\frac{x}{i}\right) - I_k(t)\Phi\left(\frac{x}{t}\right) \tag{2.3}$$

where

$$I_k(t) = \sum_{1\le i\le t} i^k = \frac{t^{k+1}}{k+t} + O(t^k).$$

When $k = 1$, we proceed as follows: Choose $t = \sqrt{x}$. The first sum on the right side of (2.3) is

$$\begin{aligned} &\sum_{1\le k\le\sqrt{x}} \left\{\frac{3}{\pi^2}\left(\frac{x}{k}\right)^2 + O\left(\frac{x}{k}\log\frac{x}{k}\right)\right\} \\ &\quad = \frac{3}{\pi^2}x^2\left\{\log\sqrt{x} + \gamma + O\left(\frac{1}{\sqrt{x}}\right)\right\} + O(x^{3/2}\log x). \end{aligned} \tag{2.4}$$

The second sum in (2.3) is

$$\sum_{1\le k\le\sqrt{x}} \varphi(k)\left\{\frac{1}{2}\left(\frac{x}{k}\right)^2 + O\left(\frac{x}{k}\right)\right\} = \frac{x^2}{2}\sum_{1\le k\le\sqrt{x}} \frac{\varphi(k)}{k^2} + O(x^{3/2}). \tag{2.5}$$

Now

$$\begin{aligned} \sum_{1\le k\le\sqrt{x}} \frac{\varphi(k)}{k^2} &= \sum_{1\le k\le\sqrt{x}} \frac{2k+1}{(k(k+1))^2}\Phi(k) + \frac{\Phi(\sqrt{x})}{[x]} \\ &= 2\sum_{1\le k\le\sqrt{x}} \frac{1}{k(k+1)^2}\left\{\frac{3}{\pi^2}k^2 + R(k)\right\} + \sum_{1\le k\le\sqrt{x}} \frac{\Phi(k)}{k^2(k+1)^2} + \frac{3}{\pi^2} + O\left(\frac{\log x}{\sqrt{x}}\right) \\ &= \frac{6}{\pi^2}\sum_{1\le k\le\sqrt{x}} \frac{k}{(k+1)^2} + 2\sum_{k=1}^{\infty} \frac{R(k)}{k(k+1)^2} + \sum_{k=1}^{\infty} \frac{\Phi(k)}{k^2(k+1)^2} + \frac{3}{\pi^2} + O\left(\frac{\log x}{\sqrt{x}}\right) \\ &= \frac{3}{\pi^2}\log x + d + O\left(\frac{\log x}{\sqrt{x}}\right) \end{aligned}$$

where

$$d = \sum_{k=1}^{\infty} \left\{ \Phi(k) + 2kR(k) - \frac{6}{\pi^2} k(2k+1) \right\} / \big(k(k+1)\big)^2 + \frac{6}{\pi^2}\left(\gamma + \frac{1}{2}\right) \tag{2.6}$$

and γ is Euler's constant. Using this in equation (2.5) yields that the second sum in (2.3) is

$$\frac{3x^2}{2\pi^2}\log x + \frac{d}{2}x^2 + O\big(x^{3/2}\log x\big). \tag{2.7}$$

The third term in (2.3) is

$$\frac{1}{2}\frac{3}{\pi^2}x^2 + O\big(x^{3/2}\log x\big). \tag{2.8}$$

Combining (2.8), (2.7) and (2.4) in (2.3) and using this in (2.2) yields:

$$\sum_{m,n\le x} (m,n) = \frac{6}{\pi^2}x^2\log x + \left(d + \frac{6}{\pi^2}\left(\gamma+\frac{1}{2}\right) - \frac{1}{2}\right)x^2 + O\big(x^{3/2}\log x\big),$$

where d is defined in (2.6).

When $k \ge 2$, the best choice of t in (2.3) is $t = 1$. A calculation very similar to the case of $k = 1$ leads to (1.3).

We now prove (1.6). Consider the sum

$$\begin{aligned}
\sum_{i,j\le x} [i,j]^k &= 2\sum_{i\le x}\sum_{j\le i}[i,j]^k + O\big(x^{k+1}\big) \\
&= 2\sum_{i\le x}\sum_{d\mid i}\sum_{j\le i}\left(\frac{ij}{d}\right)^k + O\big(x^{k+1}\big) \\
&= 2\sum_{\substack{i\le x\\ x}} i^k \sum_{d\mid i} f_k\left(\frac{i}{d}\right) + O\big(x^{k+1}\big) \\
&= 2\sum_{d=1}^{} d^k \sum_{j\le x/d} j^k f_k(j) + O\big(x^{k+1}\big).
\end{aligned} \tag{2.9}$$

Where

$$f_k(n) = \sum_{\substack{j\le n\\ (j,n)=1}} j^k.$$

We may derive another expression for $f_k(n)$ by considering the sum

$$\sum_{i=1}^{n} i^k = \frac{n^{k+1}}{k+1} + R_k(n) = n^k \sum_{d\mid n}\frac{f_k(d)}{d^k}. \tag{2.10}$$

Dividing (2.10) by n^k and inverting yields

$$\frac{f_k(n)}{n^k} = \frac{1}{k+1}\sum_{d\mid n}\mu\left(\frac{n}{d}\right)d + \sum_{d\mid n}\mu\left(\frac{n}{d}\right)\frac{R_k(d)}{d^k}$$

or

$$f_k(n) = \frac{n^k}{k+1}\varphi(n) + \sum_{d\mid n}\mu\left(\frac{n}{d}\right)\left(\frac{n}{d}\right)^k R_k(d) = \frac{n^k\varphi(n)}{k+1} + E(n).$$

When we substitute this expression for $f_k(j)$ in (2.9) we must evaluate:

$$\begin{aligned} S_1(y) &= \sum_{j\le y} j^k E(j) = \sum_{j\le y} j^k \sum_{d|j} \mu\Big(\frac{j}{d}\Big)\Big(\frac{j}{d}\Big)^k R_k(d) \\ &= \sum_{i\le y} \mu(i) i^{2k} \sum_{d\le y/i} R_k(d) d^k. \end{aligned}$$

Now $R_k(d)$ is a polynomial in d of degree k. Thus,

$$|S_1(y)| \le \sum_{i\le y} i^{2k}\Big(\frac{y}{i}\Big)^{2k+1} = O\big(y^{2k+1}\log y\big).$$

We must also evaluate

$$\begin{aligned} S_2(y) &= \frac{1}{k+1}\sum_{j\le y} j^k \varphi(j) \\ &= \frac{1}{k+1}\Big\{2k\sum_{j\le y} -j^{2k-1}\Phi(j) + O\Big(\sum_{j\le y} j^{2k-2}\Phi(j)\Big) + \Phi(y)y^{2k}\Big\} \\ &\quad - \frac{6}{\pi^2}\frac{k}{(k+1)}\ \frac{y^{2k+2}}{(2k+2)} + \frac{3}{\pi^2}\ \frac{1}{(k+1)}\ y^{2k+2} + O\big(y^{2k+1}\log y\big) \\ &= \frac{6}{\pi^2(k+1)}\Big(\frac{1}{2} - \frac{k}{2k+2}\Big)y^{2k+2} + O\big(y^{2k+1}\log y\big) \\ &= \frac{3}{\pi^2}\frac{1}{(k+1)^2}y^{2k+2} + O\big(y^{2k+1}\log y\big). \end{aligned}$$

Substituting in the right side of (2.9) we have

$$\begin{aligned} \sum_{i,j\le x} [i,j]^k &= 2\sum_{d=1}^{x} d^k\Big\{S_1\Big(\frac{x}{d}\Big) + S_2\Big(\frac{x}{d}\Big)\Big\} + O\big(x^{k+1}\big) \\ &= \frac{6}{\pi^2}\frac{1}{(k+1)^2}x^{2k+2}\sum_{d=1}^{x}\frac{1}{d^{k+2}} + O\big(x^{2k+1}\log x\big) \\ &= \frac{\zeta(k+2)}{\zeta(2)}\ \frac{x^{2k+2}}{(k+1)^2} + O\big(x^{2k+1}\log x\big). \end{aligned}$$

□

3. Miscellaneous remarks

1. If $M_1, M_2, \ldots, M_k$ are random integers chosen uniformly at random then the results stated in Christopher (1956) (see also Cohen (1960), Herzog and Stewart (1971), and Neymann (1972)) imply that

$$P_x\big\{(M_1, M_2, \ldots, M_k) = j\big\} = \frac{1}{\zeta(k)}\ \frac{1}{j^k} + O\Big(\frac{1}{x\ j^{k-1}}\Big) k \ge 3. \tag{3.1}$$

We have not tried to extend theorems 1 and 2 to the k-dimensional case.

(3.1) has an application to a problem in applied statistics. Suppose a population of n individuals is distributed into k categories with n individuals in category i. Often only the proportions $p_i = n_i/n$ are reported. A method for estimating n given p_i, $1 \le i \le k$ is described in Wallis and Roberts (1956), pp. 184–189. Briefly,

let $m = \min |\sum_{i=1}^{k} p_i b_i|$ where the minimum is taken over all k tuples $(b_1, b_2, \ldots, b_k)$, with $b_i \epsilon \{0, \pm 1, \pm 2, \ldots\}$ not all b_i equal zero. An estimate for n is $[1/m]$. This method works if the p_i are reported with enough precision and the n_i are relatively prime for then the Euclidean algorithm implies there are integers $\{b_i\}_{i=1}^{k}$ such that $\sum b_i n_i = 1$. These b_i give the minimum $m = \frac{1}{n}$. If it is reasonable to approximate the n_i as random integers then (3.1) implies that $\text{Prob}((n_1, n_2, \ldots, n_k) = 1) \doteq \frac{1}{\zeta(k)}$ and, as expected, as k increases this probability goes to 1. For example, $\frac{1}{\zeta(5)} \doteq .964$, $\frac{1}{\zeta(7)} \doteq .992$, $\frac{1}{\zeta(9)} \doteq .998$. This suggests the method has a good chance of working with a small number of categories. Wallace and Roberts (1956) give several examples and further details about practical implementation.

2. The best result we know for $R(x)$ defined in (2.1) is due to Saltykov (1960). He shows that

$$R(x) = O\big(x(\log x)^{2/3}(\log\log x)^{1+\epsilon}\big).$$

Use of this throughout leads to a slight improvement in the bounds of theorems 1 and 2.

3. The functions (M, N) and $[M, N]$ are both multiplicative in the sense of Delange (1969, 1970). It would be of interest to derive results similar to Theorems 1 and 2 for more general multiplicative functions.

References

[1] Cesàro, E. (1885). Etude Moyenne du plus grand Commun Diviseur de deux nombres, *Ann. Nat. Pura. Appl.* **13**(2), 233–268.

[2] Christopher, J. (1956). The asymptotic density of some k dimensional sets, *Amer. Math. Monthly* **63**, 399-401. MR97363

[3] Cohen, E. (1960). Arithmetical Functions of a Greatest Common Divisor I, *Proc. Amer. Math. Soc.* **11**, 164–171. MR111713

[4] Delange, H. (1969). Sur Les Fonctions De Plusiurs Entiers Strictement Positifs, *Enseignement Math.* **15**, 77–88. MR245538

[5] Delange, H. (1970). Sur Les Fonctions Multiplicative de Plusiurs Entiers, *Enseignement Math.* **16**, 219–246. MR294275

[6] Hardy, G. H. and Wright, E. M. (1960). *The Theory of Numbers*, Oxford University Press.

[7] Herzog, F. and Stewart, B. (1971). Patterns of Visible and non Visible Lattices, *Amer. Math. Monthly* **78**, 487–496. MR284403

[8] Neymann, J. E. (1972). On the Probability that k Positive Integers are Relatively Prime, *Jour. Number Th.* **4**, 469–473. MR304343

[9] Saffari, B. (1968). Sur quelques Applications de la "Méthode de l'hyperbole" de Dirichlet à la Théorie des Numbres Premiers, *Enseignement Math.* **14**, 205–224. MR268138

[10] Saltykov, A. I. (1960), On Eulers Function, *Vestnik Maskov Univ. Ser. I Mat. Meh.* **6**, 34–50. MR125088

[11] Wallis, W. A. and Roberts, H. V. (1956). *Statistics a New Approach.* New York, Free Press. MR88841

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 62–74
© Institute of Mathematical Statistics, 2004

Versions of de Finetti's Theorem with applications to damage models*

C. R. Rao[1] and D. N. Shanbhag[†1]

The Pennsylvania State University

Abstract: Alzaid et al. (1986) and Rao et al. (2002) have shown that several of the results on damage models have links with certain results on nonnegative matrices. Rao et al. (2002) have also shown that there is a connection between a specialized version of de Finetti's theorem for discrete exchangeable random variables and a potential theoretic result relative to nonnegative matrices. In the present article, we deal with integral equations met in damage model studies via specialized versions of de Finetti's theorem and extend further the theorems of Rao and Rubin (1964) and Shanbhag (1977) on damage models.

1. Introduction

The concept of damage models was first introduced by Rao (1963) and it has led to many interesting and illuminating characterizations of discrete distributions; among various noteworthy results in the area are those of Rao and Rubin (1964) and Shanbhag (1977). In mathematical terms, a damage model can be described by a random vector (X, Y) of non-negative integer-valued components, with the joint probability law of X and Y having the following structure:

$$P\{X = x, Y = y\} = S(y|x)g_x, \qquad y = 0, 1, 2, \ldots, x;\ x = 0, 1, 2, \ldots, \tag{1.1}$$

where $\{S(y|x) = P\{Y = y|X = x\} : y = 0, 1, 2, \ldots, x\}$ is a discrete probability law for each $x = 0, 1, 2, \ldots$ and $\{g_x = P\{X = x\} : x = 0, 1, 2, \ldots\}$ is the marginal probability law of X. In the context of damage models, the conditional probability law $\{S(y|x) : y = 0, 1, 2, \ldots, x\}$ is called the survival distribution. It is also natural to call Y the undamaged part of X and $X - Y$ the damaged part of X. Multivariate versions of the terminologies have also been dealt with in the literature. Rao and Rubin (1964) showed via Bernstein's theorem for absolutely monotonic functions that if the survival distribution is binomial with parameter vector (x, p) for almost all x (i.e. for each x with $g_x > 0$), where $p \in (0, 1)$ and fixed, and $g_0 < 1$, then the Rao-Rubin condition (RR(0))

$$P\{X = y\} = P\{Y = y|X = Y\}, \qquad y = 0, 1, 2, \ldots \tag{1.2}$$

is met if and only if X is Poisson. It was pointed out by Shanbhag (1977) that an extended version of the Rao–Rubin result can be deduced from the solution to a

*One of us has collaborated with Professor Herman Rubin on a result which is now known in statistical literature as the Rao–Rubin theorem.This theorem and another result known as Shanbhag's theorem have generated considerable research on characterization problems.Our paper on these theorems and some further results is dedicated to Professor Rubin in appreciation of his fundamental contributions to statistical inference.

[†]Address for correspondence: 3 Worcester Close, Sheffield S10 4JF, England, United Kingdom.

[1]Center for Multivariate Analysis, Thomas Building, The Pennsylvania State University, University Park, PA 16802, USA. e-mail: d.shanbhag@btopenworld.com

Keywords and phrases: de Finetti's theorem, Choquet–Deny theorem, Lau–Rao–Shanbhag theorems, Rao–Rubin–Shanbhag theorems, Rao's damage model, Rao–Rubin condition.

AMS 2000 subject classifications: 60E05, 62E10, 62H10.

general recurrence relation of the form

$$v_n = \sum_{m=0}^{\infty} w_m v_{m+n}, \qquad n = 0, 1, 2, \ldots, \tag{1.3}$$

where $\{w_m : m \geq 0\}$ is a given sequence of nonnegative real numbers with $w_1 > 0$ and $\{v_n : n \geq 0\}$ is a sequence of nonnegative real numbers to be determined. Using essentially a renewal theoretic approach, Shanbhag obtained a complete solution to (1.3), which provided a unified approach to a variety of characterizations of discrete distributions including, in particular, those related to damage models, strong memoryless property, order statistics, record values, etc.

Shanbhag's (1977) general result on damage models states essentially (in the notation described above) that if $g_0 < 1$ and, with $\{(a_n, b_n) : n = 0, 1, \ldots\}$ as a sequence of 2-component real vectors such that $a_n > 0$ for all $n, b_0, b_1 > 0$, and $b_n \geq 0$ for all $n \geq 2$, we have, for almost all x,

$$S(y|x) \propto a_y b_{x-y}, \qquad y = 0, 1, \ldots, x,$$

then the following are equivalent:

(i) $(1 \cdot 1)$ (i.e. RR(0)) is met;

(ii) Y and $X - Y$ are independent;

(iii) $(g_x/c_x) = (g_0/c_0)\lambda^x,\ x = 0, 1, \ldots$, for some $\lambda > 0$, where $\{c_n\}$ is the convolution of $\{a_n\}$ and $\{b_n\}$.

Characterizations of many standard discrete distributions in damage model studies follow as corollaries to this latter result. In particular, taking $a_n = p^n/n!,\ n = 0, 1, \ldots$, and $b_n = (1-p)^n/n!,\ n = 0, 1, \ldots$, where $p \in (0, 1)$ and fixed, we get the Rao–Rubin (1964) theorem as a corollary to this. There are several other interesting contributions to the literature on damage models. Rao and Shanbhag (1994; Chapter 7) have reviewed and unified most of these. More recently, Rao et al. (2002) and Rao et al. (2003) have provided systematic approaches to damage models based on nonnegative matrices and Markov chains. In particular, Rao et al. (2002) have shown that several of the findings on damage models in the literature are corollaries to a potential theoretic result, appearing as Theorem 4.4.1 in Rao and Shanbhag (1994), on nonnegative matrices; these subsume some of the results in the area based on the version of de Finetti's theorem for discrete exchangeable random variables.

The purpose of the present paper is to go beyond Rao et al. (2002) and show, amongst other things, that certain specialized versions of de Finetti's theorem or the relevant moment arguments provide us with further novel approaches to arrive at the Rao–Rubin–Shanbhag theorems or their generalizations. In the process of doing this, we also establish some new results on damage models or otherwise, including, in particular, an improved version of the crucial result of Alzaid et al. (1987a).

2. Simple integral equations in damage model studies

The link between the Choquet–Deny type integral equations and exchangeability or, in particular, certain versions of de Finetti's theorem for an infinite sequence

of exchangeable random variables is well-documented in Rao and Shanbhag (1994) and other places in the literature. Some specialized versions of de Finetti's theorem follow via simple arguments involving, among others, moments of probability distributions, or a potential theoretic result on nonnegative matrices; see, for example, Feller (1966, pp. 225–226) and Rao et al. (2002). A detailed account of the literature on de Finetti's theorem is provided by Aldous (1985); see, also, Chow and Teicher (1979) for an elegant proof of the theorem in the case of real-valued random variables.

Our main objective in this section though is to verify certain key results on functional equations with applications to damage models, as corollaries to specialized versions of de Finetti's theorem; the theorems and corollaries that we have dealt with in this section are obviously subumed by the relevant general results obtained via certain other techniques in Rao and Shanbhag (1994, Chapter 3) and Rao and Shanbhag (1998).

Theorem 2.1 (Shanbhag's Lemma [32]). *Let* $\{(v_n, w_n) : n = 0, 1, \ldots\}$ *be a sequence of* 2*-vectors with nonnegative real components, such that* $v_n > 0$ *for at least for one* $n > 0$ *and* $w_1 > 0$*. Then* (1.3) *is met if and only if, for some* $b > 0$*,*

$$v_n = v_0 b^n,\ n = 1, 2, \ldots, \qquad \text{and} \qquad \sum_{n=0}^{\infty} w_n b^n = 1. \tag{2.1}$$

Proof. The "if" part of the assertion is trivial. To prove the "only if" part of the assertion, let (1.3) be met with the stated assumptions. Since in that case we have $v_n(1 - w_0) \geq w_1 v_{n+1},\ n = 0, 1, \ldots$, it is clear that $w_0 < 1$ and $v_0 > 0$. (Note that Shanbhag (1977) observes via a slightly different argument that $v_n > 0$ for all $n \geq 0$, but, for us, it is sufficient to have that $v_0 > 0$.) Essentially from (1.3), we have then that there exists a sequence $\{X_n : n = 1, 2, \ldots\}$ of 0-1-valued exchangeable random variables satisfying

$$P\{X_1 = \cdots = X_n = 1\} = \frac{v_n}{v_0} w_1^n, \qquad n = 1, 2, \ldots. \tag{2.2}$$

(For some relevant information,see Remark 2.6.) From the corresponding specialized version of de Finetti's theorem, we have hence that $\{\frac{v_n}{v_0} w_1^n : n = 0, 1, \ldots\}$ is a moment sequence of a (bounded) nonnegative random variable, which, in turn, implies that $\{\frac{v_n}{v_0} : n = 0, 1, \ldots\}$ is a moment sequence of a (bounded) nonnegative random variable. Denoting the random variable in the latter case by Y and appealing to (1.3) in conjunction with the expression for Z, we get, in view of Fubini's theorem,or the monotone convergence theorem, that

$$E(Z) = E\big(Z^2\big) = 1, \tag{2.3}$$

where

$$Z = \sum_{n=0}^{\infty} w_n Y^n. \tag{2.4}$$

From (2.3), noting, for example, that $E\{(Z - 1)^2\} = 0$, we see that $Z = 1$ a.s.; consequently, from (2.4) and, in particular, the property that $w_0 < 1$, we get that there exists a number $b > 0$ such that $Y = b$ a.s. and $\sum_{n=0}^{\infty} w_n b^n = 1$. Since

$$\frac{v_n}{v_0} = E(Y^n),\ n = 0, 1, \ldots,$$

we then see that the "only if" part of the theorem holds. □

Theorem 2.2. *Let k be a positive integer and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and $\{(v_{\underline{n}}, w_{\underline{n}}) : \underline{n} \in \mathbb{N}_0^k\}$ be a sequence of 2-vectors of nonnegative real components such that $v_{\underline{0}} > 0, w_{\underline{0}} < 1$ and $w_{\underline{n}} > 0$ whenever $\underline{n}$ is of unit length. (The notation $\underline{0}$ stands for $\underline{n}$ with all coordinates equal to zero.) Then*

$$v_{\underline{n}} = \sum_{\underline{m} \in \mathbb{N}_0^k} v_{\underline{n}+\underline{m}} w_{\underline{m}}, \ \underline{n} \in \mathbb{N}_0^k \tag{2.5}$$

if and only if $\{v_{\underline{n}}/v_{\underline{0}}\}$ is the moment sequence relative to a k-component random vector $(Y_1, \ldots, Y_k)$ with Y_r's as nonnegative and bounded such that (in obvious notation)

$$\sum_{\underline{n} \in \mathbb{N}_0^k} w_{\underline{n}} \prod_{r=1}^{k} Y_r^{n_r} = 1 \quad a.s. \tag{2.6}$$

Proof. It is sufficient, as in the case of Theorem 2.1, to prove the "only if" part of the assertion. Clearly under the assumptions of the theorem taking for convenience $k \geq 2$, the validity of (2.5) implies the existence of a sequence $\{X_m : m = 1, 2, \ldots\}$ of exchangeable random variables, with values in $\{0, 1, \ldots, k\}$, satisfying (with obvious interpretation when some or all of the n_r's equal zero)

$P\{X_1, \ldots, X_{n_1+\ldots+n_k}$ are such that the first n_1 of these equal 1, the next n_2 equal 2, and so on$\}$

$$= \frac{v_{\underline{n}}}{v_{\underline{0}}} \prod_{r=1}^{k} w_{\underline{I}(r)}^{n_r}, \ \underline{n}\big(= (n_1, \ldots, n_k)\big) \in \mathbb{N}_0^k, \tag{2.7}$$

where $\underline{I}(r)$ is the rth row of the $k \times k$ identity matrix. (For some relevant information,see Remark 2.6.) Using the appropriate version of de Finetti's theorem and following a suitably modified version of the relevant part of the argument in the proof of Theorem 2.1, we see that there exists a random vector $(Y_1, \ldots, Y_k)$ as in the assertion with $\{v_{\underline{n}}/v_{\underline{0}}\}$ as the corresponding moment sequence; note especially that in this latter case (2.3) holds with Z given by the left hand side of (2.6). □

Corollary 2.1 (Hausdorff). *A sequence $\{\mu_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ of real numbers represents the moment sequence of some probability distribution concentrated on $[0, 1]^k$ if and only if $\mu_{\underline{0}} = 1$ and*

$$(-1)^{m_1+\ldots+m_k} \Delta_1^{m_1} \ldots \Delta_k^{m_k} \mu_{\underline{n}} \geq 0, \ (m_1, \ldots, m_k, \underline{n}) \in \mathbb{N}_0^{2k}, \tag{2.8}$$

where Δ_i is the usual difference operator acting on the ith coordinate.

Proof. Define the left hand side of the inequality under (2.8) by $v_{(m_1,\ldots,m_k,n_1,\ldots,n_k)}$. Then, we can easily verify that

$$\begin{aligned} v_{(m_1,\ldots,m_k,n_1,\ldots,n_k)} &= \frac{1}{k}\big\{v_{(m_1+1,\ldots,m_k,n_1,\ldots,n_k)} + \cdots + v_{(m_1,\ldots,m_k+1,n_1,\ldots,n_k)} \\ &\qquad + v_{(m_1,\ldots,m_k,n_1+1,\ldots,n_k)} + \cdots + v_{(m_1,\ldots,m_k,n_1,\ldots,n_k+1)}\big\}, \\ &\quad (m_1, \ldots, m_k, \underline{n}) \in \mathbb{N}_0^{2k}. \end{aligned}$$

Because of (2.8),Theorem 2.2 implies then that $\{\mu_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ (i.e. $\{v_{(\underline{0},\underline{n})} : \underline{n} \in \mathbb{N}_0^k\}$) is the moment sequence relative to a k-component random vector $(Y_1, \ldots, Y_k)$ with

Y_i's bounded and nonnegative. In view of (2.8), it follows further that $\{E(Y_r^{n_r}) : n_r = 0, 1, \ldots\}$ is decreasing and hence, it is obvious that the "if" part of the result holds. The "only if" part here is trivial and therefore we have the corollary. □

Remark 2.1. Although Theorem 2.1 is a corollary to Theorem 2.2, we have dealt with it separately because of its importance in characterization theory relative to univariate discrete distributions. Theorem 2.2, in turn, is a corollary to a result of Ressel (1985) and also to that of Rao and Shanbhag(1998) established via certain general versions of de Finetti's theorem, but its proof given by us here could appeal to the audience due to its simplicity. It may also be worth pointing out in this place that Chapter 3 of Rao and Shanbhag(1994) reviews and unifies,amongst other things,martingale approaches to certain generalized versions of Theorem 2.2,implied earlier;the cited chapter also shows,explicitly or otherwise,using partially a different route to ours that the following Corollaries 2.1 and 2.2 are consequences of the general results.

Remark 2.2. Corollary 2.1 can also be proved directly via de Finetti's theorem noting that there exists a sequence $\{X_n : n = 1, 2, \ldots\}$ of exchangeable random variables with values in $\{0, 1, \ldots, k\}$ and satisfying (2.7) with its right hand side replaced by $\mu_{\underline{n}} k^{-(n_1+\ldots+n_k)}$. Also, since $\{\mu_{\underline{n}}\}$ in Corollary 2.1 is the moment sequence relative to a probability distribution with compact support, it is obvious that it determines the distribution; in view of this, we can easily obtain the following result as a further corollary to Theorem 2.2

Corollary 2.2 (Bochner). *Let f be a completely monotonic function on $(0, \infty)^k$. Then f has the integral representation*

$$f(x) = \int_{[0,\infty)^k} \exp\{-\langle \underline{y}, \underline{x} \rangle\}\, d\nu(\underline{y}), \qquad \underline{x} \in (0, \infty)^k, \tag{2.9}$$

with ν as a uniquely determined measure on $[0, \infty)^k$.

Proof. Given any $\underline{x}_0 \in (0, \infty)^k$, Corollary 2.1, on taking into account the latter observation in Remark 2.2 and the continuity of f, implies after a minor manipulation that there exists a probability measure $\mu_{\underline{x}_0}$ on $[0, \infty)^k$ such that for all k-vectors $\underline{r}$ with positive rational components

$$f(\underline{x}_0 + \underline{r}) = f(\underline{x}_0) \int_{[0,\infty)^k} \exp\{-\langle \underline{y}, \underline{r} \rangle\}\, d\mu_{\underline{x}_0}(\underline{y}). \tag{2.10}$$

Since $f(\underline{x}_0 + \cdot)$ is continuous on $[0, \infty)^k$, (2.10) implies because of the dominated convergence theorem that

$$f(\underline{x}_0 + \underline{x}) = f(\underline{x}_0) \int_{[0,\infty)^k} \exp\{-\langle \underline{y}, \underline{x} \rangle\}\, d\mu_{\underline{x}_0}(\underline{y}), \qquad \underline{x} \in [0, \infty)^k.$$

In view of the arbitrary nature of $\underline{x}_0$ and the uniqueness theorem for Laplace-Stieltjes transforms, we have (2.9) to be valid with ν as unique and such that, irrespectively of what $\underline{x}_0$ is,

$$d\nu(\underline{y}) = f(\underline{x}_0) \exp\{\langle \underline{y}, \underline{x}_0 \rangle\}\, d\mu_{\underline{x}_0}(\underline{y}), \qquad \underline{y} \in [0, \infty)^k.$$

Hence, we have the Corollary. □

Remark 2.3. Bernstein's theorem for completely monotonic or absolutely monotonic functions is indeed a corollary to Corollary 2.2. Rao and Rubin (1964) have used this theorem to arrive at a characterization of Poisson distributions based on a damage model. There are also further applications of the theorem to damage models; see, for example, the next section of the present paper. Talwalker (1970) has given an extended version of the Rao–Rubin result via Corollary 2.2, while Puri and Rubin (1974) have given representations of relevance to reliability essentially via Corollaries 2.2 and 2.1, respectively; for certain observations on these latter results, see, for example, Shanbhag (1974) and Davies and Shanbhag (1987).

The following theorem of Rao and Shanbhag (1994, p.167), which is an extended version of the results of Rao and Rubin (1964) and Talwalker (1970) referred to in Remark 2.3 above as well as of the relevant result in Shanbhag (1977), is indeed a corollary to Theorem 2.2; this obviously tells us that Theorem 7.2.6 of Rao and Shanbhag (1994) is also subsumed by Theorem 2.2.

Theorem 2.3. *Let* $(\underline{X}, \underline{Y})$ *be a random vector such that* $\underline{X}$ *and* $\underline{Y}$ *are* k*-component vectors satisfying*

$$P\{\underline{X} = \underline{n}, \underline{Y} = \underline{r}\} = g_{\underline{n}} S(\underline{r}|\underline{n}), \qquad \underline{r} \in [\underline{0}, \underline{n}] \cap \mathbb{N}_0^k,\ \underline{n} \in \mathbb{N}_0^k$$

with $\{g_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ *as a probability distribution and, for each* $\underline{n}$ *for which* $g_{\underline{n}} > 0$,

$$S(\underline{r}|\underline{n}) = \frac{a_{\underline{r}} b_{\underline{n}-\underline{r}}}{c_{\underline{n}}}, \qquad \underline{r} \in [\underline{0}, \underline{n}] \cap \mathbb{N}_0^k,\ \underline{n} \in \mathbb{N}_0^k,$$

where $\{a_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ *and* $\{b_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ *are respectively positive and nonnegative real sequences with* $b_{\underline{0}} > 0$ *and* $b_{\underline{n}} > 0$ *if* $\underline{n}$ *is of unit length, and* $\{c_{\underline{n}} : \underline{n} \in \mathbb{N}_0^k\}$ *is the convolution of these two sequences. Then*

$$P\{\underline{Y} = \underline{r}\} = P\{\underline{Y} = \underline{r} | \underline{X} = \underline{Y}\},\ \underline{r} \in \mathbb{N}_0^k, \tag{2.11}$$

if and only if (*in obvious notation*)

$$g_{\underline{n}}/c_{\underline{n}} = \int_{[0,\infty)^k} \left(\prod_{i=1}^k \lambda_i^{n_i} \right) d\nu(\underline{\lambda}),\ \underline{n} \in \mathbb{N}_0^k, \tag{2.12}$$

with ($o^0 = 1$ *and*)ν *as a finite measure on* $[0,\infty)^k$ *such that it is concentrated for some* $\beta > 0$ *on* $\{\underline{\lambda} : \sum_{\underline{n} \in \mathbb{N}_0^k} b_{\underline{n}} \prod_{i=1}^k \lambda_i^{n_i} = \beta\}$.

The above theorem follows on noting especially that (2.11) is equivalent to

$$g_{\underline{n}}/c_{\underline{n}} \propto \sum_{m \in \mathbb{N}_0^k} b_{\underline{m}} (g_{\underline{m}+\underline{n}}/c_{\underline{m}+\underline{n}}), \qquad \underline{n} \in \mathbb{N}_0^k.$$

To provide a further generalization of the Rao–Rubin-Shanbhag theorems, consider S to be a countable Abelian semigroup with zero element, equipped with discrete topology, and $S^* \subset S$ such that given $w : S \to [0,\infty)$ with supp $(w)(= \{x : w(x) > 0\}) = S^*$, any function $v : S \to [0,\infty)$ with $v(0) > 0$ cannot be a solution to

$$v(x) = \sum_{y \in S} v(x+y) w(y), \qquad x \in S \tag{2.13}$$

unless it has an integral representation in terms of w-harmonic exponential functions, with respect to a probability measure. (By a w-harmonic exponential function here, we mean a function $e : S \to [0, \infty)$ such that $e(x + y) = e(x)e(y), x, y \in S$, and $\sum_{x \in S} e(x)w(x) = 1$.) Examples of such S, S^* have been dealt with by Rao and Shanbhag (1998) and studied implicitly or otherwise by Rao and Shanbhag (1994). Suppose now that $a : S \to (0, \infty)$ and $b : S \to [0, \infty)$ are such that $b(0) > 0$ and there exists $c : S \to (0, \infty)$ as the convolution of a and b, and Y and Z are random elements defined on a probability space, with values in S, such that

$$P\{Y = y, Z = z\} = g(y + z)\frac{a(y)b(z)}{c(y + z)}, \qquad y, z \in S,$$

where $\{g(x) : x \in S\}$ is a probability distribution. If $\text{supp}(b) = S^*$, then it easily follows that

$$P\{Y = y\} = P\{Y = y|Z = 0\}, \qquad y \in S,$$

if and only if $g(x)/c(x), x \in S$, is of the form of a constant multiple of the solution v to (2.13) with, for some $\gamma > 0$, w replaced by γb; this latter result is clearly an extended version of Theorem 2.3.

Remark 2.4. In view of Rao et al. (2002), the link between the general result relative to a countable semigroup that we have met above and Theorem 4.4.1 of Rao and Shanbhag (1994) or its specialized version appearing in Williams (1979) is obvious. The arguments in Rao and Shanbhag (1994) for solving general integral equations on semigroups, including those involving martingales obviously simplify considerably if the semigroups are countable; we shall throw further light on these issues through a separate article.

Remark 2.5. Modifying the proof of Theorem 2.1 slightly, involving in particular a further moment argument, a proof based on the version of de Finetti's theorem relative to 0-1-valued exchangeable random variables can be produced for Corollary 2.2.3 appearing on page 31 in Rao and Shanbhag (1994). (Note that the version of (1.3) in this case implies that there exists a nonnegative bounded random variable Y such that $E(Y^{mn}) = \frac{v_{mn}}{v_0}$, $n = 0, 1, \ldots$, for each m with $w_m > 0$.) This latter result is indeed a corollary to the Lau–Rao theorem ([13], [20]), and, in turn, is essentially a generalization of Shanbhag's lemma. As pointed out by Rao and Shanbhag (2004), in view of Alzaid et al. (1987b), there exists a proof for the Lau–Rao theorem based, among other things, on the version of de Finetti's theorem just referred to; there also exist possibilities of solving integral equations via this or other versions of de Finetti's theorem, elsewhere.

Remark 2.6. Suppose S is a countable Abelian semigroup with zero element, equipped with discrete topology, and v and w are nonnegative real-valued functions on S such that $v(0) > 0$, $w(0) < 1$, and (2.13) is met. Then there exists an infinite sequence $\{X'_n : n = 1, 2, \ldots\}$ of exchangeable random elements with values in S for which for each positive integer n and $x'_1, \ldots, x'_n \in S$,

$$P\{X'_1 = x'_1, X'_2 = x'_2, \ldots, X'_n = x'_n\} = \big(v(x'_1 + \cdots + x'_n)/v(0)\big) \prod_{i=1}^{n} w(x'_i). \quad (2.14)$$

If s_i, $i = 1, \ldots, k$ (with $k \geq 1$), are distinct nonzero members of S such that $w(s_i) > 0$, $i = 1, \ldots, k$, taking for example, X_n, $n = 1, 2, \ldots$, such that

$$X_n = \begin{cases} i & \text{if } X'_n = s_i,\ i = 1, \ldots, k, \\ 0 & \text{if } X'_n \notin \{s_1, \ldots, s_k\}, \end{cases}$$

we can now see that there exists a sequence $\{X_n : n = 1, 2, ...\}$ of exchangeable random variables with values in $\{0, 1, \ldots, k\}$ for which (2.7) (when its left hand side is read as that of (2.2) with n_1 in place of n if $k = 1$) is valid, provided its right hand side is now replaced by $\frac{v(n_1 s_1 + \cdots + n_k s_k)}{v(0)} \prod_{i=1}^{k} (w(s_i))^{n_i}$. Consequently, in view of the relevant version of de Finetti's theorem, it follows that even when s_i, $i = 1, \ldots, k$, are not taken to be distinct or nonzero, provided $w(s_i) > 0$, $i = 1, \ldots, k$, we have $\{\frac{v(n_1 s_1 + \cdots + n_k s_k)}{v(0)} : n_1, n_2, \ldots, n_k = 0, 1, ...\}$ to be the moment sequence of a probability distribution on $\mathbb{R}^k$, with support as a compact subset of $[0, \infty)^k$.

3. Spitzer's integral representation theorem and relevant observations

This section is devoted mainly to illustrate as to how Bernstein's theorem on absolutely monotonic functions,referred to in Remark 2.3, in conjunction with Yaglom's theorem mentioned on page 18 in Athreya and Ney (1972), leads us to an improved version of the key result of Alzaid et al. (1987a) and certain of its corollaries.

Suppose $\{Z_n : n = 0, 1, \ldots\}$ is a homogeneous Markov chain with state space $\{0, 1, \ldots\}$, such that the corresponding one-step transition probabilities are given by

$$\begin{aligned} p_{ij} &= P\{Z_{n+1} = j | Z_n = i\} \\ &= \begin{cases} c p_j^{(i)}, & \text{i=0,1,...; j=1,2,...,} \\ 1 - c + c p_0^{(i)}, & \text{i=0,1,...; j=0,} \end{cases} \end{aligned}$$

where $c \in (0, 1]$ and $\{p_j^{(i)} : j = 0, 1, \ldots\}$ is the i-fold convolution of some probability distribution $\{p_j\}$ for which $p_0 \in (0, 1)$, for $i = 1, 2, \ldots$, and the degenerate distribution at zero if $i = 0$. Clearly, this is an extended version of a Bienaymé-Galton-Watson branching process; indeed, we can view the latter as a special case of the former with $c = 1$.

Under the condition that $m = \sum_{j=1}^{\infty} j p_j < 1$ with $m^* = \sum_{j=1}^{\infty} (j \log j) p_j < \infty$, Alzaid et al. (1987a) have given an integral representation for stationary measures of the general process referred to above. A specialized version of this representation in the case of $c = 1$ was essentially established earlier by Spitzer (1967); this latter result appears also as Theorem 3 in Section 2 of Chapter II of Athreya and Ney (1972). The general representation theorem as well as its specialized version follow via Martin boundary related approaches or their alternatives involving specific tools such as Bernstein's theorem on absolutely monotonic functions, see, for example, Alzaid et al. (1987a) and Rao et al. (2002) for some relevant arguments or observations in this connection.

From a minute scrutiny of the proof provided by Alzaid et al. (1987a) for the general representation theorem, i.e. Theorem 2 in the cited reference, it has now emerged that the theorem referred to holds even when the constraint that $m^* < \infty$ is dropped. Indeed, Yaglom's theorem mentioned on page 18 in Athreya and Ney (1972) implies (in obvious notation) that if $m < 1$, then, irrespective of whether or not $m^* < \infty, \{\mathcal{B}_n\}$ converges pointwise to $\mathcal{B}$; essentially, the argument on page 1212 in Alzaid et al. (1987a) to show that a certain function, U^*, is the generating function of a nonnegative sequence then remains valid and gives us specifically the sequence to be that corresponding to a stationary measure of the process with $p_0 = 1 - m$ and $p_1 = m$, without requiring that $m^* < \infty$. (One can also, obviously, give the argument implied here in terms of f_n, the nth iterates of f, directly without

involving Q_n; note that we use, as usual, the notation f for the generating function of $\{p_j\}$.)

The original form of Spitzer's theorem, involving, amongst other things, the parameter $Q(0)$, requires the assumption of $m^* < \infty$. [Note that $f_n(s) = \mathcal{B}^{-1}(1 - m^n + m^n\mathcal{B}(s))$ and hence $Q_n(0) = \dfrac{(f_n(0)-1)}{m^n} = (\mathcal{B}^{-1}(1-m^n)-1)/m^n$ has a nonzero limit $Q(0)$ as $n \to \infty$ only of $\mathcal{B}'(1-) < \infty$ and hence only if $m^* < \infty$; see the proof of the theorem on page 70, in conjunction with the remark on page 18, in Athreya and Ney (1972).] However, from what we have observed above, it is clear that this latter theorem holds even when the assumption mentioned is deleted, provided "-1" is taken in place of "$Q(0)$" in the statement of the theorem.

As a by-product of the revelation that we have made above, it follows that if $m < 1, U(\cdot)$ is the generating function of a stationary measure of the process if and only if it is of the form $U^*(\mathcal{B}(\cdot))$ with U^* as the generating function of a stationary measure in the special case where $p_0 = 1 - m$, $p_1 = m$. This is obviously a consequence of Yaglom's theorem, in light of the extended continuity theorem of Feller (1966, page 433). The example given by Harris, appearing on page 72 of Athreya and Ney (1972), to prove the existence of stationary measures does not require $m^* < \infty$ and is of the form that we have met here; clearly it is not covered by Spitzer's original representation theorem. As implied in Alzaid et al. (1987a), a representation for U^* itself in our general case follows essentially as a consequence of Bernstein's theorem on absolutely monotonic functions or the Poisson-Martin integral representation theorem for a stationary measure; see, also, Rao et al. (2002) for some relevant observations.

Taking into account our observations, it is hence seen that the following modified version of the main result of Alzaid et al. (1987a) holds.

Theorem 3.1. *If $m < 1$, then every sequence $\{\eta_j : j = 1, 2, \ldots\}$ is a stationary measure if and only if, for some non-null finite measure ν on $[0, 1)$,*

$$\eta_j = \sum_{n=-\infty}^{\infty} c^n \int_{[0,1)} \exp\{-m^{n-t}\} \left(\sum_{k=1}^{j} \frac{m^{(n-t)k}}{k!} b_j^{(k)} \right) d\nu(t), \qquad j = 1, 2, \ldots, \tag{3.1}$$

where, for each k, $\{b_j^{(k)} : j = 1, 2, \ldots\}$ (with $b_0^{(k)} = 0$) denotes the distribution relative to the probability generating function $(\mathcal{B}(\cdot))^k$ with $\mathcal{B}(\cdot)$ as implied earlier (to be a unique probability generating function satisfying $\mathcal{B}(0) = 0$ and $\mathcal{B}(f(s)) = 1 - m + m\mathcal{B}(s)$, $s \in [-1, 1]$.) Moreover, if (3.1) is met with $m < 1$, then $\{\eta_j\}$ is a stationary measure satisfying $\sum_{j=1}^{\infty} \eta_j p_0^j = 1$. i.e. with generating function U such that $U(p_0) = 1$, if and only if, for some probability measure μ on $[0, 1)$,

$$d\nu(t) = K \, d\mu(t), \qquad t \in [0, 1), \tag{3.2}$$

with K such that

$$K^{-1} = \begin{cases} 1 & \text{if } c = 1 \\ \left(\dfrac{1-c}{c}\right) \displaystyle\sum_{n=-\infty}^{\infty} c^n \int_{[0,1)} \exp\{-m^{n-t}\} \, d\mu(t) & \text{if } c \in (0, 1). \end{cases}$$

The following theorem is of relevance to the topic of damage models especially in view of the results on damage models appearing in Talwalker (1980), Rao et al. (1980) and Alzaid et al. (1987a); this theorem is indeed a variation of Theorem 1 of Alzaid et al. (1987a).

Theorem 3.2. *Let $c \in (0,1)$ and $\{(v_n, h_n) : n = 0, 1, \ldots\}$ be a sequence of 2-vectors with nonnegative real components such that at least one v_n is nonzero and h_0 is nonzero and $h_1 < 1$. Then*

$$c \sum_{k=0}^{\infty} v_k h_j^{(k)} = v_j, \qquad j = 0, 1, \ldots, \tag{3.3}$$

where, for each $k > 0, \{h_j^{(k)}\}$ is the k-fold convolution of $\{h_j\}$, and $\{h_j^{(0)}\}$ is the probability distribution that is degenerate at zero, if and only if, for some $s_0 > 0$,

$$p_j = h_j s_0^{j-1}, \qquad j = 0, 1, \ldots, \tag{3.4}$$

is a nondegenerate probability distribution, $\{v_j s_0^j : j = 1, 2, \ldots\}$ is a stationary measure (not necessarily normalized as in Alzaid et al. (1987a)) relative to the general branching process with $\{p_j\}$ as in (3.4), and $v_0 = c(1-c)^{-1} \sum_{k=1}^{\infty} v_k h_0^k$.

Theorem 3.2 is easy to establish.

Remark 3.1. If $\{h_n\}$ of Theorem 3.2 satisfies a further condition that $h_n = 0$ for $n \geq 2$, then the assertion of the theorem holds with $s_0 = \frac{h_0}{(1-h_1)}$ and the stationary measure in it satisfying (3.1) with $b_1 = 1$ and $m = h_1$. Additionally, if we are given a priori that $\{v_j\}$ is of the form

$$v_j = g_j \alpha^j, \qquad j = 0, 1, \ldots$$

with $\{g_j\}$ as a probability distribution and $\alpha > 0$, then it is clear that (3.3) holds if and only if

$$g_j \propto \sum_{n=-\infty}^{\infty} c^n \int_{[0,1)} \exp\{-h_1^{n-t}\} \frac{h_1^{(n-t)j}}{j!} \left(\frac{1-h_1}{h_0 \alpha}\right)^j d\mu(t), \qquad j = 0, 1, \ldots$$

with μ as a probability measure on $[0, 1)$. As an immediate consequence of the latter result, Theorem 3 of Alzaid et al. (1987a) now follows.

Remark 3.2. One can extend the main result of Alzaid et al. (1986) based on the Perron–Frobenius theorem in an obvious way involving (in usual notation)

$$\begin{aligned} P\{Y = r\} &= P\{Y' = r | X' - Y' = k_0\} \\ &= P\{Y'' = r | X'' - Y'' = k_0 + k_1\}, \qquad r = 0, 1, \ldots \end{aligned}$$

with $k_0 \geq 0$ and $k_1 > 0$, such that the survival distributions corresponding to $(X, Y), (X', Y')$ and (X'', Y'') are not necessarily the same but $X \stackrel{d}{=} X' \stackrel{d}{=} X''$. This provides us with further insight into Theorem 3 of Alzaid et al. (1987a). (For an account of the Perron–Frobenius theorem with applications to Markov chains, see Seneta (1981).)

Remark 3.3. Most of the results dealt with in this article also follow via alternative arguments based on Choquet's theorem; for the details of this theorem, see Phelps (1966).

Remark 3.4. If we agree to rewrite the notation U^* as $U^*_{(c)}$, to take into account the value of the parameter c of the process, it easily follows (in obvious notation) that,given $c < 1$ and $U^*_{(c)}$, there exists an $U^*_{(1)}$ such that

$$\frac{d}{ds} U^*_{(1)}(s) \propto \left(\frac{d}{ds} U^*_{(c)}(s)\right) / (1-s)^{(\ln c)/(\ln m)}, \qquad s \in (-1, 1). \tag{3.5}$$

However, it is worth noting here that there exist cases of $U^*_{(1)}$ (such as those with $U^*_{(1)}(s) = (\ln(1-s))/(\ln(m))$, $s \in (-1,1)$) for which (3.5) with $c \in (0,1)$ is not met.

References

[1] Aldous, D. J. (1985). Exchangeability and related topics. *Lecture Notes in Mathematics*, **1117**, Springer, Berlin, 1–198. MR883646

[2] Alzaid, A. A., Rao, C. R., and Shanbhag, D. N. (1984). Solutions of certain functional equations and related results on probability distributions. *Unpublished Research Report*, University of Sheffield.

[3] Alzaid, A. A., Rao, C. R., and Shanbhag, D. N. (1986). An application of the Perron-Frobenius theorem to a damage model problem. *Sankhyā*, Series A, **48**, 43–50. MR883949

[4] Alzaid, A. A., Rao, C. R., and Shanbhag, D. N. (1987a). An extension of Spitzer's integral representation theorem with an application. *Ann. Probab.*, **15**, 1210–1216. MR893925

[5] Alzaid, A. A., Rao, C. R., and Shanbhag, D. N. (1987b). Solution of the integrated Cauchy equation using exchangeability. *Sankhyā*, Series A, **49**, 189–194. MR1055782

[6] Alzaid, A. A., Lau, K., Rao, C. R., and Shanbhag, D. N. (1988). Solution of Deny's convolution equation restricted to a half line via a random walk approach. *J. Multivariate Analysis*, **24**, 309–329. MR926359

[7] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer, Berlin. MR373040

[8] Chow, Y. S. and Teicher, H. (1979). *Probability Theory, Independence, Interchangeability, Martingales.* Springer Verlag, New York.

[9] Choquet, G. and Deny, J. (1960). Sur l'equation de convolution $\mu = \mu * \sigma$. *Com. Rendus Acad. Sci.*, Paris, **259**, 799–801. MR119041

[10] Davies, P. L. and Shanbhag, D. N. (1987). A generalization of a theorem of Deny with applications in characterization theory. *J. Math.*, Oxford, **38**(2), 13–34. MR876261

[11] Deny, J. (1961). Sur l'equation de convolution $\mu = \mu * \sigma$. *Sem. Theory Potent. M. Brelot., Fac. Sci.*, Paris, 1959–1960, 4 anne.

[12] Feller, W. (1966). *An Introduction to Probability and its Applications*, Vol 2. J. Wiley and Sons, New York. MR210154

[13] Lau, K. S. and Rao, C. R. (1982). Integrated Cauchy functional equation and characterization of the exponential law. *Sankhyā*, Series A, **44**, 72–90. MR753078

[14] Marsaglia, G. and Tubilla, A. (1975). A note on the lack of memory property of the exponential distribution. *Ann. Prob.* **3**, 352–354. MR365821

[15] Phelps, R. R. (1966). *Lecture Notes on Choquet's theorem.* Van Nostrand, Princeton, New Jersey.

[16] Puri, P. S. and Rubin, H. (1974). On a characterization of the family of distributions with constant multivariate failure rates. *Ann. Prob.*, **2**, 738–740. MR436463

[17] Ramachandran, B. and Lau, K. S. (1991). *Functional Equations in Probability Theory.* Academic Press, New York. MR1132671

[18] Rao, C. R. (1963). On discrete distributions arising out of methods of ascertainment. Paper presented at the Montreal conference on discrete distributions. Printed in *Sankhyā*, Series A, **27**, 1965, 311–324 and also in the *Proceedings of the conference*, 1965, 320–332. Ed.: G. P. Patil, Statistical Publishing Society, Calcutta. MR208736

[19] Rao, C. R. and Rubin, H. (1964). On a characterization of the Poisson distribution. *Sankhyā*, Series A, **26**, 295–298. MR184320

[20] Rao, C. R. and Shanbhag, D. N. (1986). Recent results on characterization of probability distributions: A unified approach through extensions of Deny's theorem. *Adv. Applied Probab.*, **18**, 660–678. MR857324

[21] Rao, C. R. and Shanbhag, D. N. (1991). An elementary proof for an extended version of the Choquet–Deny theorem *J. Mult. Anal.*, **38**, 141–148. MR1128941

[22] Rao, C. R. and Shanbhag, D. N. (1994). *Choquet–Deny Type Functional Equations with Applications to Stochastic Models.* John Wiley and Sons, Chichester, UK. MR1329995

[23] Rao, C. R. and Shanbhag, D. N. (1998). Further versions of the convolution equation. A paper dedicated to the memory of P. V. Sukhatme. *J. Indian Soc. Agricultural Statist.*, **51**, 361–378. MR1776587

[24] Rao, C. R. and Shanbhag, D. N. (2001). Exchangeability, Functional Equations and Characterizations. *Handbook of Statistics*, Vol. 19 (*Stochastic Processes: Theory and Methods*), 733–763, Eds.: Shanbhag, D. N. and Rao, C. R., Elsevier, North Holland. MR1861738

[25] Rao, C. R., Rao, M. B., and Shanbhag, D. N. (2002). Damage models: A Martin boundary connection. *Basu Memorial Volume, Sankhyā*, Vol. 64, 868–883. MR1981516

[26] Rao, C. R., Albassam, M., Rao, M. B. and Shanbhag, D. N. (2003). Markov chain approaches to damage modles. *Handbook of Statistics*, Vol. 21, Chap. 19, 775–794, Eds. Shanbhag, D. N. and Rao, C. R. Elsevier, North Holland. MR1973558

[27] Rao, C. R. and Shanbhag, D. N. (2004). Characterizations of stable laws based on a number of theoretic result. To appear in: *Comm. in Stat. Theory and Methods*, 33, No. 12, *Special issue on Characterizations*, Guest Editor: R. C. Gupta.

[28] Rao, C. R., Srivastava, R. C., Talwalker, S., and Edgar, G. A. (1980). Characterizations of probability distributions based on a generalized Rao–Rubin condition. *Sankhyā*, Series A, **42**, 161–169. MR656253

[29] Ressel, P. (1985). De Finetti-type theorems: An analytical approach. *Ann. Probab.*, **13**, 898–922. MR799427

[30] Seneta, E. (1981). *Nonnegative Matrices and Markov Chains*. Second Edition. Springer-Verlag, New York. MR719544

[31] Shanbhag, D. N. (1974). An elementary proof for the Rao–Rubin characterization of the Poisson distribution. *J. Appl. Prob.*, **11**, 211–215. MR359120

[32] Shanbhag, D. N. (1977). An extension of the Rao–Rubin characterization of the Poisson distribution. *J. Applied Probab.*, **14**, 640–646. MR451487

[33] Shanbhag, D. N. (1991). Extended versions of Deny's theorem via de Finetti's theorem. *Comput. Statist. Data Anal.*, **12**, 115–126. MR1131648

[34] Spitzer, F. (1967). Two explicit Martin boundary constructions. *Symposium on Probability Methods in Analysis. Lecture Notes in Math.*, **31**, 296–298, Springer. MR224165

[35] Talwalker, S. (1970). A characterization of the double Poisson distribution. *Sankhyā*, Ser. A, **34**, 191–193. MR293763

[36] Williams, D. (1979). *Diffusions, Markov Processes, and Martingales*. Vol 1: Foundations. John Wiley and Sons, Chichester, UK. MR531031

This list includes, amongst others, some items that are not cited explicitly, although implied, in the text; these deal with aspects of functional equations of relevance to the present study.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 75–91
© Institute of Mathematical Statistics, 2004

A short history of stochastic integration and mathematical finance: The early years, 1880–1970

Robert Jarrow[1] **and Philip Protter**[*1]

Cornell University

Abstract: We present a history of the development of the theory of Stochastic Integration, starting from its roots with Brownian motion, up to the introduction of semimartingales and the independence of the theory from an underlying Markov process framework. We show how the development has influenced and in turn been influenced by the development of Mathematical Finance Theory. The calendar period is from 1880 to 1970.

The history of stochastic integration and the modelling of risky asset prices both begin with Brownian motion, so let us begin there too. The earliest attempts to model Brownian motion mathematically can be traced to three sources, each of which knew nothing about the others: the first was that of T. N. Thiele of Copenhagen, who effectively created a model of Brownian motion while studying time series in 1880 [81].[2]; the second was that of L. Bachelier of Paris, who created a model of Brownian motion while deriving the dynamic behavior of the Paris stock market, in 1900 (see, [1, 2, 11]); and the third was that of A. Einstein, who proposed a model of the motion of small particles suspended in a liquid, in an attempt to convince other physicists of the molecular nature of matter, in 1905 [21](See [64] for a discussion of Einstein's model and his motivations.) Of these three models, those of Thiele and Bachelier had little impact for a long time, while that of Einstein was immediately influential.

We go into a little detail about what happened to Bachelier, since he is now seen by many as the founder of modern Mathematical Finance. Ignorant of the work of Thiele (which was little appreciated in its day) and preceding the work of Einstein, Bachelier attempted to model the market noise of the Paris Bourse. Exploiting the ideas of the Central Limit Theorem, and realizing that market noise should be without memory, he reasoned that increments of stock prices should be independent and normally distributed. He combined his reasoning with the Markov property and semigroups, and connected Brownian motion with the heat equation, using that the Gaussian kernel is the fundamental solution to the heat equation. He was able to define other processes related to Brownian motion, such as the maximum change during a time interval (for one dimensional Brownian motion), by using random walks and letting the time steps go to zero, and by then taking

[*]Supported in part by NSF grant DMS-0202958 and NSA grant MDA-904-03-1-0092.

[1]School of Operations Research and Industrial Engineering, Cornell University, 219 Rhodes Hall, Ithaca, NY 14853, USA. e-mail: pep4@cornell.edu

Keywords and phrases: stochastic integration, semimartingales, martingales, Brownian motion, Markov processes, Black-Scholes, options, warrants, contingent claims, hedging strategies, Bachelier, homogeneous chaos, history of mathematics.

AMS 2000 subject classifications: 01A60, 60H05, 60H30, 60G44, 60G35, 60G46, 91B70, 91B28, 91B99, 60J45, 60J55, 60J65.

[2]This was called to our attention by Ragnar Norberg, whom we thank, and the contributions of Thiele are detailed in a paper of Hald [30].

limits. His thesis was appreciated by his mentor H. Poincaré, but partially due to the distaste of studying economics as an application of mathematics, he was unable to join the Paris elite, and he spent his career far off in the provincial capital of Besançon, near Switzerland in Eastern France. (More details of this sad story are provided in [11]).

Let us now turn to Einstein's model. In modern terms, Einstein assumed that Brownian motion was a stochastic process with continuous paths, independent increments, and stationary Gaussian increments. He did not assume other reasonable properties (from the standpoint of physics), such as rectifiable paths. If he had assumed this last property, we now know his model would not have existed as a process. However, Einstein was unable to show that the process he proposed actually did exist as a mathematical object. This is understandable, since it was 1905, and the ideas of Borel and Lebesgue constructing measure theory were developed only during the first decade of the twentieth century.

In 1913 Daniell's approach to measure theory (in which integrals are defined before measures) appeared, and it was these ideas, combined with Fourier series, that N. Wiener used in 1923 to construct Brownian motion, justifying after the fact Einstein's approach. Indeed, Wiener used the ideas of measure theory to construct a measure on the path space of continuous functions, giving the canonical path projection process the distribution of what we now know as Brownian motion. Wiener and others proved many properties of the paths of Brownian motion, an activity that continues to this day. Two key properties relating to stochastic integration are that (1) the paths of Brownian motion have a non zero finite quadratic variation, such that on an interval (s,t), the quadratic variation is $(t-s)$ and (2) the paths of Brownian motion have infinite variation on compact time intervals, almost surely. The second property follows easily from the first. Note that if Einstein were to have assumed rectifiable paths, Wiener's construction would have essentially proved the impossibility of such a model. In recognition of his work, his construction of Brownian motion is often referred to as *the Wiener process*. Wiener also constructed a multiple integral, but it was not what is known today as the "Multiple Wiener Integral": indeed, it was K. Itô, in 1951, when trying to understand Wiener's papers (not an easy task), who refined and greatly improved Wiener's ideas [36].

The next step in the groundwork for stochastic integration lay with A. N. Kolmogorov. The beginnings of the theory of stochastic integration, from the nonfinance perspective, were motivated and intertwined with the theory of Markov processes, in which Kolmogorov, of course, played a fundamental role. Indeed, in 1931, two years before his famous book establishing a rigorous mathematical basis for Probability Theory using measure theory, Kolmogorov refers to and briefly explains Bachelier's construction of Brownian motion ([41], pp. 64, 102–103). It is this paper too in which he develops a large part of his theory of Markov processes. Most significantly, in this paper Kolmogorov showed that continuous Markov processes (diffusions) depend essentially on only two parameters: one for the speed of the drift and the other for the size of the purely random part (the diffusive component). He was then able to relate the probability distributions of the process to the solutions of partial differential equations, which he solved, and which are now known as "Kolmogorov's equations." Of course, Kolmogorov did not have the Itô integral available, and thus he relied on an analysis of the semigroup and its infinitesimal generator, and the resulting partial differential equations.[3]

[3] J. L. Doob [17] has complained that the PDE methods of Kolmogorov and Feller used to study Markov processes have often been called "analytic", whereas the method of stochastic differentials

After Kolmogorov we turn to the fascinating and tragic story of Vincent Doeblin (born Wolfgang Döblin) the son of the author Alfred Döblin, who wrote *Berlin Alexanderplatz* for example. The Döblin family fled the Nazis from Germany, first to Switzerland, and then to Paris. Wolfgang changed his name to Vincent Doeblin, and became a French citizen, finishing his schooling there and being quickly recognized as an extraordinary mathematical talent. In the late 1920's Probability Theory was becoming stylish among mathematicians, especially in the two centers, Moscow and Paris. Doeblin joined the probabilists, working on Markov chains and later Markov processes.[4] Doeblin wanted to construct a stochastic process with continuous paths that would be consistent with Kolmogorov's analytic theory of transition probabilities for Markov processes. He ultimately developed a framework to study them which was prescient in regards to future developments. However Doeblin was drafted, and he volunteered to go to the front. Before he went he sketched out his ideas and he put this work in the safe of the National Academy of Science of France, to be opened only by him or else after 100 years. As the Maginot line fell, to avoid sharing his ideas with the Nazis Doeblin first burned his notes, and then he took his own life. The academy safe was opened only in May 2000, at the request of his brother, Claude Doeblin. It was only then that the far reaching vision of his work became apparent. In those notes, he utilized the new concept of martingales proposed by J. Ville only in 1939 [84] and understood the importance of studying sample paths, instead of relying exclusively on distributional properties. One idea he had was to run Brownian motion by a random clock: what is known today as a time change. The change of time was then related to the diffusion coefficient, and in this way he was able to give a modern treatment of diffusions decades before it was developed otherwise. [5]

We turn now to Kiyosi Itô, the father of stochastic integration. We will not attempt to reproduce the beautiful summary of his work and contributions provided in 1987 by S. R. S. Varadhan and D. W. Stroock [83], but instead give a short synopsis of what we think were key moments.[6] No doubt an attempt to establish a true *stochastic differential* to be used in the study of Markov processes was one of Itô's primary motivations for studying stochastic integrals, just as it was Döblin's before him, although of course Döblin's work was secret, hidden away in the safe of the French Academy of Science. Wiener's integral did not permit stochastic processes as integrands, and such integrands would of course be needed

introduced by Itô has in contrast been called "probabilistic". Indeed, he writes, "It is considered by some mathematicians that if one deals with analytic properties and expectations then the subject is part of analysis, but that if one deals with sample sequences and sample functions then the subject is probability but not analysis". Doob then goes on to make his point convincingly that both methods are probability. (Doob's criticism is likely to have been partially inspired by comments of the second author.) Nevertheless, we contend that the methods of Itô changed the *probabilistic intuition* one develops when studying Markov processes.

[4] J. Doob references his fundamental work on Markov chains and Markov processes extensively in his book [17], for example. Paul Lévy wrote of him in an article devoted to an appreciation of his work after his death: "Je crois pouvoir dire, pour donner une idée du niveau où il convient de le situer, qu'on peut compter sur les doigts d'une seule main les mathématiciens qui, depuis Abel et Galois, sont morts si jeunes en laissant une oeuvre aussi important". Translated: 'I can say, to give an idea of Doeblin's stature, that one can count on the fingers of one hand the mathematicians who, since Abel and Galois, have died so young and left behind a body of work so important.' See [44]

[5] The second author is grateful to Marc Yor for having sent to him his beautiful article, written together with Bernard Bru [6]. This article, together with the companion (and much more detailed) article [7], are the sources for this discussion of Doeblin. In addition, the story of Doeblin has recently been turned into a book in biographical form [65].

[6] The interested reader can also consult [66].

if one were to represent (for example) a diffusion as a solution of a stochastic differential equation. Indeed, Itô has explained this motivation himself, and we let him express it: "In these papers[7] I saw a powerful analytic method to study the transition probabilities of the process, namely Kolmogorov's parabolic equation and its extension by Feller. But I wanted to study the paths of Markov processes in the same way as Lévy observed differential processes. Observing the intuitive background in which Kolmogorov derived his equation (explained in the introduction of the paper), I noticed that a Markovian particle would perform a time homogeneous differential process for infinitesimal future at every instant, and arrived at the notion of a stochastic differential equation governing the paths of a Markov process that could be formulated in terms of the differentials of a single differential process" [37].[8]

Itô's first paper on stochastic integration was published in 1944 ([34]), the same year that Kakutani published two brief notes connecting Brownian motion and harmonic functions. Meanwhile throughout the 1940's Doob, who came to probability from complex analysis, saw the connection between J. Ville's martingales and harmonic functions, and he worked to develop a martingale based probabilistic potential theory. In addition, H. Cartan greatly advanced potential theory in the mid 1940's, later followed by Deny's classic work in 1950. All these ideas swirling around were interrelated, and in the 1940s Doob, clearly explained, for the first time, what should be the strong Markov property. A few years later (in 1948) E. Hille and K. Yosida independently gave the structure of semigroups of strongly continuous operators, clarifying the role of infinitesimal generators in Markov process theory.

In his efforts to model Markov processes, Itô constructed a stochastic differential equation of the form:

$$dX_t = \sigma(X_t)dW_t + \mu(X_t)dt,$$

where of course W represents a standard Wiener process. He now had two problems: one was to make sense of the stochastic differential $\sigma(X_t)dW_t$ which he accomplished in the aforementioned article [34].[9] The second problem was to connect Kolmogorov's work on Markov processes with his interpretation. In particular, he wanted to relate the paths of X to the transition function of the diffusion. This amounted to showing that the distribution of X solves Kolmogorov's forward equation. This effort resulted in his spectacular paper [35] in 1951, where he stated and proved what is now known as **Itô's formula**:

$$f(X_t) = f'(X_t)dX_t + \frac{1}{2}f''(X_t)d[X,X]_t.$$

Here the function f is of course assumed to be $\mathcal{C}^2$, and we are using modern notation.[10] Itô's formula is of course an extension of the change of variables formula for

[7] Here Itô is referring to the papers of Kolmogorov [41] and of Feller [26].

[8] Note that while Itô never mentions the work of Bachelier in his foreword, citing instead Kolmogorov, Lévy, and Doob as his main influences, it is reasonable to think he was aware of the work of Bachelier, since it is referenced and explained in the key paper of Kolmogorov ([41]) that he lists as his one of his main inspirations. While we have found no direct evidence that Itô ever read Bachelier's work, nevertheless Hans Föllmer and Robert Merton have told the authors in private communications that Itô had indeed been influenced by the work of Bachelier. Merton has also published this observation: see page 47 of [51].

[9] Here Itô cites the work of S. Bernstein [5] as well as that of Kolmogorov [41] and W. Feller [26] as antecedents for his work.

[10] The book by H. P. McKean, Jr., published in 1969 [47], had a great influence in popularizing the Itô integral, as it was the first explanation of Itô's and others' related work in book form. But McKean referred to Itô's formula as *Itô's lemma*, a nomenclature that has persisted in some

Riemann-Stieltjes integration, and it reveals the difference between the Itô stochastic calculus and that of the classical path by path calculus available for continuous stochastic processes with paths of bounded variation on compact time sets. That formula is, of course, where A denotes such a process and f is $\mathcal{C}^1$:

$$df(A_t) = f'(A_t)dA_t.$$

It can be shown that if one wants to define a path by path integral of the form $\int_0^t H_s dA_s$ as the limit of sums, where H is any process with continuous sample paths, then as a consequence of the Banach Steinhaus theorem *A a fortiori* has sample paths of bounded variation on compacts. (See, for example, [67].) Since Brownian motion has paths of unbounded variation almost surely on *any* finite time interval, Itô knew that it was not possible to integrate all continuous stochastic processes. One of his key insights was to limit his space of integrands to those that were, as he called it, *non anticipating*. That is, he only allows integrands that are adapted to the underlying filtration of σ-algebras generated by the Brownian motion. This allowed him to make use of the independence of the increments of Brownian motion to establish the L^2 isometry

$$E\left(\left(\int_0^t H_s dW_s\right)^2\right) = E\left(\int_0^t H_s^2 ds\right).$$

Once the isometry is established for continuous non-anticipating processes H, it then extends to jointly measurable non-anticipating processes.[11]

J. L. Doob realized that Itô's construction of his stochastic integral for Brownian motion did not use the full strength of the independence of the increments of Brownian motion. In his highly influential 1953 book [16] he extended Itô's stochastic integral for Brownian motion first to processes with orthogonal increments (in the L^2 sense), and then to processes with conditionally orthogonal increments, that is, martingales. What he needed, however, was a martingale M such that $M_t^2 - F(t)$ is again a martingale, where the increasing process F is *non-random*. He established the now famous Doob decomposition theorem for submartingales: *If X_n is a (discrete time) submartingale, then there exists a unique decomposition $X_n = M_n + A_n$ where M is a martingale, and A is a process with non-decreasing paths, $A_0 = 0$, and with the special measurability property that A_n is $\mathcal{F}_{n-1}$ measurable.* Since M^2 is a submartingale when M is a martingale, he needed an analogous decomposition theorem in continuous time in order to extend further his stochastic integral. As it was, however, he extended Itô's isometry relation as follows:

$$E\left(\left(\int_0^t H_s dM_s\right)^2\right) = E\left(\int_0^t H_s^2 dF(s)\right),$$

where F is non-decreasing and non-random, $M^2 - F$ is again a martingale, and also the stochastic integral is also a martingale. (See Chapter IX of [16].)

circles to this day. Obviously this key theorem of Itô is much more important than the status the lowly nomenclature "lemma" affords it, and we prefer Itô's own description: "formula".

[11] Indeed, this is how the theory is presented in the little 1969 book of McKean [47]. Unfortunately it is not as simple as McKean thought at this early stage of the theory, to determine exactly which processes are included in this procedure; the natural σ-algebra generated by the simple integrands is today known as the *predictable* σ-algebra, and the predictably measurable processes are a strict subset of jointly measurable, non-anticipating processes. This point is clarified in (for example) the book of K. L. Chung and R. Williams [9], p. 63.

Thus it became an interesting question, if only for the purpose of extending the stochastic integral to martingales in general, to see if one could extend Doob's decomposition theorem to submartingales indexed by continuous time. However there were other reasons as well, such as the development of probabilistic potential theory, which began to parallel the development of axiomatic potential theory, especially with the publication of G. A. Hunt's seminal papers in 1957 and 1958 [31, 32, 33]. It took perhaps a decade for these papers to be fully appreciated, but in the late 1960's and early 1970's they led to even greater interest in Itô's treatment of Markov processes as solutions of stochastic differential equations, involving both Brownian motion and what is today known as Poisson random measure.

The issue was resolved in two papers by the (then) young French mathematician P. A. Meyer in 1962. Indeed, as if to underline the importance of probabilistic potential theory in the development of the stochastic integral, Meyer's first paper, establishing the existence of the Doob decomposition for continuous time submartingales [52], is written in the language of potential theory. Meyer showed that the theorem is false in general, but true if and only if one assumes that the submartingale has a uniform integrability property when indexed by stopping times, which he called "Class (D)", clearly in honor of Doob. Ornstein had shown that there were submartingales not satisfying the Class (D) property[12], and G. Johnson and L. L. Helms [40] quickly provided an example in print in 1963, using three dimensional Brownian motion. Also in 1963, P. A. Meyer established the uniqueness of the Doob decomposition [53], which today is known as the Doob-Meyer decomposition theorem. In addition, in this second paper Meyer provides an analysis of the structure of L^2 martingales, which later will prove essential to the full development of the theory of stochastic integration. Two years later, in 1965, Itô and S. Watanabe, while studying multiplicative functionals of Markov processes, define *local martingales* [39]. This turns out to be the key object needed for Doob's original conjecture to hold. That is, any submartingale X, whether it is of Class (D) or not, has a unique decomposition

$$X_t = M_t + A_t,$$

where M is a local martingale, and A is a non-decreasing, predictable process with $A_0 = 0$.

Returning however to P. A. Meyer's original paper [52], at the end of the paper, as an application of his decomposition theorem, he proposes an extension of Doob's stochastic integral, and thus *a fortiori* an extension of Itô's integral. His space of integrands is that of "well adapted" processes, meaning jointly measurable and adapted to the underlying filtration of σ-algebras. He makes the prescient remark at the end of his paper that "it seems hard to show (though it is certainly true) that the full class of well adapted processes whose "norm" is finite has been attained by this procedure." This anticipates the oversight of McKean six years later (see footnote 11), and it is this somewhat esoteric measurability issue that delays the full development of stochastic integration for martingales which have jumps, as we shall see.

Before we continue our discussion of the evolution of the theory of stochastic integration, however, let us digress to discuss the developments in economics. It is curious that Peter Bernstein, in his 1992 book [4], states "Despite its importance, Bachelier's thesis was lost until it was rediscovered quite by accident in the 1950's by Jimmie Savage, a mathematical statistician at Chicago." He goes on a little later to say "Some time around 1954, while rummaging through a university library, Savage

[12] See, for example, [59], p. 823

chanced upon a small book by Bachelier, published in 1914, on speculation and investment." We know however that Kolmogorov and also Doob explicitly reference Bachelier, and Itô certainly knew of his work too; but perhaps what was "lost" was Bachelier's contributions to economics.[13] Bernstein relates that Savage alerted the economist Paul Samuelson to Bachelier's work, who found Bachelier's thesis in the MIT library, and later remarked "Bachelier seems to have had something of a one-track mind. But what a track!" [73]. See also [74].

After a decade of lectures around the country on warrant pricing and how stock prices must be random,[14] Samuelson then went on to publish, in 1965, two papers of ground breaking work. In his paper [72] he gives his economics arguments that prices must fluctuate randomly, 65 years after Bachelier had assumed it! This paper, along with Fama's [24] work on the same topic, form the basis of what has come to be known as "the efficient market hypothesis." The efficient market hypothesis caused a revolution in empirical finance; the debate and empirical investigation of this hypothesis is still continuing today (see [25]). Two other profound insights can be found in this early paper that subsequently, but only in a modified form, became the mainstay of option pricing theory. The first idea is the belief (postulate) that discounted futures prices follow a martingale[15]. From this postulate, Samuelson proved that changes in futures prices were uncorrelated across time, a generalization of the random walk model (see [46], and also [13]). The second insight is that this proposition can be extended to arbitrary functions of the spot price, and although he did not state it explicitly herein, this forebodes an immediate application to options.

In his companion paper [71], he combined forces with H.P. McKean Jr.[16] (who the same year published his tome together with K. Itô [38]) who wrote a mathematical appendix to the paper, to show essentially that a good model for stock price movements is what is today known as geometric Brownian motion. Samuelson explains that Bachelier's model failed to ensure that stock prices always be positive, and that his model leads to absurd inconsistencies with economic principles, whereas geometric Brownian motion avoids these pitfalls. This paper also derived valuation formulas for both European and American options.[17] The derivation was almost identical to that used nearly a decade later to derive the Black-Scholes formula, except that instead of invoking a no arbitrage principle to derive the valuation formula, he again postulated the condition that the discounted options payoffs follow a martingale (see [71] p. 19), from which the valuation formulae easily followed.

[13]It is possible that L. J. Savage read Bachelier's work because Doob's book had appeared only one year earlier and had referenced it, and then he might have been surprised by the economics content of Bachelier's work. But this is pure speculation. Also, Samuelson wrote in [73] (p. 6) that "this was largely lost in the literature, even though Bachelier does receive occasional citation in standard works in probability."

[14]These lectures lead to other papers being published by researchers following up on Samuelson's ideas, for example the renowned paper of Osborne [62].

[15]See the Theorem of Mean Percentage Price Drift on page 46 and the subsequent discussion.

[16]Samuelson combined forces with McKean, and later R. C. Merton, because he did not feel comfortable with the newly developed stochastic calculus (see [4] p. 215). This insight was also confirmed by private communications with R. C. Merton.

[17] This is the paper that first coined the terms "European" and "American" options. According to a private communication with R.C. Merton, prior to writing the paper, P. Samuelson went to Wall Street to discuss options with industry professionals. His Wall Street contact explained that there were two types of options available, one more complex - that could be exercised any time prior to maturiy, and one more simple - that could be exercised only at the maturity date, and that only the more sophisticated European mind (as opposed to the American mind) could understand the former. In response, when Samuelson wrote the paper, he used these as prefixes and reversed the ordering.

The much later insights of Black, Scholes, and Merton, relating prices of options to perfect hedging strategies, is of course not discussed in this article. Furthermore, it is also noteworthy that within this paper, Samuelson and McKean determine the price of an American option by discovering the relation of an American option to a free boundary problem for the heat equation. This is the first time that this connection is made. Interestingly, Samuelson and McKean do not avail themselves of the tools of stochastic calculus, at least not explicitly. The techniques McKean uses in his appendix are partial differential equations in the spirit of Kolmogorov, coupled with stopping times and the potential theoretic techniques pioneered by G. Hunt and developed by Dynkin.

The final precursor to the Black, Scholes and Merton option pricing formulaes can be found in the paper of Samuelson and Merton [75]. Following similar mathematics to [71], instead of invoking the postulate that discounted option payoffs follow a martingale, they derived this postulate as an implication of a utility maximizing investor's optimization decision. Herein, they showed that the option's price could be viewed as its discounted expected value, where instead of using the actual probabilities to compute the expectation, one uses utility or risk adjusted probabilities[18]. These risk adjusted probabilities later became known as "risk-neutral" or "equivalent martingale" probabilities. It is interesting to note that, contrary to common belief, this use of "equivalent martingale probabilities" under another guise predated the paper by Cox and Ross [12] by nearly 10 years. In fact, Merton (footnote 5 page 218, [50]) points out that Samuelson knew this fact as early as 1953! Again, by not invoking the no arbitrage principle, this paper just missed obtaining the famous Black Scholes formula. The first use of the no arbitrage principle to prove a pricing relation between various financial securities can be found in Modigliani and Miller [60] some eleven years earlier, where they showed the equivalence between two different firms' debt and equity prices, generating the famous M&M Theorem. Both Samuelson and Merton were aware of this principle, Modigliani being a colleague at M.I.T., but neither thought to apply it to this pricing problem until many years later.

Unrelated to finance, and almost as an aside in the general tide of the development of the theory of stochastic integration, were the insights of Herman Rubin. At the Third Berkeley Symposium in 1955, Rubin gave a talk on stochastic differential equations. The following year, he presented an invited paper at the Seattle joint meetings of the Institute of Mathematical Statistics, the American Mathematical Society , the Biometric Society, the Mathematical Association of America, and the Econometrics Society. In this paper he outlined what was later to become D. L. Fisk's Ph.D. thesis, which invented both quasimartingales and what is now known as the Stratonovich integral. To quote his own recollections, "I was unhappy with the Itô integral because of the lack of invariance with nonlinear change of coordinate systems, no matter how smooth, and, observing that using the average of the right and left endpoints gave exactly the right results for the integral of XdX for any X (even discontinuous), it seemed that this was, for continuous X with sufficiently good properties, the appropriate candidate for the integral...Quasimartingales seemed the natural candidate for the class of processes, but I did not see a clear proof. I gave the problem to Fisk to work on for a Ph.D. thesis, and he did come up with what was needed" [69].

Indeed, in D. L. Fisk's thesis [27], written under Rubin when he was at Michigan State University, Fisk developed what is now known as the Stratonovich integral,

[18] See especially expression (20) on page 26.

and he also coined the phrase and developed the modern theory of quasimartingales, later used by K. M. Rao [68] to give an elegant proof that a quasimartingale is the difference of two submartingales, and also used by S. Orey [63] in a paper extending the idea and which foreshadowed modern day semimartingales. Fisk submitted his thesis for publication, but the editor did not believe there was much interest in stochastic integration, again according to the recollections of Herman Rubin [69]. So Fisk dropped that part of the thesis and did not pursue it, publishing instead only the part on quasimartingales, which appeared as [28].

Returning now to the historical development of stochastic integration, we mention that P. A. Meyer's development of the stochastic integral in [52] is skeletal at best, and a more systematic development is next put forward by Philippe Courrège in 1963 [10]. The motivation clearly arises from potential theory, and the paper of Courrège is published not in a journal, but in the (at the time) widely circulated *Séminaire Brélot-Choquet-Dény (Théorie du Potentiel)*. Many reasonable Markov processes, and in particular those treated by Hunt ([31, 32, 33]), have the property that they are *quasi-left continuous*. That is, they have paths which are right continuous with left limits a.s., and if there is a jump at a stopping time T, then that time T must be totally inaccessible. Intuitively, T must come as a complete surprise. One can formulate the condition of quasi-left continuity in terms of the underlying filtration of σ-algebras of the Markov process as well. This seems to be a reasonable property for the filtration of a time homogeneous Markov process to have, and is satisfied for a vast collection of examples.

It was natural for someone working in potential theory to make the assumption that the filtration is quasi-left continuous, and such an assumption has the fortuitous consequence to imply that if X is a submartingale and $X = M + A$ is its Doob-Meyer decomposition, then A has continuous sample paths. What this means is that in the L^2 isometry

$$E\left(\left(\int_0^t H_s\, dM_s\right)^2\right) = E\left(\int_0^t H_s^2\, dA_s\right),$$

where A is the increasing process corresponding to the submartingale $X = M^2$, one extends the Itô-Doob technique to general L^2 martingales, and the resultant increasing random process A has continuous paths. This, it turns out, greatly simplifies the theory. And it is precisely this assumption that Courrège makes. Courrège also works with integrands which have left continuous paths, and he considers the space of processes that are measurable with respect to the σ-algebra they generate, on $\mathbb{R} \times \Omega$, calling it processes which are "fortement bien adapté". Thus Courrège had, in effect, pioneered the *predictable σ-algebra*, although he did not use it as P. A. Meyer did, as we shall see. As it turns out, if dA_t is path by path absolutely continuous with respect to dt (this is usually written $dA_t << dt$), almost surely, then there ends up being essentially no difference which σ-algebra one uses: the predictable σ-algebra, or the progressive σ-algebra,[19] or even jointly measurable adapted processes. However if A is merely continuous and does not necessarily have absolutely continuous paths a.s., then one needs at least the progressive σ-algebra. We now know that what happens is that the difference between one such process and its predictable projection is a process that has a stochastic integral which is

[19]The *progressive σ-algebra* is defined later in the theory, and it has the property that if a process H_s is progressively measurable, and if τ is a finite valued stopping time, then H_τ is $\mathcal{F}_\tau$ measurable.

the zero process a.s, and this is why it does not matter. (For a detailed explanation see Liptser and Shiryaev [45], or alternatively Chung and Williams [9]).

One important thing that Courrège did not do, however, was to prove a change of variables formula, analogous to Itô's formula for stochastic integration with respect to Brownian motion. This was done in 1967 in an influential paper of H. Kunita and S. Watanabe [42]. Whereas the approach of Courrège was solidly in the tradition of Doob and Itô, that of establishing an L^2 isometry, the approach pioneered by M. Motoo and S. Watanabe two years later in 1965 was new: they treated the stochastic integral as an operator on martingales having specific properties, utilizing the Hilbert space structure of L^2 by using the Doob-Meyer increasing process to inspire an inner product through the quadratic variation of martingales. (See [61]). In the same paper Motoo and Watanabe established a martingale representation theorem which proved to be prescient of what was to come: they showed that all L^2 martingales defined on a probability space obtained via the construction of a type of Markov process named a Hunt process (in honor of the fundamental papers of G. Hunt mentioned earlier)were generated by a collection of additive functionals which were also L^2 martingales, and which were obtained in a way now associated with Dynkin's formula and "martingale problems."

The important paper of Motoo and Watanabe, however, was quickly overshadowed by the subsequent and beautifully written paper of H. Kunita and S. Watanabe, published in 1967 [42]. Here Kunita and Watanabe developed the ideas on orthogonality of martingales pioneered by P. A. Meyer, and Motoo and Watanabe, and they developed a theory of stable spaces of martingales which has proved fundamental to the theory of martingale representation, known in Finance as "market completeness." They also clarified the idea of quadratic variation as a pseudo inner product, and used it to prove a general change of variables formula, profoundly extending Itô's formula for Brownian martingales. The formula was clean and simple for martingales with continuous paths, but when it came to the general case (i.e., martingales that can have jump discontinuities in their sample paths)the authors retreated to the rich structure available to them in the Hunt process setting, and they expressed the jumps in terms of the Lévy system of the underlying Markov process. (Lévy systems for Markov processes, a structure which describes the jump behavior of a Hunt process, had only been developed a few years earlier in 1964 by S. Watanabe [85], and extended much later by A. Benveniste and J. Jacod [3]). This "retreat" must have seemed natural at the time, since stochastic integrals were, as noted previously, seen as intimately intertwined with Markov processes. And also, as an application of their change of variables formula, Kunita and Watanabe gave simple and elegant proofs of Lévy's theorem characterizing Brownian motion among continuous martingales via its quadratic variation process, as well as an extension from one to N dimensions of the spectacular 1965 theorem of L. Dubins and G. Schwarz [18] and K. E. Dambis [14] that a large class of continuous martingales can be represented as time changes of Brownian motion.

This remarkable paper of Kunita and Watanabe was quickly appreciated by P.A. Meyer, now in Strasbourg. He helped to start, with the aid of Springer-Verlag, the *Séminaire de Probabilités*, which is one of the longest running seminars to be published in Springer's famed *Lecture Notes in Mathematics* series. In the first issue, which is only Volume 39 in the *Lecture Notes* series, he published four key papers inspired by the article of Kunita and Watanabe [54, 55, 56, 57].[20] In

[20] A large number of the historically important works on stochastic integration were published in the Séminaire de Probabilités series, and these papers have been recently reprinted in a new

these papers he made two important innovations: he went beyond the "inner product" of Kunita and Watanabe (which is and was denoted $< X, Y >$, and which is tied to the Doob-Meyer decomposition), and expanding on an idea of Austin for discrete parameter martingales he created the "square bracket" (*le crochet droit*) pseudo inner product, denoted $[X, Y]$. Unlike the bracket process $< X, Y >$, which exists for all locally square integrable martingales (and therefore all continuous ones), the square bracket process exists for all martingales, and even all local martingales. This turned out to be important in later developments, such as the invention of semimartingales, and of course is key to the extension of the stochastic integral to all local martingales, and not only locally square integrable ones.

The second major insight of Meyer in these papers is his realization of the importance of the predictable σ-algebra. Going far beyond Courrège he realized that when a martingale also had paths of finite variation (of necessity a martingale with jumps), the stochastic integral should agree with a path by path construction using Lebesgue-Stieltjes integration. He showed that this holds if and only if the integrand is a predictable process. Moreover, he was able to analyze the jumps of the stochastic integral, observing that the stochastic integral has the same jump behavior as does the Lebesgue-Stieltjes integral if the integrand is predictably measurable. This laid the groundwork for the semimartingale theory that was to come a few years later.

We should further note at this point that Meyer was able to discard the Markov process framework used by Kunita and Watanabe in the first two of the four papers, and he established the general change of variables formula used today without using Lévy systems. Meyer then applied his more general results to Markov processes in the latter two of his four papers. Again, this was natural, since one of Meyer's primary interests was to resolve the many open questions raised by Hunt's seminal papers. It was research in Markov processes that was driving the interest in stochastic integration, from Itô on, up to this point. Nevertheless, Doob had begun to isolate the martingale character of processes independent of Markov processes, and Meyer's approach in his classic papers of 1962 and 1963 (already discussed [52] and [53]) was to use the techniques developed in Markov process potential theory to prove purely martingale theory results.

The development of stochastic integration as recounted so far seems to be primarily centered in Japan and France. But important parallel developments were occurring in the Soviet Union. The books of Dynkin on Markov processes appeared early, in 1960 [19] and in English as Springer Verlag books in 1965 [20]. The famed Moscow seminar (reconstituted at least once on October 18 and 19, 1996 in East Lansing, Michigan, with Dynkin, Skorohod, Wentzell, Freidlin, Krylov, etc.), and Girsanov's work on transformations of Brownian motion date to 1960 and earlier [29].[21] Stratonovich developed a version of the Itô integral which obeys the usual Riemann-Steiltjes change of variables formula, but sacrifices the martingale property as well as much of the generality of the Itô integral.[22] [80] While popular

volume of the Séminaire, with a small amount of commentary as well [23].

[21] Girsanov's work extends the much earlier work first of Cameron and Martin [8], who in 1949 transformed Brownian paths for both deterministic translations and also some random translations, keeping the old and new distributions of the processes equivalent (in the sense of having the same sets of probability zero); these ideas were extended to Markov processes first by Maruyama [48] in 1954, and then by Girsanov in 1960. It was not until 1974 that Van Schuppen and Wong [82] extended these ideas to martingales, followed in 1976 by P. A. Meyer [58] and in 1977 Lenglart [43] for the current modern versions. See also (for example) pages 132–136 of [67] for an exposition of the modern results.

[22] Indeed, the Stratonovich integral was not met with much excitement. In a book review of the

in some engineering circles, the Stratonovich integral seemed to be primarily a curiosity, until much later when it was shown that if one approximates the paths of Brownian motion with differentiable curves, the resultant integrals converge to the Stratonovich integral; this led to it being an intrinsic object in stochastic differential geometry (see, e.g., [22]).

The primary works of interest in the Soviet Union were the series of articles of Skorohod. Again mainly inspired by the developing theory of Markov processes, Skorohod generalized the Itô integral in ways startlingly parallel to those of Courrège and Kunita and Watanabe. In 1963 Skorohod, squarely in the framework of Markov processes and clearly inspired by the work of Dynkin, developed a stochastic integral for martingales which is analogous to what Courrège had done in France, although he used changes of time [76]. In 1966, while studying additive functionals of continuous Markov processes, he developed the idea of quadratic variation of martingales, as well as what is now known as the Kunita-Watanabe inequality, and the same change of variables formula that Kunita and Watanabe established [77]. He extended his results and his change of variables formula to martingales with jumps (always only those defined on Markov processes) in 1967 [79]. The jump terms in the change of variables formula are expressed with the aid of a kernel reminiscent of the Lévy systems of S. Watanabe.[23]

We close this short history with a return to France. After the paper of Kunita and Watanabe, and after P. A. Meyer's four papers extending their results, there was a hiatus of three years before the paper of C. Doléans-Dade and P. A. Meyer appeared [15]. Prior to this paper the development of stochastic integration had been tied rather intimately to Markov processes, and was perhaps seen as a tool with which one could more effectively address certain topics in Markov process theory. A key assumption made by the prior work of H. Kunita and S. Watanabe, and also of P. A. Meyer, was that the underlying filtration of σ algebras was *quasi left continuous*, alternatively stated as saying that the filtration had no fixed times of discontinuity. Doléans-Dade and Meyer were able to remove this hypothesis, thus making the theory a purely martingale theory, and casting aside its relation to Markov processes. This can now be seen as a key step that led to the explosive growth of the theory in the 1970's and also in finance to the fundamental papers of Harrison-Kreps and Harrison-Pliska, towards the end of the next decade. Last, in this same paper Doléans-Dade and Meyer coined the modern term *semimartingale*, to signify the most general process for which one knew (at that time) there existed a stochastic integral.[24]

---

time Skorohod wrote "The proposed integral, when it exists, may be expressed rather simply using the Itô integral. However the class of functions for which this integral exists is extremely narrow and artificial. Although some of the formulas are made more simple by using the symmetrized integral (while most of them are made more complicated which will be made clear from what follows), its use is extremely restricted by its domain of definition. Thus this innovation is completely unjustified."[78] The Stratonovich integral was developed simultaneously by D. Fisk in the United States, as part of his PhD thesis. However it was rejected for publication as being too trivial. In the second half of his thesis he invents quasimartingales, and that half was indeed published [28].

[23]P. A. Meyer's work ([54, 55, 56, 57]), which proved to be highly influential in the West, references Courrège, Motoo and Watanabe, Watanabe, and Kunita and Watanabe, but not Skorohod, of whose work Meyer was doubtless unaware. Unfortunately this effectively left Skorohod's work relatively unknown in the West for quite some time.

[24]As we will see in a sequel to this paper, the description of semimartingales of Doléans-Dade and Meyer of 1970 turned out to be prescient. In the late 1970's C. Dellacherie and K. Bichteler simultaneously proved a characterization of semimartingales: they showed that given a right continuous process X with left limits, if one defined a stochastic integral in the obvious way on simple predictable processes, and if one insisted on having an extremely weak version of a bounded convergence theorem, then X was *a fortiori* a semimartingale.

Acknowledgements

The authors are grateful for help with this short history given by H. Föllmer, K. Itô,[25] J. Jacod, J. Pitman, H. Rubin, A. N. Shiryaev, S. Watanabe, M. Yor, and M. Zakai.

References

[1] Bachelier, L. (1900). *Théorie de la Spéculation*, Annales Scientifiques de l'École Normale Supérieure, 21–86.

[2] Bachelier, L. (1900). *Théorie de la Spéculation*, Gauthier-Villars, Paris. {Note: This book has been reprinted by the Paris publisher Éditions Jacques Gabay (1995).} MR1397712

[3] Benveniste, A. and Jacod, J. (1973). *Systèmes de Lévy des processus de Markov*, *Invent. Math.*, **21**, 183–198. MR343375

[4] Bernstein, P. L. (1992). *Capital Ideas: The Improbable Origins of Modern Wall Street*, The Free Press, New York.

[5] Bernstein, S. (1938). *Equations différentielles stochastiques*, *Actualités Sci. Ind.*, **738**, 5–31.

[6] Bru, B. and Yor, M. (2001). *La vie de W. Doeblin et le Pli cacheté 11 668*, La Lettre de L'Académie des Sciences, **2**, 16–17.

[7] Bru, B. and Yor, M. (2002). *Comments on the life and mathematical legacy of Wolfgang Doeblin*, Finance and Stochastics **6**, 3–47. MR1885582

[8] Cameron, R. H. and Martin, W. T. (1949). *Transformation of Wiener integrals by non-linear transformations*, Transactions of the American Math. Society **66**, 253–283. MR31196

[9] Chung, K. L. and Williams, R. (1990). *Introduction to Stochastic Integration, Second Edition*, Birkhäuser, Boston. MR1102676

[10] Courrège, Ph. (1963). *Intégrales stochastiques et martingales de carré intégrable*, *Séminaire Brélot-Choquet-Dény (Théorie du Potentiel)*, **7e année, 1962/63**, 7-01–7-20.

[11] Courtault, J.-M., Kabanov, Y., Bru, B., Crépel, P., Lebon, I., and Le Marchand, A. (2000). *Louis Bachelier: On the Centenary of Théorie de la Spéculation*, *Mathematical Finance* **10**, 341–353. MR1800320

[12] Cox, J. and Ross, S. A. (1976). *The Valuation of Options for Alternative Stochastic Processes*, *Journal of Financial Economics*, **3** (1/2), 145–166.

[13] Csörgo. (1998). *Random walking around financial mathematics*, Random walks *(Budapest)*, edited by Pál Révész, Bálint Tóth. *Bolyai Soc. Math. Stud.*, **9**, 59–111. MR1752891

[14] Dambis, K. E. (1965). *On the decomposition of continuous martingales*, *Theor. Proba. Applications*, **10**, 401–410. MR202179

[25] The second author wishes to express his thanks to Junko Itô who facilitated communication with K. Itô in the preparation of [66], and this information was also used in this article.

[15] Doléans-Dade, C. and Meyer, P. A. (1970). *Intégrales stochastiques par rapport aux martingales locales*, *Séminaire de Probabilités IV*, *Lecture Notes in Mathematics*, **124** 77–107. MR270425

[16] Doob, J. L. (1953). *Stochastic Processes*, John Wiley and Sons, New York. MR58896

[17] Doob, J. L. (1996). *The Development of Rigor in Mathematical Probability (1900–1950)*, in J.-P. Pier, ed., *Development of Mathematics 1900–1950*, Birkhauser Verlag AG, Basel. MR1404084

[18] Dubins, L. and Schwarz, G. (1965). *On continuous martingales*, *Proc. National Acad. Sciences USA*, **53**, 913–916. MR178499

[19] Dynkin, E. (1960). *Theory of Markov Processes*, Pergamon Press, Oxford. MR193669

[20] Dynkin, E. (1965). *Markov Processes* (two volumes) Springer-Verlag, Berlin, 1965. MR193671

[21] Einstein, A. (1905). *On the movement of small particles suspended in stationary liquid demanded by the molecular-kinetic theory of heat*, *Ann. d. Physik* **17** {In *Investigations of the theory of Brownian movement*, ed. R. Fürth, Dover, New York, 1956}.

[22] Emery, M. (1989). *Stochastic calculus in manifolds, with an appendix by P. A. Meyer*, Springer-Verlag, Berlin. MR1030543

[23] Emery, M. and Yor, M., eds. (2002). *Séminaire de Probabilités 1967–1980: A Selection in Martingale Theory*, *Lecture Notes in Mathematics*, **1771**. MR1925827

[24] Fama, E. (1965). *The Behavior of Stock Prices*, *Journal of Business*, **38** 34–105.

[25] Fama, E. (1998). *Market Efficiency, Long Term Returns, and Behavioral Finance, Journal of Financial Economics*, **49**, 283–306.

[26] Feller, W. (1936). *Zur Theorie der Stochastichen Prozesse (existenz-und Eindeutigkeitssatze)*, *Math. Ann.* **113**.

[27] Fisk, D. (1963). *Quasi-martingales and stochastic integrals*, Ph.D. thesis, Michigan State University, Department of Statistics.

[28] Fisk, D. (1965). *Quasimartingales*, *Transactions of the American Math. Soc.*, **120**, 369–389. MR192542

[29] Girsanov, I. V. (1960). *On transforming a certain class of stochastic processes by absolutely continuous changes of measures*, *Theory Proba. Appl.*, **5**, 285–301. MR133152

[30] Hald, A. (1981). *T. N. Thiele's contributions to Statistics*, *International Statistic Review* **49**, 1–20. MR623007

[31] Hunt, G. A. (1957) *Markoff processes and potentials I*, *Illinois J. Math.* **1**, 44–93. MR91349

[32] Hunt, G. A. (1957). *Markoff processes and potentials II*, Illinois J. Math. **1**, 316–369. MR91349

[33] Hunt, G. A. (1958). *Markoff processes and potentials III*, Illinois J. Math. **2**, 151–213. MR107097

[34] Itô, K. (1944). *Stochastic Integral*, Proc. Imp. Acad. Tokyo **20**, 519–524. MR14633

[35] Itô, K. (1951). *On a formula concerning stochastic differentials*, Nagoya Math. J. **3**, 55–65. MR44063

[36] Itô, K. (1951). *Multiple Wiener integral*, J. Math. Society of Japan **3**, 157–169. MR44064

[37] Itô, K. (1987). *Foreword*, K. Itô Collected Papers, Springer-Verlag, Heidelberg, xiii–xvii. MR931775

[38] Itô, K. and McKean, H. P., Jr. (1965). *Diffusion Processes and Their Sample Paths*, Springer-Verlag, New York; new edition by Springer-Verlag, 1996.

[39] Itô, K. and Watanabe, S. (1965). *Transformation of Markov processes by multiplicative functionals*, J. Math. Kyoto Univ. **4**, 1–75. MR184282

[40] Johnson, G. and Helms, L. L. (1963). *Class (D) Supermartingales*, Bull. American Math. Society **69**, 59–62. MR142148

[41] Kolmogorov, A. N. (1931). *On Analytic Methods in Probability Theory*, in A. N. Shiryaev, ed., *Selected Works of A. N. Kolmogorov; Volume II: Probability Theory and Mathematical Statistics*, Kluwer, Dordrecht, 1992, 62–108. [Original: Uber die analytischen Methoden in der Wahrscheinlichkeitsrechnung, *Math. Ann.* **104**, 1931, 415–458.] MR1153022

[42] Kunita, H. and Watanabe, S. (1967). *On Square Integrable Martingales*, Nagoya Math. J. **30**, 209–245. MR217856

[43] Lenglart, E. (1977). *Transformation des martingales locales par changement absolument continu de probabilités*, Z. Wahrscheinlichkeitstheorie verw. Gebiete **39**, 65–70. MR448541

[44] Lévy, P. (1955). *W. Döblin (V. Doeblin) (1915-1940)* Revue d'Histoire des Sciences et de leurs Applications **8**, 107–115. MR72808

[45] Liptser, R. Sh., and Shiryaev, A. S.; A. B. Aries, translator, 1977, 1978, (2nd, revised and expanded edition, 2001) *Statistics of Random Processes*, Two volumes, Springer-Verlag, Heidelberg.

[46] Malkiel, B. G. (2003). *A Random Walk Down Wall Street*, 7th edition, WW Norton, New York.

[47] McKean, H. P., Jr. (1969). *Stochastic Integrals*, Academic Press, New York. MR247684

[48] Maruyama, G. (1954). *On the transition probability functions of Markov processes*, Nat. Sci. Rep. Ochanomizu Univ. **5**, 10–20. MR67400

[49] Maruyama, G. (1955). *Continuous time processes and stochastic equations*, *Rend. Circ. Math. Palermo* **4**, 1–43. MR71666

[50] Merton, R. C. (1990). *Continuous Time Finance*, Basil Blackwell, Cambridge, Massachusetts.

[51] Merton, R. C. (2002). *Future possibilities in Finance Theory and Finance practice, Mathematical Finance – Bachelier Congress 2000*, eds. Geman, H., Madan, D., Pliska, S. R., and T. Vorst; Springer-Verlag, Heidelberg, 47–73.

[52] Meyer, P. A. (1962). *A decomposition theorem for supermartingales*, *Ill. J. Math.* **6**, 193–205. MR159359

[53] Meyer, P. A. (1963). *Decomposition of supermartingales: the uniqueness theorem*, *Ill. J. Math.* **7**, 1–17. MR144382

[54] Meyer, P. A. (1967). *Intégrales Stochastiques I*, *Séminaire de Probabilités I*, *Lecture Notes in Mathematics*, **39**, 72–94. MR231445

[55] Meyer, P. A. (1967). *Intégrales Stochastiques II*, *Séminaire de Probabilités I*, *Lecture Notes in Mathematics*, **39**, 95–117. MR231445

[56] Meyer, P. A. (1967). *Intégrales Stochastiques III*, *Séminaire de Probabilités I*, *Lecture Notes in Mathematics*, **39**, 118–141. MR231445

[57] Meyer, P. A. (1967). *Intégrales Stochastiques IV*, *Séminaire de Probabilités I*, *Lecture Notes in Mathematics*, **39**, 142–162. MR231445

[58] Meyer, P. A. (1976). *Un cours sur les intégrales stochastiques*, *Séminaire de Probabilités X*, *Lecture Notes in Mathematics*, **511**, 246–400. MR501332

[59] Meyer, P. A. (2000). *Les Processus Stochastiques de 1950 à Nos Jours*, in *Development of Mathematics 1950-2000*, edited by Jean-Paul Pier; Birkhäuser, Boston, MA. 813–848. MR1796860

[60] Modigliani, F. and Miller, M. H. *The Cost of Capital, Corporation Finance, and the Theory of Investment*, *American Economic Review*, **48**, 261–297.

[61] Motoo, M. and Watanabe, S. (1965). *On a class of aditive functionals of Markov process*, *J. Math. Kyoto Univ.* **4**, 429–469. MR196808

[62] Osborne, M. F. M. (1959). *Brownian motion in the stock market*, Operations Research, **7**, 145–173. MR104513

[63] Orey, S. (1965). *F-processes*, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **2** 301–313, University of California Press, Berkeley. MR214124

[64] Pais, A. (1982). *'Subtle is the Lord. . . ' The Science and Life of Albert Einstein*, Oxford University Press, Oxford. MR690419

[65] Petit, M. (2003). *L'équation de Kolmogoroff*, Editions Ramsay, Paris.

[66] Protter, P. (2003). *A new prize in honor of Kiyosi Itô*, *Stochastic Processes and their Applications*, **108**, 151–153. MR2016970

[67] Protter, P. (2004). *Stochastic Integration and Differential Equations; Second Edition*, Springer Verlag, Heidelberg. MR2020294

[68] Rao, K. M. (1969). *Quasimartingales*, *Math. Scand.*, **24**, 79–92. MR275511

[69] Rubin, H. (2003). Personal communication by electronic mail.

[70] Rubin, H. (1956). *Quasi-martingales and stochastic integrals*, title of an invited talk at the Seattle Meeting of the IMS, August 21–24, 1956; see page 1206 of the *Annals Math. Statist.*, **27**, 1198–1211.

[71] Samuelson, P. (1965). *Rational Theory of Warrant Pricing*, Industrial Management Review, **6**, 13–39.

[72] Samuelson, P. (1965). *Proof That Properly Anticipated Prices Fluctuate Randomly*, Industrial Management Review, **6**, 41–49.

[73] Samuelson, P. (1973). *Mathematics of Speculative Price*, *SIAM Review*, **15**, 1–42. MR323315

[74] Samuelson, P. (2002). *Modern finance theory within one lifetime*, Mathematical Finance – Bachelier Congress 2000, eds. Geman, H., Madan, D., Pliska, S. R., and T. Vorst; Springer-Verlag, Heidelberg, 41–46. MR1960557

[75] Samuelson, P. and Merton, R. C. (1969). *A Complete Model of Warrant Pricing that Maximizes Utility,* Industrial Management Review, **10**(2), 17–46.

[76] Skorokhod, A. V. (1963). *On homogeneous continuous Markov proceses that are martingales*, Theory of Probability and its Applications, **8**, 355–365. MR158432

[77] Skorokhod, A. V. (1966). *On the local structure of continuous Markov processes*, Theory of Probability and its Applications, **11**, 336–372. MR203815

[78] Skorokhod, A. V. (1967). Review of R. L. Stratonovich, *Conditional Markov Processes and Their Application to the Theory of Optimal Control*, *Theory of Probability and its Applications*, **12**, 154–156.

[79] Skorokhod, A. V. (1967). *Homogeneous Markov processes without discontinuities of the second kind*, Theory of Probability and its Applications, **12**, 222–240. MR230372

[80] Stratonovich, R. L. (1966). *Conditional Markov Processes and Their Application to the Theory of Optimal Control*, Izd. Moscow University Press, Moscow. MR197209

[81] Thiele, T. N. (1880). *Sur la compensation de quelques erreurs quasi-systématiques par la méthode des moindres carrés*, Reitzel, Copenhagen. {Note: This article was published simultaneously in Danish and French; for the Danish reference see [30].}

[82] Van Schuppen, J. H. and Wong, E. (1974). *Transformations of local martingales under a change of law*, Annals of Probability **2**, 879–888. MR358970

[83] Varadhan, S. R. S. and Stroock, D. W. (1987). *Introduction*, *K. Itô Collected Papers*, Springer-Verlag, Heidelberg, vii–xii. MR868862

[84] Ville, J. (1939). *Étude critique de la notion de collectif*, Gauthier-Villars, Paris.

[85] Watanabe, S. (1964). *On discontinuous additive functionals and Lévy measures of a Markov process*, Japanese J. Math., **36**, 53–70. MR185675

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 92–97
© Institute of Mathematical Statistics, 2004

Non-linear filtering with Gaussian martingale noise: Kalman filter with fBm noise

L. Gawarecki[1] and V. Mandrekar[2]

Kettering University

Michigan State University

Abstract: We consider non-linear filtering problem with Gaussian martingales as a noise process, and obtain iterative equations for the optimal filter. We apply that result in the case of fractional Browian motion noise process and derive Kalman type equations in the linear case.

1. Introduction

The study of filtering of a stochastic process with a general Gaussian noise was initiated in [8]. In case the system satisfies a stochastic differential equation, we derived an iterative form for the optimal filter given by the Zakai equation ([3]). It was shown in [2] that in the case of a Gaussian noise, one can derive the FKK equation from which one can obtain the Kalman filtering equation. However in order to obtain Kalman's equation in the case of fractional Brownian motion (fBm) noise, we had to assume in [3] the form of the observation process, which was not intuitive. Using the ideas in [5], we are able to study the problem with a natural form of the observation process as in the classical work. In order to get such a result from the general theory we have to study the Bayes formula for Gaussian martingale noise and use the work in [5]. This is accomplished in Section 1. In Section 2, we obtain iterative equations for the optimal filter and in Section 3 we apply them to the case of fBm noise.

The problem of filtering with system and observation processes driven by fBm was considered in [1]. However, even the form of the Bayes formula in this case is complicated and no iterative equations for the filter can be obtained. The Bayes formula in [8] is applicable to any system process and observation process with Gaussian noise. In order to get iterative equations in non-linear case we assume that the system process is a solution of a martingale problem. This allows us to obtain an analogue of the Zakai and FKK equations. As a consequence, we easily derive the Kalman equations in the linear case. If the data about the "signal" are sent to the server and transmitted to AWACS, the resulting process has bursts [6]. We assume a particular form for this observation process (see equation (3)). In most cases, signal (missile trajectory, e.g.) is Markovian.

The work completed by D. Fisk under the guidence of Professor Herman Rubin has found applications in deriving filtering equations in the classical case [4].

[1]Department of Science and Mathematics, Kettering University, 1700 W. Third Ave., Flint, MI 48504, USA. e-mail: l.gawarec@kettering.edu

[2]Department of Statistics and Probability, Michigan State University, Wells Hall, East Lansing, MI 48824, USA. e-mail: atmash@stt.msu.edu

Keywords and phrases: non-linear filtering, Gaussian martingale noise process, Bayes fromula, FKK equation, fractional Browian motion, Kalman equations.

AMS 2000 subject classifications: 60G15, 60G35, 62M20, 93E11.

2. Bayes formula with Gaussian martingale noise

Let us consider the filtering problem with a signal or system process $\{X_t,\ 0 \le t \le T\}$, which is unobservable. Information about X_t is obtained by observing another process Y_t, which is a function of X_t, and is corrupted by noise, i.e.

$$Y_t = \beta(t, X) + N_t, \quad 0 \le t \le T,$$

where $\beta(t, \cdot)$ is measurable with respect to the σ–field $\mathcal{F}_t^X$, generated by the signal process $\{X_s,\ 0 \le s \le t\}$, and the noise $\{N_t,\ 0 \le t \le T\}$ is independent of $\{X_t,\ 0 \le t \le T\}$. The observation σ–field $\mathcal{F}_t^Y = \sigma\{Y_s,\ 0 \le s \le t\}$ contains all the available information about the signal X_t. The primary aim of filtering theory is to get an estimate for X_t based on the σ–field $\mathcal{F}_t^Y$. This is given by the conditional distribution $\hat{\Pi}_t$ of X_t given $\mathcal{F}_t^Y$ or, equivalently, by the conditional expectation $E\left(f(X_t)\,\middle|\,\mathcal{F}_t^Y\right)$ for a rich enough class of functions f. Since this estimate minimizes the squared error loss, $\hat{\Pi}_t$ is called the optimal filter.

In [8] an expression for an optimal filter was given for $\{N_t,\ 0 \le t \le T\}$, a Gaussian process and $\beta(\cdot, X) \in H(R)$, the reproducing kernel Hilbert space (RKHS) of the covariance R of the process N_t ([8]). Throughout we assume, without loss of generality, that $E(N_t) = 0$.

Let us assume that $N_t = M_t$, a continuous Gaussian martingale with the covariance function R_M. We shall first compute the form of $H(R_M)$. As we shall be using this notation exclusively for the martingale M_t, we will drop the subscript M from now on and denote the RKHS of R by $H(R)$. Let us also denote by $m(t)$ the expectation EM_t^2. Note that $m(t)$ is a non–decreasing function on $[0, T]$ and, abusing the notation, we will denote by m the associated measure on the Borel subsets $\mathcal{B}([0, T])$. With this convention, we can write

$$H(R) = \left\{ g :\ g(t) = \int_0^t g^*(u)\, dm(u),\ \ 0 \le t \le T,\ g^* \in L^2(m) \right\}.$$

The scalar product in $H(R)$ is given by $(g_1, g_2)_{H(R)} = < g_1^*, g_2^* >_{L^2(m)}$. If we denote by $H(R : t)$ the RKHS of $R|_{[0,t]\times[0,t]}$, then it follows from the above that

$$H(R : t) = \left\{ g :\ g(s) = \int_0^s g^*(u)\, dm(u),\ \ 0 \le s \le t,\ g^* \in L^2(m) \right\}.$$

It is well known (see [8], Section 2), that there exists an isometry π between $H(R)$ and $\overline{sp}^{L^2}\{M_t,\ \ 0 \le t \le T\}$, which, in case M is a martingale, is given by

$$\pi(g) = \int_0^T g^*(u)\, dM_u,$$

where the RHS denotes the stochastic integral of the deterministic function g^* with respect to M. The isometry

$$\pi_t(g) : H(R : t) \to \overline{sp}^{L^2}\{M_s,\ \ 0 \le s \le t\}$$

is given by $\pi_t(g) = \int_0^t g^*(u) dM_u$.

Suppose now

$$Y_t = \int_0^t h(s, X) dm(s) + M_t,$$

where $h(s,X)$ is $\mathcal{F}_s^X$-measurable and $h(\cdot,X) \in L^2(m)$. Then using Theorem 3.2 of [8] we get the Bayes formula for an $\mathcal{F}_T^X$-measurable and integrable function $g(T,X)$

$$E\left(g(T,X)\,\middle|\,\mathcal{F}_t^Y\right) = \frac{\int g(T,\mathbf{x}) e^{\int_0^t h(s,\mathbf{x})dY_s - \frac{1}{2}\int_0^t h^2(s,\mathbf{x})\,dm(s)}\,dP\circ X^{-1}}{\int e^{\int_0^t h(s,\mathbf{x})\,dY_s - \frac{1}{2}\int_0^t h^2(s,\mathbf{x})\,dm(s)} dP\circ X^{-1}}. \tag{1}$$

3. Equations for non-linear filter with martingale noise

In this section we derive the Zakai equation for the so-called "unconditional" measure-valued process. We follow the techniques developed in [2]. We assume that $\{X_t,\ \ 0 \le t \le T\}$ is a solution of the martingale problem. Let $C_c^2(\mathbb{R}^n)$ be the space of twice continuously differentiable functions with compact support. Let

$$(L_t f)(x) = \sum_{j=1}^n b_j(t,x)\frac{\partial f}{\partial x_j}(x) + 1/2\sum_{i,j=1}^n \sigma_{i,j}(t,x)\frac{\partial^2 f}{\partial x_i x_j}(x),$$

for $f \in C_c^2(\mathbb{R}^n)$, with $b_j(t,x)$ and $\sigma_{i,j}(t,x)$ bounded and continuous. We assume that X_t is a solution to the martingale problem, i.e., for $f \in C_c(\mathbb{R}^n)$,

$$f(X_t) - \int_0^t (L_u f)(X_u)\,du$$

is an $\mathcal{F}_t^X$–martingale with respect to the measure P. Consider the probability space $(\Omega\times\Omega, \mathcal{F}\otimes\mathcal{F}, P\otimes P')$, where P' is a probability measure given by

$$dP' = \exp\left(-\int_0^t h(s,X)dY_s + \frac{1}{2}\int_0^t h^2(s,X)dm(s)\right)\,dP.$$

Then under the measure P', the process Y_t has the same distribution as M_t and is independent of X_t. In addition, $P\circ X^{-1} = P'\circ X^{-1}$. This follows from Theorem 3.1 in [8]. Define

$$\alpha_t(\omega',\omega) = \exp\left(\int_0^t h\left(s,X(\omega')\right)\,dY_s(\omega) - 1/2\int_0^t h^2\left(s,X(\omega')\right)\,dm(s)\right).$$

Then, with a notation $g(\omega') = g(T,X(\omega'))$, equation (1) can be written as

$$E\left(g(T,X)\,\middle|\,\mathcal{F}_t^Y\right) = \frac{\int g(\omega')\alpha_t(\omega',\omega)\,dP\circ X^{-1}(\omega')}{\int \alpha_t(\omega',\omega)\,dP\circ X^{-1}(\omega')}.$$

For a function $f \in C_c^2(\mathbb{R}^n)$, denote

$$\hat{\sigma}_t(f,Y)(\omega) = \int f\left(X_t(\omega')\right)\alpha_t(\omega',\omega)\,dP(\omega').$$

Then we get the following analogue of the Zakai equation. We assume here that m is mutually absolutely continuous with respect to the Lebesgue measure.

Theorem. *The quantity $\hat{\sigma}_t(f,Y)$ defined above satisfies the equation*

$$d\hat{\sigma}_t\left(f(\cdot),Y\right) = \hat{\sigma}_t\left(L_t f(\cdot),Y\right)\,dt + \hat{\sigma}_t\left(h(t,\cdot)f(\cdot),Y\right)\,dY_t.$$

Proof. We follow the argument as in [2]. Consider $g_t(\omega') = f(X_T(\omega')) - \int_t^T (L_s f)(X_s(\omega'))\,ds$, with $f \in C_c^2(\mathbb{R}^n)$. Then

$$E_P\left(g_t \left| \mathcal{F}_t^X \right.\right) = f(X_t), \quad 0 \le t \le T.$$

We can represent $\hat{\sigma}_t(f, Y)$ as

$$\begin{aligned}
\hat{\sigma}_t(f,Y) &= \int f\left(X_t\left(\omega'\right)\right)\alpha_t(\omega',\omega)\,dP(\omega') \\
&= \int E_P\left(g_t(\omega')\alpha_t(\omega',\omega)\left|\mathcal{F}_t^X\right.\right)\,dP(\omega') \\
&= E_P\left(g_t(\omega')\alpha_t(\omega',\omega)\right) \\
&=: \sigma_t'\left(g_t, Y\right).
\end{aligned}$$

By definition of g_t,

$$dg_t = \left(L_t f\right)\left(X_t'\right)\,dt,$$

with X_t' an independent copy of X_t as a function of ω'. Using Itô's formula,

$$d\alpha_t = \alpha_t h\left(t, X'\right)\,dY_t.$$

Since $\sigma_t'\left(g_t, Y\right) = E_P(g_t\alpha_t)$, utilizing the Fubini theorem and Theorem 5.14 in [7], we rewrite the latter as

$$\begin{aligned}
E_P\left(g_t\alpha_t\right) &= E_P g_0 + \int_0^t \hat{\sigma}_s\left(L_s f, Y\right)\,ds \\
&\quad + \int\int_0^t g_s(\omega')\alpha_s(\omega',\omega)h\left(s, X(\omega')\right)\,dY_s(\omega)\,dP(\omega') \\
&= E_P g_0 + \int_0^t \hat{\sigma}_s\left(L_s f, Y\right)\,ds \\
&\quad + \int_0^t \hat{\sigma}_s\left(h\left(s, X(\omega')\right) f\left(X_s(\omega')\right), Y\right)\,dY_s.
\end{aligned}$$

It should be noted the application of Theorem 5.14 above is valid due to the fact that the martingale M_t is a time changed Brownian motion with non–singular time. □

Now we note that the optimal filter is given by

$$\hat{\Pi}_t(f) = E\left(f(X_t)\left|\mathcal{F}_t^Y\right.\right) = \frac{\hat{\sigma}(f,Y)}{\hat{\sigma}(1,Y)}.$$

Under our construction, Y_t is a continuous Gaussian martingale with the increasing process $m(t)$. Using Itô's formula we obtain

$$d\hat{\Pi}_t(f) = \hat{\Pi}_t(L_t f)dt + \left[\hat{\Pi}_t(hf) - \hat{\Pi}_t(f)\hat{\Pi}_t(h)\right]d\nu_t, \tag{2}$$

where $\nu_t = Y_t - \int_0^t \hat{\Pi}_s(h)\,dm(s)$.

4. Filtering equations in case of fractional Brownian motion noise

Let us start with the definition of fractional Brownian motion (fBm). We say that a Gaussian process $\{W_t^H,\ 0 \le t \le T\}$ on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, with continuous sample paths is a fractional Brownian motion if $W_0^H = 0$, $EW_t^H = 0$, and for $0 < H < 1$,

$$EW_s^H W_t^H = \frac{1}{2}[s^{2H} + t^{2H} - |s-t|^{2H}], \quad 0 \le s, t \le T.$$

Let us set up some notation following [5].

$$k_H(t,s) = \kappa_H^{-1} s^{1/2-H}(t-s)^{1/2-H}, \quad \text{where } \kappa_H = 2H\Gamma(3/2-H)\Gamma(H+1/2),$$
$$w_t^H = \lambda_H^{-1} t^{2-2H}, \quad \text{with } \lambda_H = \frac{2H\Gamma(3-2H)\Gamma(H+1/2)}{\Gamma(3/2-H)},$$
$$M_t^H = \int_0^t k_H(t,s)dW_s^H.$$

The integral with respect to fBm W_t^H is described in [9]. The process M_t^H is a Gaussian martingale. Define

$$Q_H^c(t) = \frac{d}{dw_t^H}\int_0^t k_H(t,s)C(s)\,ds,$$

where $C(t)$ is an $\mathcal{F}_t$–adapted process and the derivative is understood to be in terms of absolute continuity. Then the following result can be derived from [5].

Let $Y_t = \int_0^t C(s,X)\,ds + W_t^H$. Then

$$Z_t = \int_0^t Q_H^c(s)\,dw_s^H + M_t^H$$

is an $\mathcal{F}_t^Y$ semi–martingale and $\mathcal{F}_t^Y = \mathcal{F}_t^Z$. Let us now consider the filtering problem as in Section 1, with the noise $N_t = W_t^H$, and the observation process

$$Y_t = \int_0^t C(s,X)\,ds + W_t^H. \tag{3}$$

Then the equivalent filtering problem is given by the system process X_t and the observation process

$$Z_t = \int_0^t Q_H^c(s,X)\,dw_s^H + M_t^H.$$

Using results of Section 2, and assuming that X_t is a solution to the martingale problem, equation (2) reduces to

$$d\hat{\Pi}_t(f) = \hat{\Pi}_t(L_t f)dt + \left[\hat{\Pi}_t(Q_H^c f) - \hat{\Pi}_t(f)\hat{\Pi}_t(Q_H^c)\right]d\nu_t,$$

where $\nu_t = Z(t) - \int_0^t \hat{\Pi}_s(Q_H^c)\,dw_s^H$. By Theorem 2 in [5] we get that ν_t is a continuous Gaussian $\mathcal{F}_t^Y$–martingale with variance w_t^H.

Let us now assume that the system process and observation processes are given by

$$\begin{aligned} X_t &= \int_0^t b(u)X_u\,du + \int_0^t \sigma(u)\,dW_u \\ Y_t &= \int_0^t c(u)X_u\,du + W_t^H, \end{aligned}$$

where the processes W_t and W_t^H are independent. Because (X_t, Z_t) is jointly Gaussian we get

$$\begin{aligned} &\hat{\Pi}_t(X_tX_s) - \hat{\Pi}_t(X_t)\hat{\Pi}_t(X_s) \\ &\quad = E\left\{(X_t - \hat{\Pi}_t(X_t))(X_s - \hat{\Pi}_t(X_s))\,\middle|\,\mathcal{F}_t^Y\right\} \\ &\quad = E\left\{(X_t - \hat{\Pi}_t(X_t))(X_s - \hat{\Pi}_t(X_s))\right\} \\ &\quad = \Gamma(t,s). \end{aligned}$$

We obtain that

$$d\hat{\Pi}_t(X_t) = b(t)\hat{\Pi}_t(X_t)dt + \int_0^t k_H(t,s)\Gamma(t,s)\,ds\,d\nu_t. \tag{4}$$

Denote by $\gamma(t) = EX_t^2$, and $F(t) = E\left(\hat{\Pi}_t^2(X_t)\right)$. Then by the Itô formula for $f(x) = x^2$ and by taking the expectation, we get

$$\begin{aligned} d\gamma(t) &= 2b(t)\gamma(t)dt + \sigma^2(t)dt \\ \text{and} \quad dF(t) &= 2b(t)F(t)dt + \left(\int_0^t k_H(t,s)\Gamma(t,s)\,ds\right)^2 dw_t^H. \end{aligned}$$

Let us consider

$$\begin{aligned} \Gamma(t,t) &= E(X_t - \hat{\Pi}(X_t))^2 \\ &= E(X_t^2) - E(\hat{\Pi}_t^2(X_t)) \\ &= \gamma(t) - F(t). \end{aligned}$$

Then we arrive at

$$d\Gamma(t,t) = 2b(t)\Gamma(t,t)dt + \sigma^2(t)dt - \left(\int_0^t k_H(t,s)\Gamma(t,s)\,ds\right)^2 dw_s^H. \tag{5}$$

For $H = \frac{1}{2}$ this reduces to the Kalman equation.

Equations (4) and (5) give the Kalman filtering equations in the linear case.

References

[1] L. Coutin, L. Decreusefond, Abstract Nonlinear Filtering Theory in the Presence of Fractional Brownian Motion, *The Ann. Appl. Probab.* **9**, No. 4 (1999) 1058–1090. MR1728555

[2] L. Gawarecki and V. Mandrekar, Remark on "Instrumentation Problem" of A. V. Balakrishnian, *Journal of the Indian Statistical Association* (to appear).

[3] L. Gawarecki and V. Mandrekar, On the Zakai Equation of Filtering with Gaussian Noise, *Stochastics in Finite and Infinite Dimensions*, Trends in Mathematics, Birkhäuser, (2001) 145–151. MR1797085

[4] G. Kallianpur, C. Striebel, A Stochastic Differential Equation of Fisk Type for Estimation and Nonlinear Filtering Problems, *SIAM J. of Appl. Math.* **21** (1971) 61–72. MR297032

[5] M. L. Kleptsyna, A. Le Breton, M. C. Roubaud, Parameter Estimation and Optimal Filtering for Fractional Type Stochastic Systems, *Statistical Inference for Stochastic Processes* **3** (2000) 173–182. MR1819294

[6] W. E. Leland, M. Taqqu, W. Willinger, D.V. Wilson, On Self–similar Nature of Ethernet Traffic (extended version), *IEEE/ACM Trans. Networking*, **2** (1994) 1–15.

[7] R.S. Liptser and A.N. Shiryaev, Statistics of random processes, Vol. 1, Springer Verlag, N.Y. (1977).

[8] P. Mandal and V. Mandrekar, A Bayes formula for Gaussian noise processes and its applications, *SIAM J. of Control and Optimization* **39** (2000) 852–871. MR1786333

[9] I. Norros, E. Valkeila and J. Virtamo, An elementary approach to Girsanov formula and other analytical results on fractional Brownian motion, *Bernoulli* **5** (1999) 571–587. MR1704556

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 98–125
© Institute of Mathematical Statistics, 2004

Self-similar processes, fractional Brownian motion and statistical inference

B.L.S. Prakasa Rao[1]

Indian Statistical Institute (Delhi Centre)

Abstract: Self-similar stochastic processes are used for stochastic modeling whenever it is expected that long range dependence may be present in the phenomenon under consideration. After discussing some basic concepts of self-similar processes and fractional Brownian motion, we review some recent work on parametric and nonparametric inference for estimation of parameters for linear systems of stochastic differential equations driven by a fractional Brownian motion.

1. Introduction

"Asymptotic Distributions in Some Nonregular Statistical Problems" was the topic of my Ph.D. Dissertation prepared under the guidance of Prof. Herman Rubin at Michigan State University in 1966. One of the nonregular problems studied in the dissertation was the problem of estimation of the location of cusp of a continuous density. The approach adapted was to study the limiting distribution if any of the log-likelihood ratio process and then obtain the asymptotic properies of the maximum likelihood estimator. It turned out that the limiting process is a special type of a nonstationary gaussian process. The name fractional Brownian motion was not in vogue in those years and the limiting process is nothing but a functional shift of a fractional Brownian motion. Details of these results are given in Prakasa Rao (1966) and Prakasa Rao (1968). The other nonregular problems discussed in the dissertation dealt with inference under order restrictions where in it was shown that, for the existence of the limiting distribution if any for the nonparametric maximum likelihood density estimators under order restrictions such as unimodality of the density function or monotonicity of the failure rate function, one needs to scale the estimator by the cube root of n, the sample size rather than the square root of n as in the classical parametric inference (cf. Prakasa Rao (1969, 1970). These type of asymptotics are presently known as cube root asymptotics in the literature. It gives me a great pleasure to contribute this paper to the festschrift in honour of my "guruvu" Prof. Herman Rubin.

A short review of some properties of self-similar processes is given in the Section 2. Stochastic differential equations driven by a fractional Brownian motion (fBm) are introduced in the Section 3. Asymptotic properties of the maximum likelihood estimators and the Bayes estimators for parameters involed in linear stochastic differential equations driven by a fBm with a known Hurst index are reviewed in the Section 4. Methods for statistical inference such as the maximum likelihood estimation and the sequential maximum likelihood estimation are

[1]Indian Statistical Institute, 7, S. J. S. Sansanwal Marg, New Delhi, 110016. e-mail: blsp@isid.ac.in

Keywords and phrases: self-similar process, fractional Brownian motion, fractional Ornstein-Uhlenbeck type process, Girsanov-type theorem, maximum likelihood estimation, Bayes estimation, nonparametric inference, linear stochastic systems.

AMS 2000 subject classifications: primary, 62M09; secondary, 60G15.

discussed for the special case of the fractional Ornstein-Uhlenbeck type process and some new results on the method of minimum L_1-norm estimation are presented in the Section 5. Identification or nonparametric estimation of the "drift" function for linear stochastic systems driven by a fBm are studied in the Section 6.

2. Self-similar processes

Long range dependence phenomenon is said to occur in a stationary time series $\{X_n, n \geq 0\}$ if the $Cov(X_0, X_n)$ of the time series tends to zero as $n \to \infty$ and yet the condition

$$\sum_{n=0}^{\infty} |Cov(X_0, X_n)| = \infty \tag{2.1}$$

holds. In other words the covariance between X_0 and X_n tends to zero but so slowly that their sums diverge. This phenonmenon was first observed by hydrologist Hurst (1951) on projects involving the design of reservoirs along the Nile river (cf. Montanari (2003)) and by others in hydrological time series. It was recently observed that a similar phenomenon occurs in problems connected with traffic patterns of packet flows in high speed data net works such as the internet (cf. Willinger et al. (2003) and Norros (2003)). Long range dependence is also related to the concept of self-similarity for a stochastic process in that the increments of a self-similar process with stationary increments exhibit long range dependence. Long range dependence pattern is also observed in macroeconomics and finance (cf. Henry and Zafforoni (2003)). A recent monograph by Doukhan et al. (2003) discusses the theory and applications of long range dependence.

A real-valued stochastic process $Z = \{Z(t), -\infty < t < \infty\}$ is said to be *self-similar* with index $H > 0$ if for any $a > 0$,

$$\mathcal{L}(\{Z(at), -\infty < t < \infty\}) = \mathcal{L}(\{a^H Z(t), -\infty < t < \infty\}) \tag{2.2}$$

where $\mathcal{L}$ denotes the class of all finite dimensional distributions and the equality indicates the equality of the finite dimensional distributions of the process on the right side of the equation (2.2) with the corresponding finite dimensional distributions of the process on the left side of the equation (2.2). The index H is called the *scaling exponent* or the *fractal index* or the *Hurst parameter* of the process. If H is the scaling exponent of a self-similar process Z, then the process Z is called *H-self similar* process or H-ss process for short. It can be checked that a nondegenerate H-ss process cannot be a stationary process. In fact if $\{Z(t), t > 0\}$ is a H-ss process, then the process

$$Y(t) = e^{-tH} Z(e^t), -\infty < t < \infty \tag{2.3}$$

is a stationary process. Conversely if $Y = \{Y(t), -\infty < t < \infty\}$ is a stationary process, then $Z = \{t^H Y(\log t), t > 0\}$ is a H-ss process.

Suppose $Z = \{Z(t), -\infty < t < \infty\}$ is a H-ss process with finite variance and stationary increments, that is,

$$\mathcal{L}(Z(t+h) - Z(t)) = \mathcal{L}(Z(t) - Z(0)), -\infty < t, h < \infty. \tag{2.4}$$

Then the following properties hold:

(i) $Z(0) = 0$ a.s;

(ii) If $H \neq 1$, then $E(Z(t)) = 0, -\infty < t < \infty$;

(iii) $\mathcal{L}(Z(-t)) = \mathcal{L}(-Z(t))$;

(iv)$E(Z^2(t)) = |t|^{2H} E(Z^2(1))$;

(v)The covariance function $\Gamma_H(t,s)$ of the process Z is given by

$$\Gamma_H(t,s) = \frac{1}{2}\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\}. \tag{2.5}$$

(vi)The self-similarity parameter, also called the scaling exponent or fractal index H, is less than or equal to one.

(vii) If $H = 1$, then $Z(t) = tZ(1)$ a.s. for $-\infty < t < \infty$.

(viii) Let $0 < H \leq 1$. Then the function

$$R_H(s,t) = \{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\} \tag{2.6}$$

is nonnegative definite. For proofs of the above properties, see Taqqu (2003).

A gaussian process H-ss process $W^H = \{W_H(t), -\infty < t < \infty\}$ with stationary increments and with fractal index $0 < H < 1$ is called a *fractional Brownian motion* (fBm). It is said to be standard if $Var(W^H(1)) = 1$. For any $0 < H < 1$, there exists a version of the fBm for which the sample paths are continuous with probability one but are not differentiable even in the L^2-sense. The continuity of the sample paths follows from the Kolmogorov's continuity condition and the fact that

$$E|W^H(t_2) - W^H(t_1)|^\alpha = E|W^H(1)|^\alpha |t_2 - t_1|^{\alpha H} \tag{2.7}$$

from the property that the fBm is a H-ss process with stationary increments. We can choose α such that $\alpha H > 1$ to satisfy the Kolmogorov's continuity condition. Further more

$$E|\frac{W^H(t_2) - W^H(T_1)}{t_2 - t_1}|^2 = E[W^H(1)^2]|t_2 - t_1|^{2H-2} \tag{2.8}$$

and the last term tends to infinity as $t_2 \to t_1$ since $H < 1$. Hence the paths of the fBm are not L^2-differentiable. It is interesting to note that the fractional Brownian motion reduces to the Brownian motion or the Wiener process for the case when $H = \frac{1}{2}$.

As was mentioned above, self-similar processes have been used for stochastic modeling in such diverse areas as hydrology, geophysics, medicine, genetics and financial economics and more recently in modeling internet traffic patterns. Recent additional applications are given in Buldyrev et al. (1993), Ossandik et al. (1994), Percival and Guttorp (1994) and Peng et al.(1992, 1995a,b). It is important to estimate the constant H for modeling purposes. This problem has been considered by Azais (1990), Geweke and Porter-Hudak (1983), Taylor and Taylor (1991), Beran and Terrin (1994), Constantine and Hall (1994), Feuverger et al. (1994), Chen et al. (1995), Robinson (1995), Abry and Sellan (1996), Comte (1996), McCoy and Walden (1996), Hall et al. (1997), Kent and Wood (1997), and more recently in Jensen (1998), Poggi and Viano (1998) and Coeurjolly (2001).

It was observed that there are some phenomena which exhibit self-similar behaviour locally but the nature of self-similarity changes as the phenomenon evolves. It was suggested that the parameter H must be allowed to vary as function of time for modeling such data. Goncalves and Flandrin (1993) and Flandrin and Goncalves (1994) propose a class of processes which are called *locally self-similar* with dependent scaling exponents and discuss their applications. Wang et al. (2001) develop procedures using wavelets to construct local estimates for time varying scaling exponent $H(t)$ of a locally self-similar process.

3. Stochastic differential equations driven by fBm

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ be a stochastic basis satisfying the usual conditions. The natural fitration of a process is understood as the P-completion of the filtration generated by this process.

Let $W^H = \{W_t^H, t \geq 0\}$ be a normalized fractional Brownian motion (fBm) with Hurst parameter $H \in (0,1)$, that is, a gaussian process with continuous sample paths such that $W_0^H = 0, E(W_t^H) = 0$ and

$$E(W_s^H W_t^H) = \frac{1}{2}[s^{2H} + t^{2H} - |s-t|^{2H}], t \geq 0, s \geq 0. \tag{3.1}$$

Let us consider a stochastic process $Y = \{Y_t, t \geq 0\}$ defined by the stochastic integral equation

$$Y_t = \int_0^t C(s)ds + \int_0^t B(s)dW_s^H, t \geq 0 \tag{3.2}$$

where $C = \{C(t), t \geq 0\}$ is an $(\mathcal{F}_t)$-adapted process and $B(t)$ is a nonvanishing nonrandom function. For convenience, we write the above integral equation in the form of a stochastic differential equation

$$dY_t = C(t)dt + B(t)dW_t^H, t \geq 0 \tag{3.3}$$

driven by the fractional Brownian motion W^H. The integral

$$\int_0^t B(s)dW_s^H \tag{3.4}$$

is not a stochastic integral in the Ito sense but one can define the integral of a deterministic function with respect to the fBm in a natural sense (cf. Gripenberg and Norros (1996); Norros et al. (1999)). Even though the process Y is not a semimartingale, one can associate a semimartingale $Z = \{Z_t, t \geq 0\}$ which is called a *fundamental semimartingale* such that the natural filtration $(\mathcal{Z}_t)$ of the process Z coincides with the natural filtration $(\mathcal{Y}_t)$ of the process Y (Kleptsyna et al. (2000)). Define, for $0 < s < t$,

$$k_H = 2H\ \Gamma(\frac{3}{2} - H)\ \Gamma(H + \frac{1}{2}), \tag{3.5}$$

$$\kappa_H(t,s) = k_H^{-1} s^{\frac{1}{2}-H}(t-s)^{\frac{1}{2}-H}, \tag{3.6}$$

$$\lambda_H = \frac{2H\,\Gamma(3-2H)\ \Gamma(H+\frac{1}{2})}{\Gamma(\frac{3}{2}-H)}, \tag{3.7}$$

$$w_t^H = \lambda_H^{-1} t^{2-2H}, \tag{3.8}$$

and

$$M_t^H = \int_0^t \kappa_H(t,s) dW_s^H, t \geq 0. \tag{3.9}$$

The process M^H is a Gaussian martingale, called the *fundamental martingale* (cf. Norros et al. (1999)) and its quadratic variation $< M_t^H >= w_t^H$. Further more the natural filtration of the martingale M^H coincides with the natural fitration of the fBm W^H. In fact the stochastic integral

$$\int_0^t B(s) dW_s^H \tag{3.10}$$

can be represented in terms of the stochastic integral with respect to the martingale M^H. For a measurable function f on $[0,T]$, let

$$K_H^f(t,s) = -2H \frac{d}{ds} \int_s^t f(r) r^{H-\frac{1}{2}} (r-s)^{H-\frac{1}{2}} dr, 0 \leq s \leq t \tag{3.11}$$

when the derivative exists in the sense of absolute continuity with respect to the Lebesgue measure(see Samko et al. (1993) for sufficient conditions). The following result is due to Kleptsyna et al. (2000).

Theorem 3.1. *Let M^H be the fundamental martingale associated with the fBm W^H defined by (3.9). Then*

$$\int_0^t f(s) dW_s^H = \int_0^t K_H^f(t,s) dM_s^H, t \in [0,T] \tag{3.12}$$

a.s. $[P]$ whenever both sides are well defined.

Suppose the sample paths of the process $\{\frac{C(t)}{B(t)}, t \geq 0\}$ are smooth enough (see Samko et al. (1993)) so that

$$Q_H(t) = \frac{d}{dw_t^H} \int_0^t \kappa_H(t,s) \frac{C(s)}{B(s)} ds, t \in [0,T] \tag{3.13}$$

is well-defined where w^H and k_H are as defined in (3.8) and (3.6) respectively and the derivative is understood in the sense of absoulute continuity. The following theorem due to Kleptsyna et al. (2000) associates a *fundamental semimartingale* Z associated with the process Y such that the natural filtration $(\mathcal{Z}_t)$ coincides with the natural filtration $(\mathcal{Y}_t)$ of Y.

Theorem 3.2. *Suppose the sample paths of the process Q_H defined by (3.13) belong P-a.s to $L^2([0,T], dw^H)$ where w^H is as defined by (3.8). Let the process $Z = (Z_t, t \in [0,T])$ be defined by*

$$Z_t = \int_0^t \kappa_H(t,s) B^{-1}(s) dY_s \tag{3.14}$$

where the function $\kappa_H(t,s)$ is as defined in (3.6). Then the following results hold: (i) The process Z is an $(\mathcal{F}_t)$ -semimartingale with the decomposition

$$Z_t = \int_0^t Q_H(s) dw_s^H + M_t^H \tag{3.15}$$

where M^H is the fundamental martingale defined by (3.9),
(ii) the process Y admits the representation

$$Y_t = \int_0^t K_H^B(t,s)dZ_s \tag{3.16}$$

where the function $K_H^B(.,.)$ is as defined in (3.11), and
(iii) the natural fitrations of $(\mathcal{Z}_t)$ and $(\mathcal{Y}_t)$ coincide.

Kleptsyna et al. (2000) derived the following Girsanov type formula as a consequence of the Theorem 3.2.

Theorem 3.3. *Suppose the assumptions of Theorem 3.2 hold. Define*

$$\Lambda_H(T) = \exp\{-\int_0^T Q_H(t)dM_t^H - \frac{1}{2}\int_0^t Q_H^2(t)dw_t^H\}. \tag{3.17}$$

Suppose that $E(\Lambda_H(T)) = 1$. Then the measure $P^* = \Lambda_H(T)P$ is a probability measure and the probability measure of the process Y under P^* is the same as that of the process V defined by

$$V_t = \int_0^t B(s)dW_s^H, 0 \le t \le T. \tag{3.18}$$

.

4. Statistical inference for linear SDE driven by fBm

Statistical inference for diffusion type processes satisfying stochastic differential equations driven by Wiener processes have been studied earlier and a comprehensive survey of various methods is given in Prakasa Rao (1999a, b). There has been a recent interest to study similar problems for stochastic processes driven by a fractional Brownian motion for modeling stochastic phemonena with possible long range dependence. Le Breton (1998) studied parameter estimation and filtering in a simple linear model driven by a fractional Brownian motion. In a recent paper, Kleptsyna and Le Breton (2002) studied parameter estimation problems for fractional Ornstein-Uhlenbeck type process. This is a fractional analogue of the Ornstein-Uhlenbeck process, that is, a continuous time first order autoregressive process $X = \{X_t, t \ge 0\}$ which is the solution of a one-dimensional homogeneous linear stochastic differential equation driven by a fractional Brownian motion (fBm) $W^H = \{W_t^H, t \ge 0\}$ with Hurst parameter $H \in [1/2, 1)$. Such a process is the unique Gaussian process satisfying the linear integral equation

$$X_t = \theta \int_0^t X_s ds + \sigma W_t^H, t \ge 0. \tag{4.1}$$

They investigate the problem of estimation of the parameters θ and σ^2 based on the observation $\{X_s, 0 \le s \le T\}$ and prove that the maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent as $T \to \infty$.

We now discuss more general classes of stochastic processes satisfying linear stochastic differential equations driven by a fractional Brownian motion and review some recent work connected with the asymptotic properties of the maximum likelihood and the Bayes estimators for parameters involved in such processes. We will also discuss some aspects of sequential estimation and minimum distance estimation problems for fractional Ornstein-Uhlenbeck type processes in the next section.

Let us consider the stochastic differential equation

$$dX(t) = [a(t, X(t)) + \theta\, b(t, X(t))]dt + \sigma(t)dW_t^H, X(0) = 0, t \geq 0 \tag{4.2}$$

where $\theta \in \Theta \subset R, W = \{W_t^H, t \geq 0\}$ is a fractional Brownian motion with *known* Hurst parameter H and $\sigma(t)$ is a positive nonvanishing function on $[0, \infty)$. In other words $X = \{X_t, t \geq 0\}$ is a stochastic process satisfying the stochastic integral equation

$$X(t) = \int_0^t [a(s, X(s)) + \theta\, b(s, X(s))]ds + \int_0^t \sigma(s)dW_s^H, X(0) = 0, t \geq 0. \tag{4.3}$$

Let

$$C(\theta, t) = a(t, X(t)) + \theta\, b(t, X(t)), t \geq 0 \tag{4.4}$$

and assume that the sample paths of the process $\{\frac{C(\theta,t)}{\sigma(t)}, t \geq 0\}$ are smooth enough so that the the process

$$Q_{H,\theta}(t) = \frac{d}{dw_t^H} \int_0^t \kappa_H(t, s) \frac{C(\theta, s)}{\sigma(s)} ds, t \geq 0 \tag{4.5}$$

is well-defined where w_t^H and $\kappa_H(t, s)$ are as defined in (3.8) and (3.6) respectively. Suppose the sample paths of the process $\{Q_{H,\theta}, 0 \leq t \leq T\}$ belong almost surely to $L^2([0, T], dw_t^H)$. Define

$$Z_t = \int_0^t \frac{\kappa_H(t, s)}{\sigma(s)} dX_s, t \geq 0. \tag{4.6}$$

Then the process $Z = \{Z_t, t \geq 0\}$ is an $(\mathcal{F}_t)$-semimartingale with the decomposition

$$Z_t = \int_0^t Q_{H,\theta}(s)dw_s^H + M_t^H \tag{4.7}$$

where M^H is the fundamental martingale defined by (3.9) and the process X admits the representation

$$X_t = \int_0^t K_H^\sigma(t, s)dZ_s \tag{4.8}$$

where the function $K_H^\sigma(., .)$ is as defined by (3.11). Let P_θ^T be the measure induced by the process $\{X_t, 0 \leq t \leq T\}$ when θ is the true parameter. Following Theorem 3.3, we get that the Radon-Nikodym derivative of P_θ^T with respect to P_0^T is given by

$$\frac{dP_\theta^T}{dP_0^T} = \exp[\int_0^T Q_{H,\theta}(s)dZ_s - \frac{1}{2}\int_0^T Q_{H,\theta}^2(s)dw_s^H]. \tag{4.9}$$

Maximum likelihood estimation

We now consider the problem of estimation of the parameter θ based on the observation of the process $X = \{X_t, 0 \leq t \leq T\}$ and study its asymptotic properties as $T \to \infty$.

Strong consistency:

Let $L_T(\theta)$ denote the Radon-Nikodym derivative $\frac{dP_\theta^T}{dP_0^T}$. The maximum likelihood estimator (MLE) $\hat{\theta}_T$ is defined by the relation

$$L_T(\hat{\theta}_T) = \sup_{\theta \in \Theta} L_T(\theta). \tag{4.10}$$

We assume that there exists such a measurable maximum likelihood estimator. Sufficient conditions can be given for the existence of such an estimator (cf. Lemma 3.1.2, Prakasa Rao (1987)).

Note that

$$\begin{aligned} Q_{H,\theta}(t) &= \frac{d}{dw_t^H} \int_0^t \kappa_H(t,s) \frac{C(\theta,s)}{\sigma(s)} ds \qquad (4.11) \\ &= \frac{d}{dw_t^H} \int_0^t \kappa_H(t,s) \frac{a(s,X(s))}{\sigma(s)} ds + \theta \frac{d}{dw_t^H} \int_0^t \kappa_H(t,s) \frac{b(s,X(s))}{\sigma(s)} ds \\ &= J_1(t) + \theta J_2(t). (say) \end{aligned}$$

Then

$$\log L_T(\theta) = \int_0^T (J_1(t) + \theta J_2(t)) dZ_t - \frac{1}{2} \int_0^T (J_1(t) + \theta J_2(t))^2 dw_t^H \tag{4.12}$$

and the likelihood equation is given by

$$\int_0^T J_2(t) dZ_t - \int_0^T (J_1(t) + \theta J_2(t)) J_2(t) dw_t^H = 0. \tag{4.13}$$

Hence the MLE $\hat{\theta}_T$ of θ is given by

$$\hat{\theta}_T = \frac{\int_0^T J_2(t) dZ_t + \int_0^T J_1(t) J_2(t) dw_t^H}{\int_0^T J_2^2(t) dw_t^H}. \tag{4.14}$$

Let θ_0 be the true parameter. Using the fact that

$$dZ_t = (J_1(t) + \theta_0 J_2(t)) dw_t^H + dM_t^H, \tag{4.15}$$

it can be shown that

$$\frac{dP_\theta^T}{dP_{\theta_0}^T} = \exp[(\theta - \theta_0) \int_0^T J_2(t) dM_t^H - \frac{1}{2}(\theta - \theta_0)^2 \int_0^T J_2^2(t) dw_t^H]. \tag{4.16}$$

Following this representation of the Radon-Nikodym derivative, we obtain that

$$\hat{\theta}_T - \theta_0 = \frac{\int_0^T J_2(t) dM_t^H}{\int_0^T J_2^2(t) dw_t^H}. \tag{4.17}$$

Note that the quadratic variation $< Z >$ of the process Z is the same as the quadratic variation $< M^H >$ of the martingale M^H which in turn is equal to w^H. This follows from the relations (3.15) and (3.9). Hence we obtain that

$$[w_T^H]^{-1} \lim_n \Sigma [Z_{t_{i+1}^{(n)}} - Z_{t_i^{(n)}}]^2 = 1 \text{ a.s } [P_{\theta_0}]$$

where $(t_i^{(n)}$ is a partition of the interval $[0,T]$ such that $\sup|t_{i+1}^{(n)} - t_i^{(n)}|$ tends to zero as $n \to \infty$. If the function $\sigma(t)$ is an unknown constant σ, the above property can be used to obtain a strongly consistent estimator of σ^2 based on the continuous observation of the process X over the interval $[0,T]$. Here after we assume that the nonrandom function $\sigma(t)$ is known.

We now discuss the problem of maximum likelihood estimation of the parameter θ on the basis of the observation of the process X or equivalently the process Z on the interval $[0,T]$. The following result holds.

Theorem 4.1. *The maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent, that is,*

$$\hat{\theta}_T \to \theta_0 \ \ a.s \ \ [P_{\theta_0}] \ as \ \ T \to \infty \tag{4.18}$$

provided

$$\int_0^T J_2^2(t) dw_t^H \to \infty \ \ a.s \ \ [P_{\theta_0}] \ \ as \ \ T \to \infty. \tag{4.19}$$

Remark. For the case fractional Ornstein-Uhlenbeck type process investigated in Kleptsyna and Le Breton (2002), it can be checked that the condition stated in equation (4.19) holds and hence the maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent as $T \to \infty$.

Limiting distribution:

We now discuss the limiting distribution of the MLE $\hat{\theta}_T$ as $T \to \infty$.

Theorem 4.2. *Assume that the functions $b(t,s)$ and $\sigma(t)$ are such that the process $\{R_t, t \geq 0\}$ is a local continuous martingale and that there exists a norming function $I_t, t \geq 0$ such that*

$$I_T^2 < R_T >= I_T^2 \int_0^T J_2^2(t) dw_t^H \xrightarrow{p} \eta^2 \ \ as \ \ T \to \infty \tag{4.20}$$

where $I_T \to 0$ as $T \to \infty$ and η is a random variable such that $P(\eta > 0) = 1$. Then

$$(I_T R_T, I_T^2 < R_T >) \xrightarrow{\mathcal{L}} (\eta Z, \eta^2) \ \ as \ \ T \to \infty \tag{4.21}$$

where the random variable Z has the standard normal distribution and the random variables Z and η are independent.

For the proofs of Theorems 4.1 and 4.2, see Prakasa Rao (2003a).

Theorem 4.3. *Suppose the conditions stated in the Theorem 4.2 hold. Then*

$$I_T^{-1}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} \frac{Z}{\eta} \ \ as \ \ t \to \infty \tag{4.22}$$

where the random variable Z has the standard normal distribution and the random variables Z and η are independent.

Remarks. If the random variable η is a constant with probability one, then the limiting distribution of the maximum likelihood estimator is normal with mean 0 and variance η^{-2}. Otherwise it is a mixture of the normal distributions with mean zero and variance η^{-2} with the mixing distribution as that of η. The rate of convergence of the distribution of the maximum likelihood estimator is discussed in Prakasa Rao (2003b).

Bayes estimation

Suppose that the parameter space Θ is open and Λ is a prior probability measure on the parameter space Θ. Further suppose that Λ has the density $\lambda(.)$ with respect to the Lebesgue measure and the density function is continuous and positive in an open neighbourhood of θ_0, the true parameter. Let

$$\alpha_T \equiv I_T R_T = I_T \int_0^T J_2(t) dM_t^H \tag{4.23}$$

and

$$\beta_T \equiv I_T^2 < R_T >= I_T^2 \int_0^T J_2^2(t) dw_t^H. \tag{4.24}$$

We have seen earlier in (4.17) that the maximum likelihood estimator satisfies the relation

$$\alpha_T = (\hat{\theta}_T - \theta_0) I_T^{-1} \beta_T. \tag{4.25}$$

The posterior density of θ given the observation $X^T \equiv \{X_s, 0 \leq s \leq T\}$ is given by

$$p(\theta | X^T) = \frac{\frac{dP_\theta^T}{dP_{\theta_0}^T} \lambda(\theta)}{\int_\Theta \frac{dP_\theta^T}{dP_{\theta_0}^T} \lambda(\theta) d\theta}. \tag{4.26}$$

Let us write $t = I_T^{-1}(\theta - \hat{\theta}_T)$ and define

$$p^*(t | X^T) = I_T \;\; p(\hat{\theta}_T + t I_T | X^T). \tag{4.27}$$

Then the function $p^*(t|X^T)$ is the posterior density of the transformed variable $t = I_T^{-1}(\theta - \hat{\theta}_T)$. Let

$$\begin{aligned} \nu_T(t) &\equiv \frac{dP_{\hat{\theta}_T + t I_T} / dP_{\theta_0}}{dP_{\hat{\theta}_T} / dP_{\theta_0}} \\ &= \frac{dP_{\hat{\theta}_T + t I_T}}{dP_{\hat{\theta}_T}} \text{ a.s.} \end{aligned} \tag{4.28}$$

and

$$C_T = \int_{-\infty}^{\infty} \nu_T(t) \lambda(\hat{\theta}_T + t I_T) dt. \tag{4.29}$$

It can be checked that

$$p^*(t | X^T) = C_T^{-1} \nu_T(t) \lambda(\hat{\theta}_T + t I_T) \tag{4.30}$$

and

$$\begin{aligned} \log \nu_T(t) &= I_T^{-1} \alpha_T [(\hat{\theta}_T + t I_T - \theta_0) - (\hat{\theta}_T - \theta_0)] \\ &\quad - \frac{1}{2} I_T^{-2} \beta_T [(\hat{\theta}_T + t I_T - \theta_0)^2 - (\hat{\theta}_T - \theta_0)^2] \\ &= t \alpha_T - \frac{1}{2} t^2 \beta_T - t \beta_T I_T^{-1} (\hat{\theta}_T - \theta_0) \\ &= -\frac{1}{2} \beta_T t^2 \end{aligned} \tag{4.31}$$

in view of the equation (4.25).

Suppose that the convergence in the condition in the equation (4.20) holds almost surely under the measure P_{θ_0} and the limit is a constant $\eta^2 > 0$ with probability one. For convenience, we write $\beta = \eta^2$. Then

$$\beta_T \to \beta \ \text{ a.s } [P_{\theta_0}] \ \text{ as } \ T \to \infty. \tag{4.32}$$

Further suppose that $K(t)$ is a nonnegative measurable function such that, for some $0 < \varepsilon < \beta$,

$$\int_{-\infty}^{\infty} K(t) \exp[-\frac{1}{2}t^2(\beta - \varepsilon)]dt < \infty \tag{4.33}$$

and the maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent, that is,

$$\hat{\theta}_T \to \theta_0 \ \text{ a.s } \quad [P_{\theta_0}] \text{ as } \ T \to \infty. \tag{4.34}$$

In addition, suppose that the following condition holds for every $\varepsilon > 0$ and $\delta > 0$:

$$\exp[-\varepsilon I_T^{-2}] \int_{|u|>\delta} K(uI_T^{-1})\lambda(\hat{\theta}_T + u)du \to 0 \ \text{ a.s.}[P_{\theta_0}] \ \text{ as } \ T \to \infty. \tag{4.35}$$

Then we have the follwing theorem which is an analogue of the Bernstein - von Mises theorem proved in Prakasa Rao (1981) for a class of processes satisfying a linear stochastic differential equation driven by the standard Wiener process.

Theorem 4.4. *Let the assumptions (4.32) to (4.35) hold where $\lambda(.)$ is a prior density which is continuous and positive in an open neighbourhood of θ_0, the true parameter. Then*

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} K(t)|p^*(t|X^T) - (\frac{\beta}{2\pi})^{1/2} \exp(-\frac{1}{2}\beta t^2)|dt = 0 \ \ a.s \ \ [P_{\theta_0}]. \tag{4.36}$$

As a consequence of the above theorem, we obtain the following result by choosing $K(t) = |t|^m$, for some integer $m \geq 0$.

Theorem 4.5. *Assume that the following conditions hold:*

$$(C1) \ \ \hat{\theta}_T \to \theta_0 \ \ a.s \ \ [P_{\theta_0}] \ \ as \ \ T \to \infty, \tag{4.37}$$

$$(C2) \ \ \beta_T \to \beta > 0 \ \ a.s \ \ [P_{\theta_0}] \ \ as \ \ T \to \infty. \tag{4.38}$$

Further suppose that
(C3)$\lambda(.)$ is a prior probability density on Θ which is continuous and positive in an open neighbourhood of θ_0, the true parameter and

$$(C4) \ \ \int_{-\infty}^{\infty} |\theta|^m \lambda(\theta)d\theta < \infty \tag{4.39}$$

for some integer $m \geq 0$. Then

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} |t|^m |p^*(t|X^T) - (\frac{\beta}{2\pi})^{1/2} \exp(-\frac{1}{2}\beta t^2)|dt = 0 \ \ a.s \ \ [P_{\theta_0}]. \tag{4.40}$$

In particular, choosing $m = 0$, *we obtain that*

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} |p^*(t|X^T) - (\frac{\beta}{2\pi})^{1/2} \exp(-\frac{1}{2}\beta t^2)|dt = 0 \ \ a.s \ \ [P_{\theta_0}] \tag{4.41}$$

whenver the conditions (C1), (C2) and (C3) hold. This is the analogue of the Bernstein-von Mises theorem for a class of diffusion processes proved in Prakasa Rao (1981) and it shows the asymptotic convergence in the L_1*-mean of the posterior density to the normal distribution.*

For proofs of above results, see Prakasa Rao (2003a).

As a Corollory to the Theorem 4.5, we also obtain that the conditional expectation, under P_{θ_0}, of $[I_T^{-1}(\hat{\theta}_T - \theta)]^m$ converges to the corresponding m-th abosolute moment of the normal distribution with mean zero and variance β^{-1}.

We define a *regular Bayes estimator* of θ, corresponding to a prior probability density $\lambda(\theta)$ and the loss function $L(\theta, \phi)$, based on the observation X^T, as an estimator which minimizes the posterior risk

$$B_T(\phi) \equiv \int_{-\infty}^{\infty} L(\theta, \phi) p(\theta|X^T) d\theta. \tag{4.42}$$

over all the estimators ϕ of θ. Here $L(\theta, \phi)$ is a loss function defined on $\Theta \times \Theta$.

Suppose there exists a measurable regular Bayes estimator $\tilde{\theta}_T$ for the parameter θ (cf. Theorem 3.1.3, Prakasa Rao (1987).) Suppose that the loss function $L(\theta, \phi)$ satisfies the following conditions:

$$L(\theta, \phi) = \ell(|\theta - \phi|) \geq 0 \tag{4.43}$$

and the function $\ell(t)$ is nondecreasing for $t \geq 0$. An example of such a loss function is $L(\theta, \phi) = |\theta - \phi|$. Suppose there exist nonnegative functions $R(t)$, $J(t)$ and $G(t)$ such that

$$(D1) \ \ R(t)\ell(tI_T) \leq G(t) \ \ \text{for all} \ \ T \geq 0, \tag{4.44}$$

$$(D2) \ \ R(t)\ell(tI_T) \to J(t) \ \ \text{as} \ \ T \to \infty \tag{4.45}$$

uniformly on bounded intervals of t. Further suppose that the function

$$(D3) \ \ \int_{-\infty}^{\infty} J(t+h) \exp[-\frac{1}{2}\beta t^2] dt \tag{4.46}$$

has a strict minimum at $h = 0$, and
(D4)the function $G(t)$ satisfies the conditions similar to (4.33) and (4.35).

We have the following result giving the asymptotic properties of the Bayes risk of the estimator $\tilde{\theta}_T$.

Theorem 4.6. *Suppose the conditions (C1) to (C3) in the Theorem 4.5 and the conditions (D1) to (D4) stated above hold. Then*

$$I_T^{-1}(\tilde{\theta}_T - \hat{\theta}_T) \to 0 \ \ a.s \ \ [P_{\theta_0}] \ \ as \ \ T \to \infty \tag{4.47}$$

and

$$\begin{aligned} \lim_{T\to\infty} R(T) B_T(\tilde{\theta}_T) &= \lim_{T\to\infty} R(T) B_T(\hat{\theta}_T) \\ &= (\frac{\beta}{2\pi})^{1/2} \int_{-\infty}^{\infty} K(t) \exp[-\frac{1}{2}\beta t^2] dt \ \ a.s \ \ [P_{\theta_0}]. \end{aligned} \tag{4.48}$$

This theorem can be proved by arguments similar to those given in the proof of Theorem 4.1 in Borwanker et al. (1971).

We have observed earlier that

$$I_T^{-1}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} N(0, \beta^{-1}) \text{ as } T \to \infty. \tag{4.49}$$

As a consequence of the Theorem 4.6, we obtain that

$$\tilde{\theta}_T \to \theta_0 \text{ a.s } [P_{\theta_0}] \text{ as } T \to \infty \tag{4.50}$$

and

$$I_T^{-1}(\tilde{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} N(0, \beta^{-1}) \text{ as } T \to \infty. \tag{4.51}$$

In other words, the Bayes estimator is asymptotically normal and has asymptotically the same distribution as the maxiumum likelihood estimator. The asymptotic Bayes risk of the estimator is given by the Theorem 4.6.

5. Statistical inference for fractional Ornstein–Uhlenbeck type process

In a recent paper, Kleptsyna and Le Breton (2002) studied parameter estimation problems for fractional Ornstein-Uhlenbeck type process. This is a fractional analogue of the Ornstein-Uhlenbeck process, that is, a continuous time first order autoregressive process $X = \{X_t, t \geq 0\}$ which is the solution of a one-dimensional homogeneous linear stochastic differential equation driven by a fractional Brownian motion (fBm) $W^H = \{W_t^H, t \geq 0\}$ with Hurst parameter $H \in (1/2, 1)$. Such a process is the unique Gaussian process satisfying the linear integral equation

$$X_t = \theta \int_0^t X_s ds + \sigma W_t^H, t \geq 0. \tag{5.1}$$

They investigate the problem of estimation of the parameters θ and σ^2 based on the observation $\{X_s, 0 \leq s \leq T\}$ and prove that the maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent as $T \to \infty$. It is well known that the sequential estimation methods might lead to equally efficient estimators, as compared to the maximum likelihood estimators, from the process observed possibly over a shorter expected period of observation time. Novikov (1972) investigated the asymptotic properties of a sequential maximum likelihood estimator for the drift parameter in the Ornstein-Uhlenbeck process. Maximum likelihood estimators are not robust. Kutoyants and Pilibossian (1994) developed a minimum L_1-norm estimator for the drift parameter. We now discuss the asymptotic properties of a sequential maximum likelihood estimators and minimum L_1-norm estimators for the drift parameter for a fractional Ornstein-Uhlenbeck type process.

Maximum likelihood estimation

Let

$$K_H(t,s) = H(2H-1)\frac{d}{ds}\int_s^t r^{H-\frac{1}{2}}(r-s)^{H-\frac{3}{2}} dr, 0 \leq s \leq t. \tag{5.2}$$

The sample paths of the process $\{X_t, t \geq 0\}$ are smooth enough so that the process Q defined by

$$Q(t) = \frac{d}{dw_t^H}\int_0^t \kappa_H(t,s) X_s ds, t \in [0,T] \tag{5.3}$$

is well-defined where w^H and $\kappa_H(t,s)_H$ are as defined in (3.8) and (3.6) respectively and the derivative is understood in the sense of absolute continuity with respect to the measure generated by w^H. More over the sample paths of the process Q belong to $L^2([0,T], dw^H)$ a.s. [P]. Define the process Z as in (4.6).

As an application of the Girsanov type formula given in Theorem 3.3 for the fractional Brownian motions derived by Kleptsyna et al. (2000) , it follows that the Radon-Nikodym derivative of the measure P_θ^T, generated by the stochastic process X when θ is the true parameter, with respect to the measure generated by the process X when $\theta = 0$, is given by

$$\frac{dP_\theta^T}{dP_0^T} = \exp[\theta \int_0^T Q(s)dZ_s - \frac{1}{2}\theta^2 \int_0^T Q^2(s)dw_s^H]. \tag{5.4}$$

Further more the quadratic variation $< Z >_T$ of the process Z on $[0,T]$ is equal to $\sigma^2 w_T^H$ a.s. and hence the parameter σ^2 can be estimated by the relation

$$\lim_n \Sigma[Z_{t_{i+1}^{(n)}} - Z_{t_i^{(n)}}]^2 = \sigma^2 w_T^H \text{ a.s.} \tag{5.5}$$

where $(t_i^{(n)})$ is an appropriate partition of $[0,T]$ such that

$$\sup_i |t_{i+1}^{(n)} - t_i^{(n)}| \to 0$$

as $n \to \infty$. Hence we can estimate σ^2 almost surely from any small interval as long as we have a continuous observation of the process. For further discussion, we assume that $\sigma^2 = 1$.

We consider the problem of estimation of the parameter θ based on the observation of the process $X = \{X_t, 0 \le t \le T\}$ for a fixed time T and study its asymptotic properties as $T \to \infty$. The following results are due to Kleptsyna and Le Breton (2002) and Prakasa Rao (2003a).

Theorem 5.1. *The maximum likelihood estimator θ from the observation $X = \{X_t, 0 \le t \le T\}$ is given by*

$$\hat{\theta}_T = \{\int_0^T Q^2(s)dw_s^H\}^{-1} \int_0^T Q(s)dZ_s. \tag{5.6}$$

Then the estimator $\hat{\theta}_T$ is strongly consistent as $T \to \infty$, that is,

$$\lim_{T\to\infty} \hat{\theta}_T = \theta \text{ a.s. } [P_\theta] \tag{5.7}$$

for every $\theta \in R$.

We now discuss the limiting distribution of the MLE $\hat{\theta}_T$ as $T \to \infty$.

Theorem 5.2. *Let*

$$R_T = \int_0^T Q(s)dZ_s. \tag{5.8}$$

Assume that there exists a norming function $I_t, t \ge 0$ such that

$$I_T^2 \int_0^T Q^2(t)dw_t^H \xrightarrow{p} \eta^2 \text{ as } T \to \infty \tag{5.9}$$

where $I_T \to 0$ as $T \to \infty$ and η is a random variable such that $P(\eta > 0) = 1$. Then

$$(I_T R_T, I_T^2 < R_T >) \xrightarrow{\mathcal{L}} (\eta Z, \eta^2) as \ T \to \infty \tag{5.10}$$

where the random variable Z has the standard normal distribution and the random variables Z and η are independent.

Observe that

$$I_T^{-1}(\hat{\theta}_T - \theta_0) = \frac{I_T R_T}{I_T^2 < R_T >} \tag{5.11}$$

Applying the Theorem 5.2, we obtain the following result.

Theorem 5.3. *Suppose the conditions stated in the Theorem 5.2 hold. Then*

$$I_T^{-1}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} \frac{Z}{\eta} \ as \ T \to \infty \tag{5.12}$$

where the random variable Z has the standard normal distribution and the random variables Z and η are independent.

Remarks. If the random variable η is a constant with probability one, then the limiting distribution of the maximum likelihood estimator is normal with mean 0 and variance η^{-2} Otherwise it is a mixture of the normal distributions with mean zero and variance η^{-2} with the mixing distribution as that of η. Berry-Esseen type bound for the MLE is discussed in Prakasa Rao (2003b) when the limiting distribution of the MLE is normal.

Sequential maximum likelihood estimation

We now consider the problem of sequential maximum likelihood estimation of the parameter θ. Let h be a nonnegative number. Define the stopping rule $\tau(h)$ by the rule

$$\tau(h) = \inf\{t : \int_0^t Q^2(s) dw_s^H \geq h\}. \tag{5.13}$$

Kletptsyna and Le Breton (2002) have shown that

$$\lim_{t \to \infty} \int_0^t Q^2(s) dw_s^H = +\infty \ \text{ a.s. } [P_\theta] \tag{5.14}$$

for every $\theta \in R$. Then it can be shown that $P_\theta(\tau(h) < \infty) = 1$. If the process is observed up to a previuosly determined time T , we know that the maximum likelihood estimator is given by

$$\hat{\theta}_T = \{\int_0^T Q^2(s) dw_s^H\}^{-1} \int_0^T Q(s) dZ_s. \tag{5.15}$$

The estimator

$$\begin{aligned} \hat{\theta}(h) &\equiv \hat{\theta}_{\tau(h)} \\ &= \{\int_0^{\tau(h)} Q^2(s) dw_s^H\}^{-1} \int_0^{\tau(h)} Q(s) dZ_s \\ &= h^{-1} \int_0^{\tau(h)} Q(s) dZ_s \end{aligned} \tag{5.16}$$

is called the *sequential maximum likelihood estimator* of θ. We now study the asymptotic properties of the estimator $\hat{\theta}(h)$.

The following lemma is an analogue of the Cramer-Rao inequality for sequential plans $(\tau(X), \hat{\theta}_\tau(X))$ for estimating the parameter θ satisfying the property

$$E_\theta\{\hat{\theta}_\tau(X)\} = \theta \tag{5.17}$$

for all θ.

Lemma 5.4. *Suppose that differentiation under the integral sign with respect to θ on the left side of the equation (5.17) is permissible. Further suppose that*

$$E_\theta\{\int_0^{\tau(X)} Q^2(s)dw_s^H\} < \infty \tag{5.18}$$

for all θ. Then

$$Var_\theta\{\hat{\theta}_\tau(X)\} \geq \{E_\theta\{\int_0^{\tau(X)} Q^2(s)dw_s^H\}^{-1} \tag{5.19}$$

for all θ.

A sequential plan $(\tau(X), \hat{\theta}_\tau(X))$ is said to be *efficient* if there is equality in (5.19) for all θ. We now have the following result.

Theorem 5.5. *Consider the fractional Ornstein-Uhlenbeck type process governed by the stochastic differential equation (5.1) with $\sigma = 1$ driven by the fractional Brownian motion W^H with $H \in [\frac{1}{2}, 1)$. Then the sequential plan $(\tau(h), \hat{\theta}(h))$ defined by the equations (5.13) and (5.16) has the following properties for all θ.*

(i) *$\hat{\theta}(h) \equiv \hat{\theta}_{\tau(h)}$ is normally distributed with $E_\theta(\hat{\theta}(h)) = \theta$ and $Var_\theta(\hat{\theta}(h)) = h^{-1}$;*

(ii) *the plan is efficient; and*

(iii) *the plan is closed, that is, $P_\theta(\tau(h) < \infty) = 1$.*

For proof, see Prakasa Rao (2004a).

Minimum L_1-norm estimation

In spite of the fact that maximum likelihood estimators (MLE) are consistent and asymptotically normal and also asymptotically efficient in general, they have some short comings at the same time. Their calculation is often cumbersome as the expression for the MLE involve stochastic integrals which need good approximations for computational purposes. Further more the MLE are not robust in the sense that a slight perturbation in the noise component will change the properties of the MLE substantially. In order to circumvent such problems, the minimum distance approach is proposed. Properties of the minimum distance estimators (MDE) were discussed in Millar (1984) in a general frame work.

We now obtain the minimum L_1-norm estimates of the drift parameter of a fractional Ornstein-Uhlenbeck type process and investigate the asymptotic properties of such estimators following the work of Kutoyants and Pilibossian (1994).

We now consider the problem of estimation of the parameter θ based on the observation of fractional Ornstein-Uhlenbeck type process $X = \{X_t, 0 \leq t \leq T\}$ satisfying the stochastic differential equation

$$dX_t = \theta X(t)dt + \varepsilon dW_t^H, X_0 = x_0, 0 \leq t \leq T \tag{5.20}$$

for a fixed time T where $\theta \in \Theta \subset R$ and study its asymptotic properties as $\varepsilon \to 0$.

Let $x_t(\theta)$ be the solution of the above differential equation with $\varepsilon = 0$. It is obvious that

$$x_t(\theta) = x_0 e^{\theta t}, 0 \le t \le T. \tag{5.21}$$

Let

$$S_T(\theta) = \int_0^T |X_t - x_t(\theta)| dt. \tag{5.22}$$

We define θ_ε^* to be a *minimum L_1-norm estimator* if there exists a measurable selection θ_ε^* such that

$$S_T(\theta_\varepsilon^*) = \inf_{\theta \in \Theta} S_T(\theta). \tag{5.23}$$

Conditions for the existence of a measurable selection are given in Lemma 3.1.2 in Prakasa Rao (1987). We assume that there exists a measurable selection θ_ε^* satisfying the above equation. An alternate way of defining the estimator θ_ε^* is by the relation

$$\theta_\varepsilon^* = \arg \inf_{\theta \in \Theta} \int_0^T |X_t - x_t(\theta)| dt. \tag{5.24}$$

Consistency:

Let $W_T^{H*} = \sup_{0 \le t \le T} |W_t^H|$. The self-similarity of the fractional Brownian motion W_t^H implies that the random variables W_{at}^H and $a^H W_t$ have the same probability distribution for any $a > 0$. Further more it follows from the self-similarity that the supremum process W^{H*} has the property that the random variables W_{at}^{H*} and $a^H W_t^{H*}$ have the same probability distribution for any $a > 0$. Hence we have the following observation due to Novikov and Valkeila (1999).

Lemma 5.6. *Let $T > 0$ and the process $\{W_t^H, 0 \le t \le T\}$ be a fBm with Hurst index H. Let $W_T^{H*} = \sup_{0 \le t \le T} W_t^H$. Then*

$$E(W_T^{H*})^p = K(p, H) T^{pH} \tag{5.25}$$

for every $p > 0$, where $K(p, H) = E(W_1^{H*})^p$.

Let θ_0 denote the true parameter, For any $\delta > 0$, define

$$g(\delta) = \inf_{|\theta - \theta_0| > \delta} \int_0^T |x_t(\theta) - x_t(\theta_0)| dt. \tag{5.26}$$

Note that $g(\delta) > 0$ for any $\delta > 0$.

Theorem 5.7. *For every $p > 0$, there exists a positive constant $K(p, H)$ such that, for every $\delta > 0$,*

$$\begin{aligned} P_{\theta_0}^{(\varepsilon)}\{|\theta_\varepsilon^* - \theta_0| > \delta\} &\le 2^p T^{pH+p} K(p, H) e^{|\theta_0| T p} (g(\delta))^{-p} \varepsilon^p \\ &= O((g(\delta))^{-p} \varepsilon^p). \end{aligned} \tag{5.27}$$

Proof. Let $\|.\|$ denote the L_1-norm. Then

$$\begin{aligned} P_{\theta_0}^{(\varepsilon)}\{|\theta_\varepsilon^* - \theta_0| > \delta\} &= P_{\theta_0}^{(\varepsilon)}\{\inf_{|\theta-\theta_0|\leq\delta}||X - x(\theta)|| > \inf_{|\theta-\theta_0|>\delta}||X - x(\theta)||\} \\ &\leq P_{\theta_0}^{(\varepsilon)}\{\inf_{|\theta-\theta_0|\leq\delta}(||X - x(\theta_0)|| + ||x(\theta) - x(\theta_0)||) \\ &> \inf_{|\theta-\theta_0|>\delta}(||x(\theta) - x(\theta_0)|| - ||X - x(\theta_0)||)\} \\ &= P_{\theta_0}^{(\varepsilon)}\{2||X - x(\theta_0)|| > \inf_{|\theta-\theta_0|>\delta}||x(\theta) - x(\theta_0)||\} \\ &= P_{\theta_0}^{(\varepsilon)}\{||X - x(\theta_0)|| > \frac{1}{2}g(\delta)\}. \end{aligned} \tag{5.28}$$

Since the process X_t satisfies the stochastic differential equation (5.20), it follows that

$$\begin{aligned} X_t - x_t(\theta_0) &= x_0 + \theta_0\int_0^t X_s ds + \varepsilon W_t^H - x_t(\theta_0) \\ &= \theta_0\int_0^t (X_s - x_s(\theta_0))ds + \varepsilon W_t^H \end{aligned} \tag{5.29}$$

since $x_t(\theta) = x_0 e^{\theta t}$. Let $U_t = X_t - x_t(\theta_0)$. Then it follows from the above equation that

$$U_t = \theta_0\int_0^t U_s \ ds + \varepsilon W_t^H. \tag{5.30}$$

Let $V_t = |U_t| = |X_t - x_t(\theta_0)|$. The above relation implies that

$$V_t = |X_t - x_t(\theta_0)| \leq |\theta_0|\int_0^t V_s ds + \varepsilon|W_t^H|. \tag{5.31}$$

Applying Gronwall-Bellman Lemma, we obtain that

$$\sup_{0\leq t\leq T}|V_t| \leq \varepsilon e^{|\theta_0 T|}\sup_{0\leq t\leq T}|W_t^H|. \tag{5.32}$$

Hence

$$\begin{aligned} P_{\theta_0}^{(\varepsilon)}\left\{||X - x(\theta_0)|| > \frac{1}{2}g(\delta)\right\} &\leq P\left\{\sup_{0\leq t\leq T}|W_t^H| > \frac{e^{-|\theta_0 T|}g(\delta)}{2\varepsilon T}\right\} \\ &= P\left\{W_T^{H*} > \frac{e^{-|\theta_0 T|}g(\delta)}{2\varepsilon T}\right\}. \end{aligned} \tag{5.33}$$

Applying the Lemma 5.6 to the estimate obtained above, we get that

$$\begin{aligned} P_{\theta_0}^{(\varepsilon)}\{|\theta_\varepsilon^* - \theta_0| > \delta\} &\leq 2^p T^{pH+p}K(p,H)e^{|\theta_0 T|p}(g(\delta))^{-p}\varepsilon^p \\ &= O((g(\delta))^{-p}\varepsilon^p). \end{aligned} \tag{5.34}$$

□

Remarks. As a consequence of the above theorem, we obtain that θ_ε^* converges in probability to θ_0 under $P_{\theta_0}^{(\varepsilon)}$-measure as $\varepsilon \to 0$. Further more the rate of convergence is of the order $(O(\varepsilon^p))$ for every $p > 0$.

Asymptotic distribution

We will now study the asymptotic distribution if any of the estimator θ_ε^* after suitable scaling. It can be checked that

$$X_t = e^{\theta_0 t}\{x_0 + \int_0^t e^{-\theta_0 s}\varepsilon dW_s^H\} \tag{5.35}$$

or equivalently

$$X_t - x_t(\theta_0) = \varepsilon e^{\theta_0 t}\int_0^t e^{-\theta_0 s} dW_s^H. \tag{5.36}$$

Let

$$Y_t = e^{\theta_0 t}\int_0^t e^{-\theta_0 s} dW_s^H. \tag{5.37}$$

Note that $\{Y_t, 0 \le t \le T\}$ is a gaussian process and can be interpreted as the "derivative" of the process $\{X_t, 0 \le t \le T\}$ with respect to ε. Applying Theorem 3.1, we obtain that, P-a.s.,

$$Y_t e^{-\theta_0 t} = \int_0^t e^{-\theta_0 s} dW_s^H = \int_0^t K_H^f(t,s) dM_s^H, t \in [0,T] \tag{5.38}$$

where $f(s) = e^{-\theta_0 s}, s \in [0,T]$ and M^H is the fundamental gaussian martingale associated with the fBm W^H. In particular it follows that the random variable $Y_t e^{-\theta_0 t}$ and hence Y_t has normal distribution with mean zero and further more, for any $h \ge 0$,

$$\begin{aligned} Cov(Y_t, Y_{t+h}) &= e^{2\theta_0 t + \theta_0 h} E[\int_0^t e^{-\theta_0 u} dW_u^H \int_0^{t+h} e^{-\theta_0 v} dW_v^H] \\ &= e^{2\theta_0 t + \theta_0 h} H(2H-1)\int_0^t \int_0^t e^{-\theta_0(u+v)}|u-v|^{2H-2} du dv \\ &= e^{2\theta_0 t + \theta_0 h}\gamma_H(t) \ \text{(say)}. \end{aligned} \tag{5.39}$$

In particular

$$Var(Y_t) = e^{2\theta_0 t}\gamma_H(t). \tag{5.40}$$

Hence $\{Y_t, 0 \le t \le T\}$ is a zero mean gaussian process with $Cov(Y_t, Y_s) = e^{\theta_0(t+s)}\gamma_H(t)$ for $s \ge t$.

Let

$$\zeta = \arg \inf_{-\infty < u < \infty} \int_0^T |Y_t - utx_0 e^{\theta_0 t}| dt. \tag{5.41}$$

Theorem 5.8. *As $\varepsilon \to 0$, the random variable $\varepsilon^{-1}(\theta_\varepsilon^* - \theta_0)$ converges in probability under the probability measure P_{θ_0} to a random variable whose probability distribution is the same as that of the random variable ζ under P_{θ_0}.*

Proof. Let $x_t'(\theta) = x_0 t e^{\theta t}$ and let

$$Z_\varepsilon(u) = ||Y - \varepsilon^{-1}(x(\theta_0 + \varepsilon u) - x(\theta_0))|| \tag{5.42}$$

and

$$Z_0(u) = ||Y - ux'(\theta_0)||. \tag{5.43}$$

Further more, let

$$A_\varepsilon = \{\omega : |\theta_\varepsilon^* - \theta_0| < \delta_\varepsilon\}, \delta_\varepsilon = \varepsilon^\tau, \tau \in (\frac{1}{2}, 1), L_\varepsilon = \varepsilon^{\tau-1}. \tag{5.44}$$

Observe that the random variable $u_\varepsilon^* = \varepsilon^{-1}(\theta_\varepsilon^* - \theta_0)$ satisfies the equation

$$Z_\varepsilon(u_\varepsilon^*) = \inf_{|u|<L_\varepsilon} Z_\varepsilon(u), \omega \in A_\varepsilon. \tag{5.45}$$

Define

$$\zeta_\varepsilon = \arg \inf_{|u|<L_\varepsilon} Z_0(u). \tag{5.46}$$

Observe that, with probability one,

$$\begin{aligned} \sup_{|u|<L_\varepsilon} |Z_\varepsilon(u) - Z_0(u)| &= |||Y - ux'(\theta_0) - \frac{1}{2}\varepsilon u^2 x''(\tilde{\theta})|| - ||Y - ux'(\theta_0)||| \\ &\leq \frac{\varepsilon}{2} L_\varepsilon^2 \sup_{|\theta-\theta_0|<\delta_\varepsilon} \int_0^T |x''(\theta)|dt \\ &\leq C\varepsilon^{2\tau-1}. \end{aligned} \tag{5.47}$$

Here $\tilde{\theta} = \theta_0 + \alpha(\theta - \theta_0)$ for some $\alpha \in (0, 1)$. Note that the last term in the above inequality tends to zero as $\varepsilon \to 0$. Further more the process $\{Z_0(u), -\infty < u < \infty\}$ has a unique minimum u^* with probability one. This follows from the arguments given in Theorem 2 of Kutoyants and Pilibossian (1994). In addition, we can choose the interval $[-L, L]$ such that

$$P_{\theta_0}^{(\varepsilon)}\{u_\varepsilon^* \in (-L, L)\} \geq 1 - \beta g(L)^{-p} \tag{5.48}$$

and

$$P\{u^* \in (-L, L)\} \geq 1 - \beta g(L)^{-p} \tag{5.49}$$

where $\beta > 0$. Note that $g(L)$ increases as L increases. The processes $Z_\varepsilon(u), u \in [-L, L]$ and $Z_0(u), u \in [-L, L]$ satisfy the Lipschitz conditions and $Z_\varepsilon(u)$ converges uniformly to $Z_0(u)$ over $u \in [-L, L]$. Hence the minimizer of $Z_\varepsilon(.)$ converges to the minimizer of $Z_0(u)$. This completes the proof. □

Remarks. We have seen earlier that the process $\{Y_t, 0 \leq t \leq T\}$ is a zero mean gaussian process with the covariance function

$$Cov(Y_t, Y_s) = e^{\theta_0(t+s)}\gamma_H(t)$$

for $s \geq t$. Recall that

$$\zeta = \arg \inf_{-\infty<u<\infty} \int_0^T |Y_t - utx_0 e^{\theta_0 t}|dt. \tag{5.50}$$

It is not clear what the distribution of ζ is. Observe that for every u, the integrand in the above integral is the absolute value of a gaussian process $\{J_t, 0 \leq t \leq T\}$ with the mean function $E(J_t) = -utx_0 e^{\theta_0 t}$ and the covariance function

$$Cov(J_t, J_s) = e^{\theta_0(t+s)}\gamma_H(t)$$

for $s \geq t$.

6. Identification for linear stochasic systems driven by fBm

We now discuss the problem of nonparametric estimation or identification of the "drift" function $\theta(t)$ for a class of stochastic processes satisfying a stochastic differential equation

$$dX_t = \theta(t)X_t dt + dW_t^H, X_0 = \tau, t \geq 0 \tag{6.1}$$

where τ is a gaussian random variable independent of the process $\{W_t^H\}$ which is a fBm with known Hurst parameter. We use the method of sieves and study the asymptotic properties of the estimator. Identification of nonstationary diffusion models by the method of sieves is studied in Nguyen and Pham (1982).

Estimation by the method of sieves

We assume that $\theta(t) \in L^2([0,T], dt)$. In other words $X = \{X_t, t \geq 0\}$ is a stochastic process satisfying the stochastic integral equation

$$X(t) = \tau + \int_0^t \theta(s)X(s)ds + W_t^H, 0 \leq t \leq T. \tag{6.2}$$

where $\theta(t) \in L^2([0,T], dt)$. Let

$$C_\theta(t) = \theta(t)\ X(t), 0 \leq t \leq T \tag{6.3}$$

and assume that the sample paths of the process $\{C_\theta(t), 0 \leq t \leq T\}$ are smooth enough so that the process

$$Q_{H,\theta}(t) = \frac{d}{dw_t^H} \int_0^t \kappa_H(t,s)C_\theta(s)ds, 0 \leq t \leq T \tag{6.4}$$

is well-defined where w_t^H and $\kappa_H(t,s)$ are as defined in (3.8) and (3.6) respectively. Suppose the sample paths of the process $\{Q_H(t), 0 \leq t \leq T\}$ belong almost surely to $L^2([0,T], dw_t^H)$. Define

$$Z_t = \int_0^t \kappa_H(t,s)dX_s, 0 \leq t \leq T. \tag{6.5}$$

Then the process $Z = \{Z_t, 0 \leq t \leq T\}$ is an $(\mathcal{F}_t)$-semimartingale with the decomposition

$$Z_t = \int_0^t Q_{H,\theta}(s)dw_s^H + M_t^H \tag{6.6}$$

where M^H is the fundamental martingale defined by (3.9) and the process X admits the representation

$$X_t = X_0 + \int_0^t K_H(t,s)dZ_s \tag{6.7}$$

where the function K_H is as defined by (3.11) with $f \equiv 1$. Let P_θ^T be the measure induced by the process $\{X_t, 0 \leq t \leq T\}$ when $\theta(.)$ is the true "drift" function. Following Theorem 3.3, we get that the Radon-Nikodym derivative of P_θ^T with respect to P_0^T is given by

$$\frac{dP_\theta^T}{dP_0^T} = \exp[\int_0^T Q_{H,\theta}(s)dZ_s - \frac{1}{2}\int_0^T Q_{H,\theta}^2(s)dw_s^H]. \tag{6.8}$$

Suppose the process X is observable on $[0,T]$ and $X_i, 1 \leq i \leq n$ is a random sample of n independent observations of the process X on $[0,T]$. Following the representation of the Radon-Nikodym derivative of P_θ^T with respect to P_0^T given above, it follows that the log-likelihood function corresponding to the observations $\{X_i, 1 \leq i \leq n\}$ is given by

$$\begin{aligned} L_n(X_1,\ldots,X_n;\theta) &\equiv L_n(\theta) \\ &= \sum_{i=1}^{n}\Big(\int_0^T Q_{H,\theta}^{(i)}(s)dZ_i(s) - \frac{1}{2}\int_0^T [Q_{H,\theta}^{(i)}]^2(s)dw_s^H\Big). \end{aligned} \tag{6.9}$$

where the process $Q_{H,\theta}^{(i)}$ is as defined by the relation (6.4) for the process X_i. For convenience in notation, we write $Q_{i,\theta}(s)$ hereafter for $Q_{H,\theta}^{(i)}(s)$.

Let $\{V_n, n \geq 1\}$ be an increasing sequence of subspaces of finite dimensions $\{d_n\}$ such that $\cup_{n\geq 1}V_n$ is dense in $L^2([0,T],dt)$. The method of sieves consists in maximizing $L_n(\theta)$ on the subspace V_n. Let $\{e_i\}$ be a set of linearly independent vectors in $L^2([0,T],dt)$ such that the set of vectors $\{e_1,\ldots,e_{d_n}\}$ is a basis for the subspace V_n for every $n \geq 1$. For $\theta \in V_n$, $\theta(.) = \sum_{j=1}^{d_n}\theta_j e_j(.)$, we have

$$\begin{aligned} Q_{i,\theta}(t) &= \frac{d}{dw_t^H}\int_0^t \kappa_H(t,s)\theta(s)X_i(s)ds \\ &= \frac{d}{dw_t^H}\int_0^t \kappa_H(t,s)[\sum_{j=1}^{d_n}\theta_j e_j(s)]X_i(s)ds \\ &= \sum_{j=1}^{d_n}\theta_j\frac{d}{dw_t^H}\int_0^t \kappa_H(t,s)e_j(s)X_i(s)ds \\ &= \sum_{j=1}^{d_n}\theta_j\Gamma_{i,j}(t) \ \text{(say)}. \end{aligned} \tag{6.10}$$

Further more

$$\begin{aligned} \int_0^T Q_{i,\theta}(t)dZ_i(t) &= \int_0^T [\sum_{j=1}^{d_n}\theta_j\Gamma_{i,j}(t)]dZ_i(t) \\ &= \sum_{j=1}^{d_n}\theta_j\int_0^T \Gamma_{i,j}(t)dZ_i(t) \\ &= \sum_{j=1}^{d_n}\theta_j R_{i,j} \ \text{(say)} \end{aligned} \tag{6.11}$$

and

$$\begin{aligned} \int_0^T Q_{i,\theta}^2(t)dw_t^H &= \int_0^T [\sum_{j=1}^{d_n}\theta_j\Gamma_{i,j}(t)]^2 dw_t^H \\ &= \sum_{j=1}^{d_n}\sum_{k=1}^{d_n}\theta_j\theta_k\int_0^T \Gamma_{i,j}(t)\Gamma_{i,k}(t)dw_t^H \\ &= \sum_{j=1}^{d_n}\sum_{k=1}^{d_n}\theta_j\theta_k < R_{i,j}, R_{i,k} > \end{aligned} \tag{6.12}$$

where $< ., . >$ denotes the quadratic covariation. Therefore the log-likelihood function corresponding to the observations $\{X_i, 1 \leq i \leq n\}$ is given by

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n}(\int_0^T Q_{i,\theta}(t)dZ_i(t) - \frac{1}{2}\int_0^T Q_{i,\theta}^2(t)dw_t^H) \\ &= \sum_{i=1}^{n}[\sum_{j=1}^{d_n}\theta_j R_{i,j} - \frac{1}{2}\sum_{j=1}^{d_n}\sum_{k=1}^{d_n}\theta_j\theta_k < R_{i,j}, R_{i,k} >] \\ &= n[\sum_{j=1}^{d_n}\theta_j B_j^{(n)} - \frac{1}{2}\sum_{j=1}^{d_n}\sum_{k=1}^{d_n}\theta_j\theta_k A_{j,k}^{(n)}] \end{aligned} \tag{6.13}$$

where

$$B_j^{(n)} = n^{-1}\sum_{i=1}^{n} R_{i,j}, \quad 1 \leq j \leq d_n \tag{6.14}$$

and

$$A_{j,k}^{(n)} = n^{-1}\sum_{i=1}^{n} < R_{i,j}, R_{i,k} >, \quad 1 \leq j, k \leq d_n. \tag{6.15}$$

Let $\theta^{(n)}, B^{(n)}$ and $A^{(n)}$ be the vectors and the matrix with elements $\theta_j, j = 1, \ldots, d_n$, $B_j^{(n)}, j = 1, \ldots, d_n$ and $A_{j,k}^{(n)}, j, k = 1, \ldots, d_n$ as defined above. Then the log-likelihood function can be written in the form

$$L_n(\theta) = n[B^{(n)}\theta^{(n)} - \frac{1}{2}\theta^{(n)\prime}A^{(n)}\theta^{(n)}]. \tag{6.16}$$

Here α' denotes the transpose of the vector α. The restricted maximum likelihood estimator $\hat{\theta}^{(n)}(.)$ of $\theta(.)$ is given by

$$\hat{\theta}^{(n)}(.) = \sum_{j=1}^{d_n}\hat{\theta}_j^{(n)}e_j(.) \tag{6.17}$$

where

$$\hat{\theta}^{(n)} = (\hat{\theta}_1^{(n)}, \ldots, \hat{\theta}_{d_n}^{(n)}) \tag{6.18}$$

is the solution of the equation

$$A^{(n)}\hat{\theta}^{(n)} = B^{(n)}. \tag{6.19}$$

Assuming that $A^{(n)}$ is invertible, we get that

$$\hat{\theta}^{(n)} = (A^{(n)})^{-1}B^{(n)}. \tag{6.20}$$

Asymptotic properties of the estimator $\hat{\theta}^{(n)}(.)$ are studied in Prakasa Rao (2004b). We do not go into the details here.

7. Remarks

(1)We have considered the stochastic differential equations of the type

$$dY_t = C(t)dt \ + \ B(t)dW_t^H, t \geq 0 \tag{7.1}$$

driven by a fBm where $B(.)$ is a nonrandom function. As was mentioned earlier, one can define a stochastic integral of a nonrandom function with respect to a

fractional Brownian motion via a suitable limit of Riemann-Steiltjes type approximating sums as was described in Gripenberg and Norros (1996). However it is not possible to extend this approach to define stochastic integration of a large class of random functions with respect to a fractional Brownian motion in view of the fact that the fractional Brownian is not a semimartingale. It is known that if a stochastic process $\{Z_t, t \geq 0\}$ has the property that the stochastic integral $\int B_t dZ_t$ is well-defined for a large class of integrands $\{B_t, t \geq 0\}$ and satisfies reasonable conditions such as linearity, dominated convergencge theorems as satisfied by integrals with respect to σ-finite measures, then the process $\{Z_t, t \geq 0\}$ has to be a semimartingale (cf. Metivier and Pellaumail (1980)). Hence the classical theory of stochastic integration with respect to a Brownian motion cannot be extended to define stochastic integration with respect to a fBm for random integrands in the usual manner. Lin (1995) and Dai and Heyde (1996) defined stochastic integrals with respect to fBm and extended the Ito formula. Their definition of a stochastic integral leads to a stochastic integral of Stratonovich type and the corresponding Ito formula is the standard chain rule for differentiation. The stochastic integral $\int B_t dZ_t$ defined by them however does not satisfy the property $E(\int B_t dZ_t) = 0$ in general which is essential for modeling purposes. Duncan et al (2000) defined stochastic integration of a random function $\{B_t, t \geq 0\}$ with respect to a fBm $\{W_t^H, t \geq 0\}, H \in (\frac{1}{2}, 1)$ using the concept of Wick product and this integral satisfies the condition $E(\int B_t dW_t^H) = 0$ whenever it is well-defined. They have also developed the correponding Ito type formula in their work. Using the notion of Skorokhod integral, Decreusefond and Ustunel (1999) developed a stochastic integral with respect to a fBm (cf. Decreusefond (2003)).

(2) We have assumed through out the Section 4 to Section 6 that a complete path of the process $\{X_t, 0 \leq t \leq T\}$ is observable and that the process is driven by a fBm with *known* Hurst index H. The problem of estimation of the index H has been studied well and a discussion is given Section 2. The problem of estimation of the parameters in the absence of knowledge of the Hurst index H remains open. It would be interesting to find whether it is possible to estimate the parameters and the index H simultaneously from a complete path of the process $\{X_t, 0 \leq t \leq T\}$. From a practical point of view, it is clear that the assumption, that a complete path of the process $\{X_t, 0 \leq t \leq T\}$ is observable, is not tenable. Suppose the process $\{X_t, 0 \leq t \leq T\}$ is observed at some discrete set of times $\{t_i, 1 \leq i \leq n\}$ in the interval $[0, T]$ where the time points $\{t_i, 1 \leq i \leq n\}$ could be nonrandom or random as well as equally spaced or irregularly spaced. If the process is observed at a set of discrete times, then the problems of estimation of the parameters involved as well as the estimation of Hurst index in case it is unknown remain open. It would be interesting to study these problems for the models discussed in this paper. A general discussion on statistical inference from sampled data for stochastic processes is given in Prakasa Rao (1988). Results for the special case of diffusion type processes are studied in Prakasa Rao (1999a).

References

Abry, P. and Sellan, F. (1996) The wavelet-based synthesis for fractional Brownian motion proposed by F.Sellan and Y.Meyer: Remarks and fast implementation. *Applied and Computational Harmonic Analysis*, **3**, 377–383. MR1420505

Azais, J.M. (1990) Condition for convergence of number of crossings to the lo-

cal time. Application to stable processes with independent increments and to gaussian processes. *Probab. Math. Stat.*, **11**, 19–36.

Beran, J. and Terrin, N. (1994) Estimation of the long-memory parameter based on multivariate central limit theorem, *J. Time Ser. Anal.* ,**15**, 269–278.

Borwanker, J.D., Kallianpur, G. and Prakasa Rao, B.L.S. (1971) The Bernstein-von Mises theorem for Markov processes, *Ann. Math. Statist.*, **42**, 1241–1253. MR298811

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Stanley, H.E., Stanley, M.H.R. and Simons, M. (1993) Fractal landscapes and molecular evolution: modeling the Moyosin heavy chain gene family. *Biophys. J.*, **65**, 2673–2679.

Chen, G., Hall, P. and Poskitt, D.S. (1995) Periodogram-based estimators of fractal properties. *Ann. Statist.*, **23**, 1684–1711.

Coeurjolly, Jean-Francois. (2001) Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths, *Statist. Infer. Stoch. Proc.*, **4**, 199–207.

Comte, F. (1996) Simulation and estimation of long memory continuous time models, *J. Time Ser. Anal.*, **17**, 19–36.

Constantine, A.G. and Hall, P. (1994) Characterizing surface smoothness via estimation of effective fractal dimension, *J. Roy. Statist. Soc.*, **56**, 97–113.

Dai, W. and Heyde, C.C. (1996) Ito formula with respect fractional Brownian motion and its application, *J. Appl. Math. Stochastic Anal.*, **9**, 439–448.

Decreusefond, L. and Ustunel, A.S. (1999) Stochastic analysis of the fractional Brownian motion, *Potential Anal.*, **10**, 177–214. MR1696137

Decreusefond, L. (2003) Stochastic integration with respect to fractional Brownian motion, In *Theory of Long-Range Dependence* , Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp. 203–226.

Doukhan, P., Oppenheim, G. and Taqqu, M.S. (2003)*Theory of Long-Range Dependence*, Birkhauser, Boston.

Duncan, T., Hu, Y., and Pasik-Duncan, B. (2000) Stochastic calculus for fractional Brownian motion, *SIAM J. Control. Opt.*, **38**, 582–612.

Feuverger, A., Hall. P. and Wood, A.T.A. (1994) Estimation of the fractal index and fractal dimension of a gaussian process by counting the number of level crossings, *J. Time Ser. Anal.* , **15**, 587–606.

Flandrin, P., and Goncalves, P. (1994) From wavelets to time-scale energy distributions, In *Recent Advances in wavelet Analysis*, Ed. L.L.Schumaker and G.Webb, Academic Press, Boston, pp. 309–334. MR1244601

Geweke, J. and Porter-Hudak, S. (1983) The estimation and application of long-memory time series models, *J. Time Ser. Anal.*, **4**, 221–238.

Goncalves, P and Flandrin, P. (1993) Bilinear time-scale analysis applied to a local scaling exponents estimation, In *Progress in Wavelet Analysis and Applications*, Ed. Y.Meyer and S.Roques, Frontieres, Paris, pp. 271–276.

Gripenberg, G. and Norros, I. (1996) On the prediction of fractional Brownian motion, *J. Appl. Prob.*, **33**, 400–410.

Hall, P., Koul, H.L. and Turlach, B.A. (1997) Note on convergence rates of semiparametric estimators of dependence index, *Ann. Statist.*, **25**, 1725–1739.

Henry, M. and Zafforoni, P. (2003) The long-range dependence pardigm for macroeconomics and finance, In *Theory of Long-Range Dependence*, Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp. 417–438. MR1956041

Hurst, H.E. (1951) Long term storage capacity of reservoirs (with discussion), *Trans. Amer. Soc. Civ. Eng.*, **116**, 770–808.

Jensen, M. (1998) An approximate wavelet MLE of short and long memory parameters, *Studies in Nonlinear Dynamics and Econometrics*, **3**, 239–253.

Kent, J.T. and Wood, A,T.A. (1997) Estimating the fractal dimension of a locally self-similar gaussian process by using increments, *J. Roy. Statist. Soc. B* , **59**, 679–700.

Kleptsyna, M.L. and Le Breton, A. (2002) Statistical analysis of the fractional Ornstein-Uhlenbeck type process, *Statist. Infer. Stoch. Proc.*, **5**, 229–248.

Kleptsyna, M.L. and Le Breton, A. and Roubaud, M.-C. (2000) Parameter estimation and optimal filtering for fractional type stochastic systems, *Statist. Infer. Stoch. Proc.*, **3**, 173–182.

Kutoyants, Yu. and Pilibossian, P. (1994) On minimum L_1-norm estimate of the parameter of the Ornstein-Uhlenbeck process, *Statist. Probab. Lett.*, **20**, 117–123.

Le Breton, A. (1998) Filtering and parameter estimation in a simple linear model driven by a fractional Brownian motion, *Statist. Probab. Lett.*, **38**, 263–274.

Lin, S.J. (1995) Stochastic analysis of fractional Brownian motions, *Stochastics Stochastics Rep.*, **55**, 121–140.

Mc Coy, E.J. and Walden, A.T. (1996) Wavelet analysis and synthesis of stationary long memory processes, *J. Comput. Graphical Statist.*, **5**, 26–56.

Metivier, M. and Pellaumail, J. (1980) *Stochastic Integration*, Academic Press, New York.

Millar, P.W. (1984) A general approach to the optimality of the minimum distance estimators, *Trans. Amer. Math. Soc.*, **286**, 249–272.

Montanari, A. (2003) Long-range dependence in hydrology, In *Theory of Long-Range Dependence*, Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp. 461–472. MR1956041

Nguyen, Hung T. and Pham, Tuan D. (1982) Identification of nonstationary diffusion model by the method of sieves, *SIAM J. Control and Optimization*, **20**, 603–611.

Norros, I. (2003) Large deviations of queues with long-range dependent input, In *Theory of Long-Range Dependence*, Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp. 409–415. MR1956041

Norros, I., Valkeila, E., and Virtamo, J. (1999) An elementary approach to a Girsanov type formula and other analytical results on fractional Brownian motion, *Bernoulli*, **5**, 571–587.

Novikov, A.A. (1972) Sequential estimation of the parameters of diffusion proocesses, *Mathematical Notes*, **12**, 812–818.

Novikov, A.A. and Valkeila, E. (1999) On some maximal inequalities for fractional Brownian motion, *Statist. Probab. Lett.*, **44**, 47–54. MR1706362

Ossandik, S.M. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.-K., Simons, M. and Stanley, H.E. (1993) Correlation approach to identify coding regions in DNA sequences, *Biophys. J.*, **67**, 64–70.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M and Stanley, H.E. (1992) Long-range correlation in nucleotide sequences, *Nature*, **356**, 168–170.

Peng, C.-K., Hausdorff, J.M., Mietus, J.E., Havlin,S., Stanley, H.E. and Goldberger, A.L. (1995a) Fractals in physiological control from heartbeat to gait, In *Levy Flights and Related Phenomena in Physics*, Ed. M.F. Shlesinger,, G.M. Zaslavsky, and U.Frisch, Proceedings of the 1994 International Conference on Levy flights, Springer, Berlin, pp. 315–330. MR1381481

Peng, C.-K., Havlin, S., Stanley, H.E., and Goldberger, A.L. (1995b) Quantification of scaling exponents and crossover phenomena in noinstationary heartbeat time series, *Chaos*, **5**, 82–87.

Percival, D.P. and Guttorp, P. (1994) Long-memory processes, the Allan variance and wavelets, In *Wavelets in Geophysics*, Ed. Foufoula-Georgiou, E. and Kumar, P., Academic Press, New York, pp. 325–357. MR1295674

Poggi, Jean-Michel. and Viano, Marie-Claude. (1998) An estimate of the fractal index using multiscale aggregates, *J. Time Ser. Anal.*, **19**, 221–233.

Prakasa Rao, B.L.S. (1966) *Asymptotic Distributions in Some Nonregular Statistical Problems*, Ph.D. Dissertation, Michigan State University.

Prakasa Rao, B.L.S. (1968) Estimation of the location of the cusp of a continuous density, *Ann. Math. statist.*, **39**, 76–87.

Prakasa Rao, B.L.S. (1969) Estimation of a unimodal density, *Sankhya Ser.A*, **31**, 23–36.

Prakasa Rao, B.L.S. (1970) Estimation for distributions with monotone failure rate, *Ann. Math. Statist.*, **41**, 507–519.

Prakasa Rao, B.L.S. (1981) The Bernstein- von Mises theorem for a class of diffusion processes, *Teor. Sluch. Proc.*, **9**, 95–101 (In Russian).

Prakasa Rao, B.L.S. (1987) *Asymptotic Theory of Statistical Inference*, Wiley, New York.

Prakasa Rao, B.L.S. (1988) Statistical inference from sampled data for stochastic processes, In *Statistical Inference from Stochastic Processes*, Contemporary Mathematics, Ed. N.U.Prabhu, Vol.**80**, Amer. Math. Soc., pp. 249–284. MR999004

Prakasa Rao, B.L.S. (1999a) *Statistical Inference for Diffusion Type Processes*, Arnold, London and Oxford University Press, New York.

Prakasa Rao, B.L.S. (1999b) *Semimartingales and Their Statistical Inference*, CRC Press, Boca Raton and Chapman and Hall, London.

Prakasa Rao, B.L.S. (2003a) Parametric estimation for linear stochastic differential equations driven by fractional Brownian motion, *Random Operators and Stochastic Equations*, **11**, 229–242.

Prakasa Rao, B.L.S. (2003b) Berry-Esseen bound for MLE for linear stochastic differential equations driven by fractional Brownian motion, Preprint, Indian Statistical Institute, New Delhi.

Prakasa Rao, B.L.S. (2003c) Minimum L_1-norm estimation for fractional Ornstein-Uhlenbeck type process, *Theory of Probability and Mathematical Statistics* (To appear).

Prakasa Rao, B.L.S. (2004a) Sequential estimation for fractional Ornstein-Uhlenbeck type process, *Sequential Analysis* **23** 33–44.

Prakasa Rao, B.L.S. (2004b) Identification for linear stochastic systems driven by fractional Brownian motion, *Stoch. Anal. Appl.* (To appear).

Robinson, P. (1995) Gaussian semiparametric estimation of long range dependence, *Ann. Statist.*, **23**, 1630–1661.

Samko, S.G., Kilbas, A.A., and Marichev, O.I. (1993) *Fractional Integrals and Derivatives: Theory and Applications*, Gordon and Breach Science, Yverdon.

Taqqu, M. (2003) Fractional Brownian motion and long-range dependence, In *Theory of Long-Range Dependence*, Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp.5–38. MR1956041

Taylor, C.C. and Taylor, S.J. (1991) Estimating the dimension of a fractal, *J. Roy. Statist. Soc. B*, **53**, 353–364.

Wang, Y., Cavanaugh. J.E., and Song, C. (2001) Self-similarity index estimation via wavelets for locally self-similar processes, *J. Stat. Plan. Infer.*, **99**, 91–110. MR1858712

Willinger, W., Paxson, V., Riedi, R.H. and Taqqu, M. (2003) Long-range dependence and data network traffic, In *Theory of Long-Range Dependence*, Ed. Doukhan, P., Oppenheim, G. and Taqqu, M.S., Birkhauser, Boston, pp. 373–407. MR1956041

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 126–137
© Institute of Mathematical Statistics, 2004

Some properties of the arc-sine law related to its invariance under a family of rational maps*

Jim Pitman[1] and Marc Yor[2]

University of California, Berkeley

Université Pierre et Marie Curie – Boîte

Abstract: This paper shows how the invariance of the arc-sine distribution on $(0,1)$ under a family of rational maps is related on the one hand to various integral identities with probabilistic interpretations involving random variables derived from Brownian motion with arc-sine, Gaussian, Cauchy and other distributions, and on the other hand to results in the analytic theory of iterated rational maps.

1. Introduction

Lévy [20, 21] showed that a random variable A with the arc-sine law

$$P(A \in da) = \frac{da}{\pi\sqrt{a(1-a)}} \qquad (0 < a < 1) \tag{1}$$

can be constructed in numerous ways as a function of the path of a one-dimensional Brownian motion, or more simply as

$$A = \tfrac{1}{2}(1 - \cos\Theta) \stackrel{d}{=} \tfrac{1}{2}(1 - \cos 2\Theta) = \cos^2\Theta \tag{2}$$

where Θ has uniform distribution on $[0, 2\pi]$ and $\stackrel{d}{=}$ denotes equality in distribution. See [31, 7] and papers cited there for various extensions of Lévy's results. In connection with the distribution of local times of a Brownian bridge [29], an integral identity arises which can be expressed simply in terms of an arc-sine variable A. Section 5 of this note shows that this identity amounts to the following property of A, discovered in a very different context by Cambanis, Keener and Simons [6, Proposition 2.1]: for all real a and c

$$\frac{a^2}{A} + \frac{c^2}{1-A} \stackrel{d}{=} \frac{(|a| + |c|)^2}{A}. \tag{3}$$

As shown in [6], where (3) is applied to the study of an interesting class of multivariate distributions, the identity (3) can be checked by a computation with densities, using (2) and trigonometric identities. Here we offer some derivations of (3) related

*Research supported in part by N.S.F. Grants 97-03961 and DMS-0071448.

[1]Department of Statistics, University of California, 367 Evans Hall # 3860, Berkeley, CA 94720-3860, USA. e-mail: pitman@Stat.Berkeley.edu

[2]Laboratoire de Probabilités et Modèles aléatoires, Université Pierre et Marie Curie – Boîte courrier 188, 75252 Paris Cedex 05, France. e-mail: deaproba@proba.jussieu.fr

Keywords and phrases: invariant measure, harmonic measure, Brownian motion, conformal invariance, Cauchy distribution, inner function.

AMS 2000 subject classifications: primary 58F11; secondary 31A15, 60J65, 30D05.

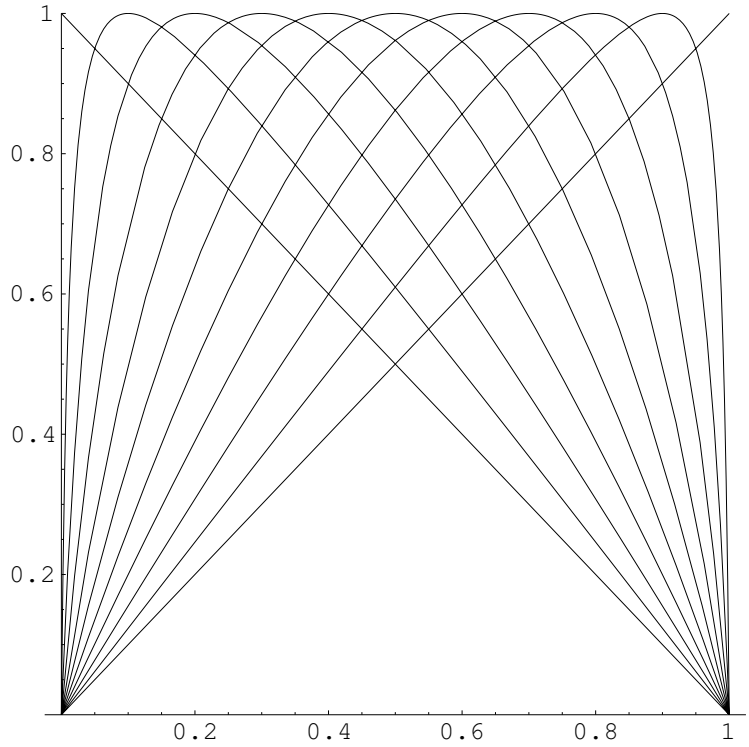


Figure 1: Graphs of $Q_u(a)$ for $0 \le a \le 1$ and $u = k/10$ with $k = 0, 1, \ldots, 10$.

to various other characterizations and properties of the arc-sine law. For $u \in [0,1]$ define a rational function

$$Q_u(a) := \left(\frac{u^2}{a} + \frac{(1-u)^2}{1-a} \right)^{-1} = \frac{a(1-a)}{u^2 + (1-2u)a} \tag{4}$$

So (3) amounts to $Q_u(A) \stackrel{d}{=} A$, as restated in the following theorem. It is easily checked that Q_u increases from 0 to 1 over $(0,u)$ and decreases from 1 to 0 over $(u,1)$, as shown in the above graphs.

Theorem 1. *For each $u \in (0,1)$ the arc-sine law is the unique absolutely continuous probability measure on the line that is invariant under the rational map $a \to Q_u(a)$.*

The conclusion of Theorem 1 for $Q_{1/2}(a) = 4a(1-a)$ is a well known result in the theory of iterated maps, dating back to Ulam and von Neumann [38]. As indicated in [3] and [22, Example 1.3], this case follows immediately from (2) and the ergodicity of the Bernoulli shift $\theta \mapsto 2\theta \pmod{2\pi}$. This argument shows also, as conjectured in [15, p. 464 (A3)] and contrary to a footnote of [37, p. 233], that the arc-sine law is not the only non-atomic law of A such that $4A(1-A) \stackrel{d}{=} A$. For the argument gives $4A(1-A) \stackrel{d}{=} A$ if $A = (1 - \cos 2\pi U)/2$ for any distribution of U on $[0,1]$ with $(2U \bmod 1) \stackrel{d}{=} U$, and it is well known that such U exist with singular continuous distributions, for instance $U = \sum_{m=1}^{\infty} X_m 2^{-m}$ for X_m independent Bernoulli(p) for any $p \in (0,1)$ except $p = 1/2$. See also [15] and papers

cited there for some related characterizations of the arc-sine law, and [13] where this property of the arc-sine law is related to duplication formulae for various special functions defined by Euler integrals.

Section 2 gives a proof of Theorem 1 based on a known characterization of the standard Cauchy distribution. In terms of a complex Brownian motion Z, the connection between the two results is that the Cauchy distribution is the hitting distribution on $\mathbb{R}$ for $Z_0 = \pm i$, while the arc-sine law is the hitting distribution on $[0,1]$ for $Z_0 = \infty$. The transfer between the two results may be regarded as a consequence of Lévy's theorem on the conformal invariance of the Brownian track. In Section 4 we use a closely related approach to generalize Theorem 1 to a large class of functions Q instead of Q_u. The result of this section for rational Q can also be deduced from the general result of Lalley [18] regarding Q-invariance of the equilibrium distribution on the Julia set of Q, which Lalley obtained by a similar application of Lévy's theorem.

2. Proof of Theorem 1

Let A have the arc-sine law (1), and let C be a standard Cauchy variable, that is

$$P(C \in dy) = \frac{dy}{\pi(1+y^2)} \qquad (y \in \mathbb{R}). \tag{5}$$

We will exploit the following elementary fact [33, p. 13]:

$$A \stackrel{d}{=} 1/(1+C^2). \tag{6}$$

Using (6) and $C \stackrel{d}{=} -C$, the identity (3) is easily seen to be equivalent to

$$uC - (1-u)/C \stackrel{d}{=} C. \tag{7}$$

This is an instance of the result of E. J. G. Pitman and E. J. Williams [28] that for a large class of meromorphic functions G mapping the half plane $\mathbb{H}^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ to itself, with boundary values mapping $\mathbb{R}$ (except for some poles) to $\mathbb{R}$, there is the identity in distribution

$$G(C) \stackrel{d}{=} \operatorname{Re} G(i) + (\operatorname{Im} G(i))C \tag{8}$$

where $i = \sqrt{-1}$ and $z = \operatorname{Re} z + i \operatorname{Im} z$. Kemperman [14] attributes to Kesten the remark that (8) follows from Lévy's theorem on the conformal invariance of complex Brownian motion Z, and the well known fact that for τ the hitting time of the real axis by Z, the distribution of Z_τ given $Z_0 = z$ is that of $\operatorname{Re} z + (\operatorname{Im} z)C$. As shown by Letac [19], this argument yields (8) for all *inner functions on* $\mathbb{H}^+$, that is all holomorphic functions G from $\mathbb{H}^+$ to $\mathbb{H}^+$ such that the boundary limit $G(x) := \lim_{y \downarrow 0} G(x+iy)$ exists and is real for Lebesgue almost every real x. In particular, (8) shows that

$$\text{if } G \text{ is inner on } \mathbb{H}^+ \text{ with } G(i) = i, \text{ then } G(C) \stackrel{d}{=} C. \tag{9}$$

As indicated by E. J. Williams [39] and Kemperman [14], for some inner G on $\mathbb{H}^+$ with $G(i) = i$, the property $G(C) \stackrel{d}{=} C$ characterizes the distribution of C among all absolutely continuous distributions on the line. These are the G whose action

on $\mathbb{R}$ is ergodic relative to Lebesgue measure. Neuwirth [26] showed that an inner function G with $G(i) = i$ is ergodic if it is not one to one. In particular,

$$G_u(z) := uz - (1-u)/z \tag{10}$$

as in (7) is ergodic. The above transformation from (3) to (7) amounts to the semi-conjugacy relation

$$Q_u \circ \gamma = \gamma \circ G_u \text{ where } \gamma(w) := 1/(1+w^2). \tag{11}$$

So Q_u acts ergodically as a measure preserving transformation of $(0,1)$ equipped with the arc-sine law. It is easily seen that for $u \in (0,1)$ a Q_u-invariant probability measure must be concentrated on $[0,1]$, and Theorem 1 follows.

See also [35, p. 58] for an elementary proof of (7), [1, 23, 24, 2] for further study of the ergodic theory of inner functions, [16, 19] for related characterizations of the Cauchy law on $\mathbb{R}$ and [17, 9] for extensions to $\mathbb{R}^n$.

3. Further interpretations

Since $w \mapsto 1/(1+w^2)$ maps i to ∞, another application of Lévy's theorem shows that the arc-sine law of $1/(1+C^2)$ is the hitting distribution on $[0,1]$ of a complex Brownian motion plane started at ∞ (or uniformly on any circle surrounding $[0,1]$). In terms of classical planar potential theory [32, Theorem 4.12], the arc-sine law is thus identified as the *normalized equilibrium distribution* on $[0,1]$. The corresponding characterization of the distribution of $1-2A$ on $[-1,1]$ appears in Brolin [5], in connection with the invariance of this distribution under the action of Chebychev polynomials, as discussed further in the next section. Equivalently by inversion, the distribution of $1/(1-2A)$ is the hitting distribution on $(-\infty,1]\cup[1,\infty)$ for complex Brownian motion started at 0. Spitzer [36] found this hitting distribution, which he interpreted further as the hitting distribution of $(-\infty,1]\cup[1,\infty)$ for a Cauchy process starting at 0. This Cauchy process is obtained from the planar Brownian motion watched only when it touches the real axis, via a time change by the inverse local time at 0 of the imaginary part of the Brownian motion. The arc-sine law can be interpreted similarly as the limit in distribution as $|x| \to \infty$ of the hitting distribution of $[0,1]$ for the Cauchy process started at $x \in \mathbb{R}$. See also [30] for further results in this vein.

4. Some generalizations

We start with some elementary remarks from the perspective of ergodic theory. Let $\lambda(a) := 1-2a$, which maps $[0,1]$ onto $[-1,1]$. Obviously, a Borel measurable function $f^\dagger$ has the property

$$f^\dagger(A) \stackrel{d}{=} A \tag{12}$$

for A with arc-sine law if and only if

$$\tilde{f}(1-2A) \stackrel{d}{=} 1-2A \text{ where } \tilde{f} = \lambda \circ f^\dagger \circ \lambda^{-1}. \tag{13}$$

Let $\rho(z) := \frac{1}{2}(z+z^{-1})$, which projects the unit circle onto $[-1,1]$. It is easily seen from (2) that (13) holds if and only if there is a measurable map f from the circle to itself such that

$$f(U) \stackrel{d}{=} U \text{ and } \tilde{f} \circ \rho(u) = \rho \circ f(u) \text{ for } |u| = 1 \tag{14}$$

where U has uniform distribution on the unit circle. In the terminogy of ergodic theory [27], every transformation $f^\dagger$ of $[0,1]$ which preserves the arc-sine law is thus a *factor* of some non-unique transformation f of the circle which preserves Lebesgue measure. Moreover, this f can be taken to be *symmetric*, meaning

$$f(\overline{z}) = \overline{f(z)}.$$

If f acts ergodically with respect to Lebesgue measure on the circle, then $f^\dagger$ acts ergodically with respect to Lebesgue measure on $[0,1]$, hence the arc-sine law is the unique absolutely continuous $f^\dagger$-invariant measure on $[0,1]$. This argument is well known in case $f(z) = z^d$ for $d = 2, 3, \ldots$, when it is obvious that (14) holds and well known that f is ergodic. Then $\tilde{f}(x) = T_d(x)$, the dth *Chebychev polynomial* [34] and we recover from (14) the well known result ([3],[34, Theorem 4.5]) that

$$T_d(1-2A) \stackrel{d}{=} 1-2A \qquad (d = 1, 2, \ldots). \tag{15}$$

Let $\mathbb{D} := \{z : |z| < 1\}$ denote the unit disc in the complex plane. An *inner function on* $\mathbb{D}$ is a function defined and holomorphic on $\mathbb{D}$, with radial limits of modulus 1 at Lebesgue almost every point on the unit circle. Let $\phi(z) := i(1+z)/(1-z)$ denote the Cayley bijection from $\mathbb{D}$ to the upper half-plane $\mathbb{H}^+$. It is well known that the inner functions G on $\mathbb{H}^+$, as considered in Section 2, are the conjugations $G = \phi \circ f \circ \phi^{-1}$ of inner functions f on $\mathbb{D}$. So either by conjugation of (9), or by application of Lévy's theorem to Brownian motion in $\mathbb{D}$ started at 0,

$$\text{if } f \text{ is inner on } \mathbb{D} \text{ with } f(0) = 0, \text{ then } f(U) \stackrel{d}{=} U \tag{16}$$

where U is uniform on the unit circle. If f is an inner function on $\mathbb{D}$ with a fixed point in $\mathbb{D}$, and f is not one-to-one, then f acts ergodically on the circle [26]. The only one-to-one inner functions with $f(0) = 0$ are $f(z) = cz$ for some c with $|c| = 1$. By combining the above remarks, we obtain the following generalization of (15), which is the particular case $f(z) = z^d$:

Theorem 2. *Let f be a symmetric inner function on $\mathbb{D}$ with $f(0) = 0$. Define the transformation $\tilde{f}$ on $[-1,1]$ via the semi-conjugation*

$$\tilde{f} \circ \rho(z) = \rho \circ f(z) \text{ for } |z| = 1, \text{ where } \rho(z) := \tfrac{1}{2}\big(z + z^{-1}\big). \tag{17}$$

If A has arc-sine law then

$$\tilde{f}(1-2A) \stackrel{d}{=} 1-2A. \tag{18}$$

Except if $f(z) = z$ or $f(z) = -z$, the arc-sine law is the only absolutely continuous law of A on $[0,1]$ with this property.

It is well known that every inner function f which is continuous on the closed disc is a *finite Blaschke product*, that is a rational function of the form

$$f(z) = c \prod_{i=1}^{d} \frac{z - a_i}{1 - \overline{a}_i z} \tag{19}$$

for some complex c and a_i with $|c| = 1$ and $|a_i| < 1$. Note that $f(0) = 0$ iff some $a_i = 0$, and that f is symmetric iff $c = \pm 1$ with some a_i real and the rest of the a_i forming conjugate pairs. In particular, if we take $c = 1, a_1 = 0, a_2 = a \in (-1,1)$, we find that the degree two Blaschke product

$$f_a(z) := z\frac{(z-a)}{(1-az)} = \frac{z-a}{z^{-1}-a}$$

for $a = 1 - 2u$ is the conjugate via the Cayley map $\phi(z) := i(1+z)/(1-z)$ of the function $G_u(w) = uw - (1-u)/w$ on $\mathbb{H}^+$, which appeared in Section 2. For $f = f_{1-2u}$ the semi-conjugation (17) is the equivalent via conjugation by ϕ of the semi-conjugation (11). So for instance

$$Q_u \circ \gamma \circ \phi = \gamma \circ \phi \circ f_{1-2u} \quad \text{where} \quad \gamma \circ \phi(z) = \frac{-(1-z)^2}{4z} \tag{20}$$

so that

$$\gamma \circ \phi(z) = \tfrac{1}{2}(1 - \operatorname{Re} z) \text{ if } |z| = 1,$$

and Theorem 1 can be read from Theorem 2.

Consider now a rational function R as a mapping from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$ where $\overline{\mathbb{C}}$ is the Riemann sphere. A subset A of $\overline{\mathbb{C}}$ is *completely R-invariant* if A is both forward and backward invariant under R: for $z \in \overline{\mathbb{C}}$, $z \in A \Leftrightarrow R(z) \in A$. Beardon [4, Theorem 1.4.1] showed that for R a polynomial of degree $d \geq 2$, the interval $[-1,1]$ is completely R-invariant iff R is T_d or $-T_d$. A similar argument yields

Proposition 3. *Let f be a symmetric finite Blaschke product of degree d. Then there exists a unique rational function $\tilde{f}$ which solves the functional equation*

$$\tilde{f} \circ \rho(z) = \rho \circ f(z) \text{ for } z \in \overline{\mathbb{C}}, \text{ where } \rho(z) := \tfrac{1}{2}\big(z + z^{-1}\big). \tag{21}$$

This $\tilde{f}$ has degree d, and $[-1,1]$ is completely $\tilde{f}$-invariant. Conversely, if $[-1,1]$ is completely R-invariant for a rational function R, then $R = \tilde{f}$ for some such f.

Proof. Note that ρ maps the circle with ± 1 removed in a two to one fashion to $(-1,1)$, while ρ fixes ± 1, and maps each of $\mathbb{D}$ and $\mathbb{D}^* := \{z : |z| > 1\}$ bijectively onto $[-1,1]^c := \overline{\mathbb{C}} \backslash [-1,1]$. Let us choose to regard

$$\rho^{-1}(w) = w + i\sqrt{1-w^2}$$

as mapping $[-1,1]^c$ to $\mathbb{D}$. Then $\tilde{f} := \rho \circ f \circ \rho^{-1}$ is a well defined mapping of $[-1,1]^c$ to itself. Because f is continuous and symmetric on the unit circle, this $\tilde{f}$ has a continuous extension to $\overline{\mathbb{C}}$, which maps $[-1,1]$ to itself. So $\tilde{f}$ is continuous from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$, and holomorphic on $[-1,1]^c$. It follows that $\tilde{f}$ is holomorphic from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$, hence $\tilde{f}$ is rational. Clearly, $\tilde{f}$ leaves $[-1,1]$ completely invariant.

Conversely, if $[-1,1]$ is completely R-invariant for a rational function R, then we can define $f := \rho^{-1} \circ R \circ \rho$ as a holomorphic map $\mathbb{D}$ to $\mathbb{D}$. Because R preserves $[-1,1]$ we find that f is continuous and symmetric on the boundary of $\mathbb{D}$. Hence f is a Blaschke product, which must be symmetric also on $\mathbb{D}$ by the Cauchy integral representation of f. □

As a check, Proposition 3 combines with Theorem 2 to yield the special case $K = [-1,1]$ of the following result:

Theorem 4. Lalley [18] *Let K be a compact non-polar subset of $\mathbb{C}$, and suppose that K is completely R-invariant for a rational mapping R with $R(\infty) = \infty$. Then the equilibrium distribution on K is R-invariant.*

Proof. Lalley gave this result for $K = J(R)$, the Julia set of a rational mapping R, as defined in any of [5, 22, 4, 18], assuming that $R(\infty) = \infty \notin J(R)$. Then K is necessarily compact, non-polar, and completely R-invariant. His argument, which we now recall briefly, shows that these properties of K are all that is required

for the conclusion. The argument is based on the fact [32, Theorem 4.12] that the normalized equilibrium distribution on K is the hitting distribution of K for a Brownian motion Z on $\overline{\mathbb{C}}$ started at ∞. Stop Z at the first time τ that it hits K. By Lévy's theorem, and the complete R-invariance of K, the path $(R(Z_t), 0 \le t \le \tau)$ has (up to a time change) the same law as does $(Z_t, 0 \le t \le \tau)$. So the distribution of the endpoint Z_τ is R-invariant. □

According to a well known result of Fatou [22, p. 57], the Julia set of a Blaschke product f is either the unit circle or a Cantor subset of the circle. According to Hamilton [11, p. 281], the former case obtains iff the action of f on the circle is ergodic relative to Lebesgue measure. Hamilton [12, p. 88] states that a rational map R has $[-1, 1]$ as its Julia set iff R is of the form described in Proposition 3 for some symmetric and ergodic Blaschke product f. In particular, for the Chebychev polynomial T_d it is known [4] that $J(T_d) = [-1, 1]$ for all $d \ge 2$, and [25, Theorem 4.3 (ii)] that $J(Q_u) = [0, 1]$ for all $0 < u < 1$. Typically of course, the Julia set of a rational function is very much more complicated than an interval or smooth curve [22, 4, 8].

Returning to consideration of the arc-sine law, it can be shown by elementary arguments that if Q preserves the arc-sine law on $[0, 1]$ and $Q(a) = P_2(a)/P_1(a)$ with P_i a polynomial of degree i, then $Q = Q_u$ or $1 - Q_u$ for some $u \in [0, 1]$. This and all preceding results are consistent with the following:

Conjecture 5. *Every rational function R which preserves the arc-sine law on $[0, 1]$ is of the form $R(a) = \frac{1}{2}(1 - \tilde{f}(1-2a))$ where $\tilde{f}$ is derived from a symmetric Blaschke product f with $f(0) = 0$, as in Theorem 2.*

5. Some integral identities

Let $(B_t, t \ge 0)$ denote a standard one-dimensional Brownian motion. Let

$$\varphi(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; \qquad \overline{\Phi}(x) := \int_x^\infty \varphi(z)\, dz = P(B_1 > x).$$

According to formula (13) of [29], the following identity gives two different expressions for the conditional probability density $P(B_U \in dx \,|\, B_1 = b)/dx$ for U with uniform distribution on $[0, 1]$, assumed independent of $(B_t, t \ge 0)$:

$$\int_0^1 \frac{1}{\sqrt{u(1-u)}}\, \varphi\left(\frac{x - bu}{\sqrt{u(1-u)}}\right) du = \frac{\overline{\Phi}(|x| + |b - x|)}{\varphi(b)}. \tag{22}$$

The first expression reflects the fact that B_u given $B_1 = b$ has normal distribution with mean bu and variance $u(1 - u)$, while the second was derived in [29] by consideration of Brownian local times. Multiply both sides of (22) by $\sqrt{2/\pi}$ to obtain the following identity for A with the arc-sine law (1): for all real x and b

$$E\left[\exp\left(-\frac{1}{2}\frac{(x - bA)^2}{A(1-A)}\right)\right] = 2\, e^{b^2/2}\, \overline{\Phi}(|x| + |b - x|). \tag{23}$$

Now

$$\frac{(x - bA)^2}{A(1-A)} = \frac{x^2}{A} + \frac{(x-b)^2}{1-A} - b^2 \overset{d}{=} \frac{(|x| + |b - x|)^2}{A} - b^2 \tag{24}$$

where the equality in distribution is a restatement of (3). So (23) amounts to the identity

$$E\left[\exp\left(-\frac{1}{2}\left(\frac{x^2}{A} + \frac{y^2}{1-A}\right)\right)\right] = 2\, \overline{\Phi}(|x| + |y|) \tag{25}$$

for arbitrary real x, y. Moreover, the identity in distribution (3) allows (25) to be deduced from its special case $y = 0$, that is

$$E\left[\exp\left(-\frac{x^2}{2A}\right)\right] = 2\overline{\Phi}(|x|), \tag{26}$$

which can be checked in many ways. For instance, $P(1/A \in dt) = dt/(\pi t\sqrt{t-1})$ for $t > 1$ so (26) reduces to the known Laplace transform [10, 3.363]

$$\frac{1}{2\pi}\int_1^\infty \frac{1}{t\sqrt{t-1}}\, e^{-\lambda t}\, dt = \overline{\Phi}(\sqrt{2\lambda}) \qquad (\lambda \geq 0). \tag{27}$$

This is verified by observing that both sides vanish at $\lambda = \infty$ and have the same derivative with respect to λ at each $\lambda > 0$. Alternatively, (26) can be checked as follows, using the Cauchy representation (6). Assuming that C is independent of B_1, we can compute for $x \geq 0$

$$E\left[\exp\left(-\frac{1}{2}\frac{x^2}{A}\right)\right] = e^{-\frac{1}{2}x^2} E\left[\exp(ixCB_1)\right] = e^{-\frac{1}{2}x^2} E\left[\exp(-x|B_1|)\right] = 2\overline{\Phi}(x). \tag{28}$$

We note also that the above argument allows (24) and hence (3) to be deduced from (23) and (26), by uniqueness of Laplace transforms.

By differentiation with respect to x, we see that (25) is equivalent to

$$E\left[\frac{x}{A}\exp\left(-\frac{1}{2}\left(\frac{x^2}{A}+\frac{y^2}{1-A}\right)\right)\right] = \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x+y)^2} \qquad (x > 0, y \geq 0). \tag{29}$$

That is to say, for each $x > 0$ and $y \geq 0$ the following function of $u \in (0,1)$ defines a probability density on $(0,1)$:

$$f_{x,y}(u) := \frac{x}{\sqrt{2\pi u^3(1-u)}}\, \exp\left[\frac{1}{2}\left((x+y)^2 - \frac{x^2}{u} - \frac{y^2}{1-u}\right)\right]. \tag{30}$$

This was shown by Seshadri [35, §p. 123], who observed that $f_{x,y}$ is the density of $T_{x,y}/(1+T_{x,y})$ for $T_{x,y}$ with the inverse Gaussian density of the hitting time of x by a Brownian motion with drift y. In particular, $f_{x,0}$ is the density of $x^2/(x^2+B_1^2)$. See also [29, (17)] regarding other appearances of the density $f_{x,0}$.

6. Complements

The basic identity (3) can be transformed and checked in another way as follows. By uniqueness of Mellin transforms, (3) is equivalent to

$$\frac{u^2}{A\varepsilon_2} + \frac{(1-u)^2}{(1-A)\varepsilon_2} \overset{d}{=} \frac{1}{A\varepsilon_2} \tag{31}$$

where ε_2 is an exponential variable with mean 2, assumed independent of A. But it is elementary and well known that $A\varepsilon_2$ and $(1-A)\varepsilon_2$ are independent with the same distribution as B_1^2. So (31) amounts to

$$\frac{u^2}{X^2} + \frac{(1-u)^2}{Y^2} \overset{d}{=} \frac{1}{X^2} \tag{32}$$

where X and Y are independent standard Gaussian. But this is the well known result of Lévy[20] that the distribution of $1/X^2$ is stable with index $\frac{1}{2}$. The same

argument yields the following multivariate form of (3): if $(W_1, \ldots, W_n)$ is uniformly distributed on the surface of the unit sphere in $\mathbb{R}^n$, then for $a_i \geq 0$

$$\sum_{i=1}^{n} \frac{a_i^2}{W_i^2} \stackrel{d}{=} \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{W_1^2}. \tag{33}$$

This was established by induction in [6, Proposition 3.1]. The identity (32) can be recast as

$$\frac{X^2Y^2}{a^2X^2 + c^2Y^2} \stackrel{d}{=} \frac{X^2}{(a+c)^2} \qquad (a, c > 0). \tag{34}$$

This is the identity of first components in the following bivariate identity in distribution, which was derived by M. Mora using the property (7) of the Cauchy distribution: for $p > 0$

$$\left(\frac{(XY(1+p))^2}{X^2 + p^2Y^2}, \frac{(X^2 - p^2Y^2)^2}{X^2 + p^2Y^2}\right) \stackrel{d}{=} \left(X^2, Y^2\right). \tag{35}$$

See Seshadri [35, §2.4, Theorem 2.3] regarding this identity and related properties of the inverse Gaussian distribution of the hitting time of $a > 0$ by a Brownian motion with positive drift. Given (X^2, Y^2), the signs of X and Y are chosen as if by two independent fair coin tosses, so (34) is further equivalent to

$$\frac{XY}{\sqrt{a^2X^2 + c^2Y^2}} \stackrel{d}{=} \frac{X}{a+c} \qquad (a, c > 0). \tag{36}$$

As a variation of (26), set $x = \sqrt{2\lambda}$ and make the change of variable $z = \sqrt{2\lambda u}$ in the integral to deduce the following curious identity: if X is a standard Gaussian then for all $x > 0$

$$E\left(\frac{x}{X\sqrt{X^2 - x^2}} \,\middle|\, X > x\right) \equiv \sqrt{\frac{\pi}{2}} \qquad (x > 0) \tag{37}$$

As a check, (37) for large x is consistent with the elementary fact that the distribution of $(x(X-x) \,|\, X > x)$ approaches that of a standard exponential variable ε_1 as $x \to \infty$. The distribution of $(x/(X\sqrt{X^2 - x^2}) \,|\, X > x)$ therefore approaches that of $1/\sqrt{2\varepsilon_1}$ as $x \to \infty$, and $E(1/\sqrt{2\varepsilon_1}) = \sqrt{\pi/2}$.

By integration with respect to $h(x)dx$, formula (37) is equivalent to the following identity: for all non-negative measurable functions h

$$\sqrt{\frac{2}{\pi}} E\left[\int_0^X \frac{x h(x)\, dx}{X\sqrt{X^2 - x^2}} 1(X \geq 0)\right] = E\left[\int_0^X h(x)\, dx\, 1(X \geq 0)\right].$$

That is to say, for U with uniform $(0,1)$ distribution, assumed independent of X,

$$\sqrt{\frac{1}{2\pi}} E\left[h\left(\sqrt{1-U^2}\,|X|\right)\right] = E\left[\,|X|\, h(\,|X|\, U)\right].$$

Equivalently, for arbitrary non-negative measurable g

$$E\left[g\left((1-U^2)X^2\right)\right] = \sqrt{2\pi} E\left[\,|X|\, h(X^2U^2)\right]. \tag{38}$$

Now $X^2 \stackrel{d}{=} A\varepsilon_2$ where ε_2 is exponential with mean 2, independent of A; and when the density of X^2 is changed by a factor of $\sqrt{2\pi}|X|$ we get back the density of ε_2. So the identity (38) reduces to

$$(1-U^2)A\varepsilon_2 \stackrel{d}{=} U^2\varepsilon_2$$

and hence to

$$(1-U^2)A \stackrel{d}{=} U^2.$$

This is the particular case $a = b = c = 1/2$ of the well known identity

$$\beta_{a+b,c}\,\beta_{a,b} \stackrel{d}{=} \beta_{a,b+c}$$

for $a, b, c > 0$, where $\beta_{p,q}$ denotes a random variable with the beta(p, q) distribution on $(0, 1)$ with density at u proportional to $u^{p-1}(1-u)^{q-1}$, and it is assumed that $\beta_{a+b,c}$ and $\beta_{a,b}$ are independent.

Acknowledgment

Thanks to Steven Lalley and Curtis McMullen for helpful suggestions and references to the ergodic theory of rational maps, and to Gérard Letac for pointing out [6] and [35].

References

[1] Aaronson, J. (1978). Ergodic theory for inner functions of the upper half plane. *Ann. Inst. Henri Poincaré*, **14**, 233–253. MR508928

[2] Aaronson, J. (1997). *An introduction to infinite ergodic theory*. American Mathematical Society, Providence, RI. MR1450400

[3] Adler, R. L. and Rivlin, T. J. (1964). Ergodic and mixing properties of Chebyshev polynomials. *Proc. Amer. Math. Soc.*, **15**, 794–796. MR202968

[4] Beardon, A. F. (1991). *Iteration of Rational Functions*. Springer-Verlag, New York. MR1128089

[5] Brolin, H. (1965). Invariant sets under iteration of rational functions. *Ark. Mat.*, **6**, 103–144. MR194595

[6] Cambanis, S., Keener, R., and Simons, G. (1983). On α-symmetric multivariate distributions. *J. Multivariate Analysis*, **13**, 213–233. MR705548

[7] Carmona, P., Petit, F., and Yor, M. (1994). Some extensions of the arc sine law as partial consequences of the scaling property of Brownian motion. *Probab. Th. Rel. Fields*, **100**, 1–29. MR1292188

[8] Devaney, R. L. (1994). The complex dynamics of quadratic polynomials. In R. L. Devaney, Ed., *Complex dynamical systems*, volume 49 of *Proc. Symp. in Applied Math.*, pp. 1–30. Amer. Math. Soc. MR1315532

[9] Dunau, J.-L. and Senateur, H. (1987). An elementary proof of the Knight-Meyer characterization of the Cauchy distribution. *Journal of Multivariate Analysis*, **22**, 74–78. MR890883

[10] Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Table of Integrals, Series and Products (corrected and enlarged edition)*. Academic Press, New York. MR582453

[11] Hamilton, D. H. (1995). Absolutely continuous conjugacies of Blaschke products. II. *J. London Math. Soc. (2)*, **51**, 279–285. MR1325571

[12] Hamilton, D. H. (1995). Length of Julia curves, *Pacific J. Math.*, **169**, 75–93. MR1346247

[13] Ismail, M. E. H. and Pitman, J. (2000). Algebraic evaluations of some Euler integrals, duplication formulae for Appell's hypergeometric function F_1, and Brownian variations. *Canad. J. Math.*, **52**, 961–981. MR1782335

[14] Kemperman, J. H. B. (1975). The ergodic behavior of a class of real transformations. In *Stochastic Processes and Related Topics I (Proc. Summer Research Inst. on Statistical Inference for Stochastic Processes, Indiana Univ., Bloomington, 1974)*, pp. 249–258. Academic Press, New York. MR372156

[15] Kemperman, J. H. B. and Skibinsky, M. (1982). On the characterization of an interesting property of the arcsin distribution. *Pacific J. Math.*, **103**(2), 457–465. MR705244

[16] Knight, F. B. (1976). A characterization of Cauchy type. *Proc. Amer. Math. Soc.*, **55**, 130–135. MR394803

[17] Knight, F. B. and Meyer, P. A. (1976). Une caractérisation de la loi de Cauchy, *Z. Wahrsch. Verw. Gebiete*, **34**, 129–134. MR397831

[18] Lalley, S. P. (1992). Brownian motion and the equilibrium measure on the Julia set of a rational mapping. *Ann. Probab.*, **20**, 1932–1967. MR1188049

[19] Letac, G. (1977). Which functions preserve Cauchy laws? *Proc. Amer. Math. Soc.*, **67**, 277–286. MR584393

[20] Lévy, P. (1939). Sur certains processus stochastiques homogènes. *Compositio Math.*, **7**, 283–339. MR919

[21] Lévy, P. (1965). *Processus Stochastiques et Mouvement Brownien.* Gauthier-Villars, Paris. (first ed. 1948). MR190953

[22] Lyubich, M. Yu. (1986). The dynamics of rational transforms: the topological picture. *Russian Math. Surveys*, **41**, 43–117. MR863874

[23] Martin, N. F. G. (1983). On finite Blaschke products whose restrictions to the unit circle are exact endomorphisms. *Bull. London Math. Soc.*, **15**(4), 343–348. MR703758

[24] Martin, N. F. G. (1986). Classification of some finite Blaschke products as metric endomorphisms. *J. Math. Anal. Appl.*, **114**(1) 205–209. MR829124

[25] Nakata, T. and Nakamura, M. (1996). On the Julia sets of rational functions of degree two with two real parameters. *Hiroshima Math. J.*, **26**(2) 253–275. MR1400539

[26] Neuwirth, J. H. (1978). Ergodicity of some mappings of the circle and the line. *Israel J. Math.*, **31**, 359–367. MR516157

[27] Petersen, K. (1983). *Ergodic Theory.* Cambridge Univ. Press, Cambridge. MR833286

[28] Pitman, E. J. G. and Williams, E. J. (1967). Cauchy-distributed functions of Cauchy variates. *The Annals of Mathematical Statistics*, **38**, 916–918. MR210166

[29] Pitman, J. (1999). The distribution of local times of Brownian bridge. In *Séminaire de Probabilités XXXIII*, pp. 388–394. Springer, 1999. Lecture Notes in Math. 1709. MR1768012

[30] Pitman, J. and Yor, M. (1986). Level crossings of a Cauchy process. *Annals of Probability*, **14**, 780–792. MR841583

[31] Pitman, J. and Yor, M. (1992). Arcsine laws and interval partitions derived from a stable subordinator. *Proc. London Math. Soc. (3)*, **65**, 326–356. MR1168191

[32] Port, S. C. and Stone, C. J. (1978). *Brownian motion and classical potential theory*. Academic Press, New York.

[33] Revuz, D. and Yor, M. (1994). *Continuous martingales and Brownian motion*. Springer, Berlin-Heidelberg, 2nd edition. MR1303781

[34] Rivlin, T. J. (1990). *Chebyshev polynomials*. Pure and Applied Mathematics. John Wiley & Sons Inc., New York, second edition. MR1060735

[35] Seshadri, V. (1993). *The inverse Gaussian distribution*. The Clarendon Press, Oxford University Press, New York. MR1306281

[36] Spitzer, F. (1958). Some theorems concerning 2-dimensional Brownian motion. *Trans. Amer. Math. Soc.*, **87**, 187–197. MR104296

[37] Stroock, D. W. (1993). *Probability Theory, an Analytic View*. Cambridge Univ. Press. MR1267569

[38] Ulam, S. M. and von Neumann, J. (1947). On combination of stochastic and deterministic processes. *Bull. Amer. Math. Soc.*, **53**, 1120.

[39] Williams, E. J. (1969). Cauchy-distributed functions and a characterization of the Cauchy distribution. *The Annals of Mathematical Statistics*, **40**, 1083–1085. MR243657

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 138–139
© Institute of Mathematical Statistics, 2004

On time changing continuous martingales to Brownian motion

Burgess Davis[1]

Purdue University

Abstract: A short variation of the original proof of Dubins and Schwarz of their result, that all continuous martingales can be time changed to Brownian motion, is given.

It would be hard to overstate my debt to Herman Rubin, whose office is just down the hall. My discussions with him range from twenty minutes on, say, infinitely divisible processes or the Fisk-Stratonovich integral (Fisk was Herman's student.) which completely changed the way I understood subjects I thought I knew, to countless small enlightenments, to providing an instant solution to a homework problem, due that day, that I realized, right before class, was a lot harder than I thought.

This paper provides a short variant of the original proof of the result of Dubins and Schwarz [2] that continuous martingales with unbounded paths can be time changed to standard Brownian motion. See [3] for a discussion of this theorem. We first consider the case that the paths of M are not constant on any open interval, and then discuss the general case. The embedding scheme used here was also used in [2]. The novelty is the use of the lemma below.

Theorem. *Let M_t, $t \geq 0$, be a continuous martingale satisfying $M_0 = 0$, $\sup_t |M_t| = \infty$, and $P(M_s = M_a, a < s < b) = 0$ for all $0 < a < b$. Then there are stopping times η_t, $t \geq 0$, which strictly and continuously increase from 0 to infinity, such that M_{η_t}, $t \geq 0$, is Brownian motion.*

Proof. Let $u_0^M = 0$, and $u_{k+1}^M = inf\{t > u_k : |M_t - M_{u_k^M}| = 1\}$, $k \geq 0$, and let $v_{n,j}^M = u_j^{2^n M}$, if $n, j \geq 0$. We drop the superscript M for the rest of this paragraph. Then M_{u_j}, $j \geq 0$, is a fair random walk, and $M_{v_{n,j}}$, $j \geq 0$, has the distribution of a fair random walk divided by 2^n. Of course the distribution of the $v_{n,j}$ is different for different martingales, but the distribution of the ordering of these times is not. To be precise, the probability of any event in the algebra of events generated by the events $\{v_{i,j} < v_{k,l}\}$ has the same probability for all martingales M. To see this, it helps to first check that $P(v_{1,3} < v_{0,1}) = 1/2$, since the random walk $M_{v_{0,j}}$, $j \geq 0$, is embedded in the random walk $M_{v_{1,j}}$, $j \geq 0$ by the discrete analog of the times $v_{0,k}$, and the probability of the analogous event for these walks is $1/2$. Now since the walks $M_{v_{k,j}}$, $j \geq 0$, can for $0 \leq k < n$ all be embedded in the walk $M_{v_{n,j}}$, $j \geq 0$, which is of course the same walk for any M, the probability of an event in the algebra is the probability of an event for discrete random walk. □

[1]Statistics Department, Purdue University, West Lafayette, IN 47906, USA. e-mail: bdavis@stat.purdue.edu

Keywords and phrases: continuous martingale, Brownian motion.
AMS 2000 subject classifications: primary 60G44, 60J65.

Lemma. *For $0 \le n < \infty$, let $t_{n,j}, 0 \le j < \infty$, be a sequence, and suppose*

(i) $0 = t_{0,n}$, $n \ge 0$,

(ii) $t_{n,j} < t_{n,j+1}$, $j \ge 0$,

(iii) *for all j and n, $t_{n,j}$ is one of the numbers $t_{n+1,k}$, $k \ge 0$,*

(iv) *the set of all the $t_{n,j}$ is dense in $[0, \infty)$.*

Then a sequence a_n, $n \ge 0$, of nonnegative numbers converges if and only if given m there is a j such that $t_{m,j} < a_k < t_{m,j+2}$ for all large enough k. Furthermore if K is a positive integer, an increasing nonnegative function f on $[0, K]$ is continuous if and only if given $n > 0$ there is m such that for each i, $0 \le i < Km$, there is $j = j(i)$ such that $t_{n,j} < f(i/m)$ and $f((i+1)/m) < t_{n,j+2}$.

This lemma is obvious. Now let $v^M_{n,2^{2n}}$ play the role of a_n and $v^M_{n,j}$ have the role of $t_{n,j}$ in this lemma. The conditions (i)–(iv) are easy to check, using the absence of flat spots for iv). The lemma implies that whether or not $v^M_{n,2^{2n}}$ converges (a.s.) depends only on the distribution of the order of the $v^M_{i,j}$. Since this latter distribution does not depend on M, we have either convergence for all M or no M. But if M is a Brownian motion B, we do have convergence, to 1. For, following Skorohod (see [1]), $v^B_{n,2^{2n}}$ has the distribution of the average of 2^{2n} iid random variables each having the distribution of $u^B_1 := u$. Since $Eu = 1$ and since the variance of u is finite, easily shown upon noting that $P(u > k+1 | u > k) < P(|Z| < 2)$, $k > 0$, where Z is standard normal, Chebyshev's inequality gives this convergence to 1. Similarly $\lim_{n\to\infty} v^M_{n,[t2^{2n}]} := \eta^M_t$ exists, where $[\,]$ denotes the greatest integer function. Now the distribution of $M_{\eta^M_1}$ is the limit of the distributions of $M_{v^M_{n,2^{2n}}}$, since M has continuous paths, and thus is the same for all martingales M, and this limit can be identified, by taking $M = B$, as standard normal. All the joint distributions can be similarly treated, and so $M_{\eta^M_t}$ is Brownian motion. This implies that η^M_t is strictly increasing. To see that it is continuous, use the last sentence of the lemma. An argument like that just given shows that continuity on $[0, K]$ for any K, and thus continuity on $[0, \infty)$, either holds for all or no M. And $\eta^B_t = t$. Finally, since η^M_t is continuous and strictly increasing, $\eta^M_t = sup_{k>0} \lim_{n\to\infty} v^M_{n,[(1-\frac{1}{k})2^{2n}t]}$, and so is a stopping time. □

In case the paths of M have flat spots, remove them. Let A stand for the union of the open intervals on which M is constant. Let $h(t) = \inf\{y : |(0,y) \bigcap A^c| = t\}, 0 \le t \le \infty$, where $|\,|$ is Lebesgue measure and the c denotes complement, so that if we define $N_t = M_{h(t)}$, N is continuous with no flat spots. Whether or not N is a martingale, random walks can be embedded in it, since they can be embedded M. Thus just as above, $N_{\eta^N_t}$ is Brownian motion. Put $\mu_t = h(\eta^N_t)$. Then μ is left continuous and strictly increasing, and M_{μ_t} is Brownian motion. And $\mu_t = sup_{k>0} \limsup_{n\to\infty} v^M_{n,[(1-\frac{1}{k})2^{2n}t]}$, so μ_t is a stopping time.

References

[1] Breiman, L. Probability. Addison Wesley, New York, 1968.

[2] Dubins, L., and Schwarz, G. On continuous martingales. PNAS 53 (1965) pp. 913–916. MR178499

[3] Revuz, D., and Yor, M. Continuous martingales and Brownian motion, third edition. Springer, Berlin, 1999.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 140–152
© Institute of Mathematical Statistics, 2004

On counts of Bernoulli strings and connections to rank orders and random permutations

Jayaram Sethuraman[1] **and Sunder Sethuraman**[2]

Florida State University and Iowa State University

Abstract: A sequence of independent random variables $\{X_1, X_2, \ldots\}$ is called a $B-$harmonic Bernoulli sequence if $P(X_i = 1) = 1 - P(X_i = 0) = 1/(i + B)$ $i = 1, 2, \ldots$, with $B \geq 0$. For $k \geq 1$, the count variable Z_k is the number of occurrences of the k-string $(1, \underbrace{0, \ldots, 0}_{k-1}, 1)$ in the Bernoulli sequence... This paper gives the joint distribution P_B of the count vector $\mathbf{Z} = (Z_1, Z_2, \ldots)$ of strings of all lengths in a $B-$harmonic Bernoulli sequence. This distribution can be described as follows. There is random variable V with a Beta$(B, 1)$ distribution, and given $V = v$, the conditional distribution of $\mathbf{Z}$ is that of independent Poissons with intensities $(1 - v)$, $(1 - v^2)/2$, $(1 - v^3)/3, \ldots$.

Around 1996, Persi Diaconis stated and proved that when $B = 0$, the distribution of Z_1 is Poisson with intensity 1. Emery gave an alternative proof a few months later. For the case $B = 0$, it was also recognized that $Z_1, Z_2, \ldots, Z_n$ are independent Poissons with intensities $1, \frac{1}{2}, \ldots, \frac{1}{n}$. Proofs up until this time made use of hard combinational techniques. A few years later, Joffe et al, obtained the marginal distribution of Z_1 as a Beta-Poisson mixture when $B \geq 0$. Their proof recognizes an underlying inhomogeneous Markov chain and uses moment generating functions.

In this note, we give a compact expression for the joint factorial moment of $(Z_1, \ldots, Z_N)$ which leads to the joint distribution given above. One might feel that if Z_1 is large, it will exhaust the number of 1's in the Bernoulli sequence $(X_1, X_2, \ldots)$ and this in turn would favor smaller values for Z_2 and introduce some negative dependence. We show that, on the contrary, the joint distribution of $\mathbf{Z}$ is positively associated or possesses the FKG property.

1. Introduction and summary

Let $\{X_i : i \geq 1\}$ be a sequence of independent Bernoulli random variables with success probabilities $p_i = P(X_i = 1) = 1 - P(X_i = 0)$ for $i \geq 1$. For integers $k \geq 1$, the sequence $(1, \underbrace{0, \ldots, 0}_{k-1}, 1)$ will be called a k-string. Such a k-string represents a wait of length k for an "event" to happen since the last time it happened, or a run of length $k-1$ of "non-events." Let Z_k be the count (which may possibly be infinite) of such k strings in the Bernoulli sequence $\{X_1, X_2, \ldots\}$. This paper is concerned with the joint distribution of the count vector $\mathbf{Z} \stackrel{def}{=} (Z_1, Z_2, \ldots)$ of all k-strings. Such problems appear in many areas such as random permutations, rank orders, genetics, abundance of species, etc.

[1]Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA. e-mail: sethu@stat.fsu.edu

[2]Department of Mathematics, 430 Carvel Hall, Iowa State University, Ames, IA 50011, USA. e-mail: sethuram@iastate.edu

Keywords and phrases: rank order statistics, records, random permutations, factorial moments, cycles, strings, Bernoulli.

AMS 2000 subject classifications: primary, 60C35; secondary 60E05.

Let $Y_{i,k}$ be the indicator variable that a k-string has occurred at time i,

$$Y_{i,k} \;=\; X_i \prod_{j=1}^{i+k-1} (1 - X_{i+j}) X_{i+k} \;=\; I\Big((X_i, X_{i+1}, \ldots, X_{i+k}) = (1, \underbrace{0, \ldots, 0}_{k-1}, 1) \Big), \tag{1}$$

for $i \geq 1, k \geq 1$, where as usual, an empty product is defined to be equal to 1. A simple expression for Z_k is then given by

$$Z_k \;=\; \sum_{i=1}^{\infty} Y_{i,k} \quad \text{for } k \geq 1. \tag{2}$$

While Z_k is not a sum of independent summands, it can be easily expressed as the sum of k series, each of which has independent summands. From this observation and Kolomogorov's three-series theorem we can state the following remark which gives a necessary and sufficient condition that the random variable Z_k be finite a.s.

Remark 1. The count random variable Z_k of k-strings is finite a.s. if and only if $E[Z_k] = \sum_{i \geq q} p_i \prod_{j=1}^{k-1} (1 - p_{i+j}) p_{i+k} < \infty$.

In this paper, we will concentrate exclusively on independent Bernoulli sequences, with a particular type of "harmonic" sequence for $\{p_i\}$, which allows for explicit computations and also, in some cases, connects the count vector $(Z_1, Z_2, \ldots)$ with the study of rank order statistics and random permutations. In fact, we will assume that $\{p_i\}$ satisfies

$$p_i(1 - p_{i+1}) = p_{i+1} \;\text{ or equivalently }\; p_i - p_{i+1} = p_i p_{i+1} \;\text{ for } n \geq 1. \tag{3}$$

We will avoid the case $p_1 = 0$, since then the only solution to (3) is the the trivial solution $p_i \equiv 0$. We will therefore assume, for the rest of this paper, that $p_1 = 1/(1 + B)$, with $B \geq 0$, so that from (3)

$$p_i = \frac{1}{i + B} \quad \text{for } i \geq 1. \tag{4}$$

We will refer to an independent Bernoulli sequence with $\{p_i\}$ given in (4) as a *$B-$harmonic Bernoulli sequence.* Occasionally, when we wish to emphasize the dependence on B, we will write $Z_{k,B}$ for the count variable Z_k, and $\mathbf{Z}_B$ for the count vector $\mathbf{Z}$. From Remark 1,

$$\begin{aligned} E[Z_{k,B}] \;&\leq\; \sum_{i \geq 1} p_i p_{i+k} \\ &=\; \sum_{i \geq 1} \frac{1}{(i + B)(i + k + B)} < \infty, \end{aligned}$$

and thus $Z_{k,B}$ is finite, for all $k \geq 1$, a.s.

When the counts $(Z_1, Z_2, \ldots)$ are almost-surely finite, their joint distribution becomes an object of interest, especially its dependence on the sequence of probabilities $\{p_i\}$. Around 1996, Persi Diaconis observed that, for $0-$harmonic Bernoulli sequences, the distribution of the count variable Z_1 is Poisson with intensity 1. A few months later [Emery (1996)] gave another proof in an unpublished manuscript. It is known that the count vector $(Z_1, \ldots, Z_k)$ of a $0-$harmonic Bernoulli

sequence can be thought of the limit of the vector $(C_1(n), \ldots, C_k(n))$ of numbers of cycles of different orders among permutations of $\{1, 2, \ldots, n\}$. (More details are given in the next section.) This fact coupled with the classical results (see [Arratia et al. (2003)], [Arratia (1992)]) establish that the joint distribution of the count vector $(Z_1, Z_2, \ldots, Z_k)$, from a $0-$harmonic Bernoulli sequence, is that of independent Poissons with intensities $(1, \frac{1}{2}, \ldots, \frac{1}{k})$, respectively. All these proofs mentioned are based on combinatorial methods.

[Joffe et al. (2002)] considered general $B-$harmonic Bernoulli sequences and obtained the moment generating function of of $Z_{1,B}$ by noticing that $\{(S_i, X_{i+1}),\ i = 1, 2, \ldots\}$ forms an inhomogeneous Markov chain, where $S_i = \sum_{m=1}^{i} X_m X_{m+1}$. From this they identified the distribution of Z_1 as a Generalized Hypergeometric Factorial (GHF) law which is more easily stated as a Beta-mixture of Poisson distributions.

In this paper we consider general $B-$harmonic Bernoulli sequences and obtain the joint distribution P_B of the count vector $\mathbf{Z}_B = (Z_{1,B}, Z_{2,B}, \ldots)$. With the addition of another random variable V, the joint distribution Q_B of $(V, \mathbf{Z}_B)$ can be described as follows: the distribution of V is Beta with parameters $(B, 1)$ and the conditional distribution $P_{B,v}$ of $\mathbf{Z}_B$ given $V = v$, is that of independent Poissons with intensities $(1 - v),\ (1 - v^2)/2,\ (1 - v^3)/3, \ldots$. These results are contained in Theorem 2.

We also compute the covariance of $Z_{k,B}$ and $Z_{m,B}$ for $k \leq m$ and note that it is positive for $B > 0$ in Corollary 2. We also show that P_B has the FKG or the positive association property in Theorem 3. There are intuitions for both positive and negative correlations between $Z_{k,B}$ and $Z_{m,B}$ and so this result is perhaps of interest. A plausible justification for positive correlations arises from the feeling that more completed k-strings allow one to "start over" more times in the Bernoulli sequence and so can lead to more strings of length m. Although with the interpretation of $Z_{k,B}$ as the number of cycles of length k among random permutations of $E_{n,B} = \{1, 2, \ldots, n + B\}$ when $B \geq 0$ is an integer (see the next section), the "age-dependent"-cycle count mapping gives perhaps the opposite interpretation. Namely, with more k-cycles formed, there should be less "room" for m-cycles to form in $E_{n,B}$, leading to negative association between $Z_{k,B}$ and $Z_{m,B}$. One may think, however, for fixed $k < m$ much smaller than $n \uparrow \infty$, that such "boundary" considerations are negligible and the first explanation is more reasonable given that the mixture distribution is of Beta type which has interpretations with respect to "reinforcement" dynamics (e.g. Polya urns). On the other hand, since the asymptotic joint distribution depends on B, we know that the "boundary" is not completely ignored in the limit, thereby confusing the matter once more. It would be of interest to have a better understanding of these dependence issues.

Our methods avoid the use of combinatorial techniques. We first show, in Lemma 2, that factorial powers of count variables $Z_{k,B}$, which are sums of indicator variables $Y_{i,k}$ (see (2)) can be expressed as simple sums of products of the $Y_{i,k}$'s. For $B-$harmonic Bernoulli sequences, many products of the form $Y_{i,k}Y_{j,k}$ vanish and there are some independence properties among the $Y_{i,k}$'s; see (6), (7) and (8). These are exploited in Lemma 1, Lemma 2 and Lemma 3 to obtain the joint factorial moments of $(Z_{1,B}, \ldots, Z_{n,B})$ in the main theorem (Theorem 1) which is further simplified in Theorem 2 by recognizing it as the sum of probabilities of inequalities among independent exponential variables. The joint distribution of $(Z_{1,B}, \ldots, Z_{n,B})$ can be deduced from this simplified expression for the factorial moments.

Even though the frequency of wait times between 1's of all orders are finite a.s., it is interesting to note that there are infinitely many 1's in the original Bernoulli

sequence (since $\sum_{i\geq 1} p_i = \sum_{i\geq 1} 1/(i+B) = \infty$). However, the events (i.e. 1's) are so sparse that the wait to the first event has infinite mean when $B > 0$. Let $N = \inf\{i \geq 1 : X_i = 1\}$ be the wait to the first event. Then $P(N = k) = B/[(k-1+B)(k+B)]$ when $B > 0$, and though $P(N < \infty) = 1$ we have $E[N] = \infty$. In a similar fashion, when $B = 0$, $X_1 = 1$ a.s. and the wait for the second event has infinite expectation. It is also not difficult to see that, no matter the value of $B \geq 0$, the number of 1's, $N_n = \sum_{i=1}^{n} X_i$, satisfies $N_n/\log n \to 1$ a.s., and $(N_n - \log n)/\sqrt{\log n} \xrightarrow{d} N(0,1)$ (cf. Example 4.6, Ch. 2 [Durrett (1995)]).

Finally, a statistician may ask whether the value of B can be consistently estimated from the count vector Z of all k-strings. To say that this can be done is the same as saying that P_B and $P_{B'}$ are mutually singular for $B \neq B'$. Let M_B be the joint distribution of a $B-$harmonic Bernoulli sequence $\{X_i, i = 1, 2, \ldots\}$. We show in Theorem 4, by use of Kakutani's criterion, that M_B and $M_{B'}$ are absolutely continuous with respect to each other for $B \neq B'$. This implies the same for P_B and $P_{B'}$, and thus B cannot be consistently estimated from $\mathbf{Z}$.

2. Related areas

Count vectors of k-strings as described above, apart from being objects of intrinsic research interest, have concrete interpretations with respect to combinatorics, genetics, ecology, statistics, and other areas (cf. [Arratia et al. (2003)], [Johnson et al. (1992)], and [Antzoulakos and Chadjiconstantinidis (2001)] and references therein). We will describe some connections to rank orders, record values and permutations for the case when $B \geq 0$ is an integer. In both situations, there is an embedded sequence of independent Bernoulli r.v.'s with respect to which the counts of k-strings have various interpretations.

Rank orders and record values. Let $\{\xi_n : n \geq 1\}$ be a sequence of i.i.d. r.v.'s with common continuous distribution function F. One might think of ξ_n as the amount of rainfall or the flood level in the nth year. Let $\xi_{1,n} < \xi_{2,n} < \cdots < \xi_{nn}$ be the ordered values of $\{\xi_i : 1 \leq i \leq n\}$ and define $R_n = j$ if $\xi_n = \xi_{j,n}$. It is a well known theorem of Renyi that $\{R_n : n \geq 1\}$ are independent and uniformly distributed on their respected ranges (cf. Example 6.2, Ch. 1 [Durrett (1995)]). Let $\{a_1, a_2, \ldots\}$ be a sequence of integers such that $1 \leq a_n \leq n$ and define $X_n = I(R_n = a_n)$. The sequence $\{X_n,\ n \geq 1\}$ is an example of a $0-$harmonic Bernoulli sequence, for any choice of the sequence $\{a_1, a_2, \ldots\}$. The sequence $\{X_{n,B} = X_{n+B}, n \geq 1\},\ n \geq 1\}$ is an example of a $B-$harmonic Bernoulli sequence when $B \geq 0$ is an integer.

In the special case $a_n = n$ for $n \geq 1$, the event $X_{n,B} = 1$ means that a record, with respect to the rainfall amounts in the first B years (which were lost or not properly recorded), was set during the year $n + B$. In this case, $Z_{k,B}$ is the number of times records were set after a wait of $k - 1$ years from a previous record.

Of course, by choosing $\{a_n\}$ differently, one can vary the interpretation of $Z_{n,B}$.

Random permutations. For $B \geq 0$ an integer, let $E_{n,B} = \{1, 2, \ldots, n + B\}$. We now describe the "Feller" algorithm which chooses a permutation $\pi : E_{n,B} \to E_{n,B}$ uniformly from the $(n+B)!$ possible permutations (cf. Section 4 [Joffe et al. (2002)], Chapter 1 of [Arratia et al. (2003)]).

1. Draw the first element uniformly from $E_{n,B}$ and call it $\pi(1)$. If $\pi(1) = 1$, a cycle of length 1 has been completed. If $\pi(1) = j \neq 1$, make a second draw uniformly from $E_{n,B} \setminus \{\pi(1)\}$ and call it $\pi(\pi(1)) = \pi(j)$. Continue drawing elements naming

them $\pi(\pi(j))), \pi((\pi(\pi(j)))), \ldots$ from the remaining numbers until 1 is drawn, at which point a cycle (of some length) is completed.

2. From the elements left after the first cycle is completed, $E_{n,B} \setminus \{\pi(1), \ldots, 1\}$, follow the process in step 1 with the smallest remaining number taking the role of "1." Repeat until all elements of $E_{n,B}$ are exhausted.

When $B = 0$, n such Feller draws produces a random permutation, $\pi : E_{n,0} \to E_{n,0}$. However, when $B > 0$, in n such Feller draws, $\pi : E_{n,B} \to E_{n,B}$ is only injective, and there may be the possibility that no cycle of any length is completed.

Let now $\{I_i^{(n)} : 1 \leq i \leq n\}$ be the indicators of when a cycle is completed at the ith drawing in n Feller draws from $E_{n,B}$. It is not difficult to see that $\{I_i^{(n)}\}$ are independent Bernoulli random variables with $P(I_i^{(n)} = 1) = 1/(n + B - i + 1)$, since at time i, independent of the past, there is exactly one choice among the remaining $n + B - i + 1$ members left in $E_{n,B}$ to complete the cycle (to paraphrase Example 5.4, Ch. 1 [Durrett (1995)]).

For $1 \leq k \leq n$, let $D_{k,B}^{(n)}$ be the number of cycles of length k in the first n Feller draws from $E_{n,B}$. It is easy to see that

$$D_{k,B}^{(n)} \xrightarrow{p} Z_{k,B} \quad \text{for } k \geq 1$$

and we give a quick proof below.

Indeed, since a cycle of length k is finished on the mth draw, for $m \geq k + 1$, exactly when $I_{m-k}^{(n)}(1 - I_{m-k+1}^{(n)}) \cdots (1 - I_{m-1}^{(n)}) I_m = 1$, and also since the first cycle is a k-cycle exactly when $(1 - I_1^{(n)})(1 - I_2^{(n)}) \cdots (1 - I_{k-1}^{(n)}) I_k^{(n)} = 1$, we have

$$D_{k,B}^{(n)} = (1 - I_1^{(n)})(1 - I_2^{(n)}) \cdots (1 - I_{k-1}^{(n)}) I_k^{(n)} + \sum_{i=1}^{n-k} I_i^{(n)} (1 - I_{i+1}^{(n)}) \cdots (1 - I_{i+k-1}^{(n)}) I_{i+k}^{(n)}.$$

Let $\{X_i : i \geq 1\}$ be independent Bernoulli random variables defined on a common space with $P(X_i = 1) = 1/(i + B)$, so that $X_i = I_{n-i+1}^{(n)}$ in law for $1 \leq i \leq n$. We can then write $D_{k,B}^{(n)}$ equivalently in distribution as

$$D_{k,B}^{(n)} \stackrel{d}{=} \sum_{i=1}^{n-k} X_i (1 - X_{i+1}) \cdots (1 - X_{i+k-1}) X_{i+k} \; + \; X_{n-k+1} \prod_{j=n-k+2}^{n} (1 - X_j).$$

As $\lim_{n \to \infty} X_{n-k+1}(1 - X_{n-k+2}) \cdots (1 - X_n) = 0$ in probability, we have

$$D_{k,B}^{(n)} \xrightarrow{p} \sum_{i \geq 1} X_i (1 - X_{i+1}) \cdots (1 - X_{i+k-1}) X_{i+k} \; = Z_{k,B}. \tag{5}$$

We see from this construction, that $Z_{k,B}$ represents the asymptotic number of "young" or "age-dependent" k-cycle numbers, that is, those formed in the first n Feller draws from sets of size $n + B$.

3. Preliminary lemmas

We will use the following standard definition of the factorial power of order r of an integer a:

$$a^{[r]} = \begin{cases} a(a-1) \cdots (a-r+1) & \text{when } a, r \geq 1 \\ 1 & \text{when } r = 0 \\ 0 & \text{when } a = 0. \end{cases}$$

Equation (2) gives a representation for the count variable Z_k of k-strings as a series of dependent summands $Y_{i,k}$, defined in (1) in terms of the $B-$harmonic Bernoulli sequence $\{X_i, i \geq 1\}$. The summands $\{Y_{i,k}, i \geq 1\}$ are indicator variables with the following useful properties

$$Y_{i,k}^2 = Y_{i,k}, \quad Y_{i,k}Y_{i,k'} = 0 \text{ if } k \neq k', \quad Y_{i,k}Y_{i',k'} = 0 \text{ for } i+1 \leq i' < i+k, \tag{6}$$

$$Y_{i,k} \text{ and } Y_{i+k+j,m} \text{ are independent for } j \geq 1, \tag{7}$$

$$\left.\begin{array}{l} E(Y_{i,k}) = \frac{1}{(i+k-1+B)(i+k+B)}, \text{ and} \\ E(Y_{i,k}Y_{i+k,m}) = \frac{1}{(i+k-1+B)(i+k+m-1+B)(i+k+m+B)}. \end{array}\right\} \tag{8}$$

These properties allow us to give simplified expressions for products of factorial powers of the count vector $(Z_1, \ldots, Z_n)$ in terms of $\{Y_{i,k}\}$.

The following lemma gives a representation for the factorial power of a sum of arbitrary indicator variables.

Lemma 1. *Let $(I_1, I_2, \ldots)$ be indicator variables, and let $Z = \sum_{i\geq 1} I_i$ be their sum. Then for integers $r \geq 1$, the factorial powers of Z have the following representation:*

$$Z^{[r]} = \sum_{\substack{i_1,\ldots,i_r \\ \text{distinct}}} I_{i_1} I_{i_2} \cdots I_{i_r} = r! \sum_{1 \leq i_1 < \ldots < i_r} I_{i_1} I_{i_2} \cdots I_{i_r}. \tag{9}$$

Proof. The proof is by induction. For $r = 1$, the identity in (9) is obvious. Now assume that the same identity holds for $r-1$, with $r \geq 2$. Write

$$\begin{aligned} Z^{[r]} &= (Z-(r-1)) \cdot Z^{[r-1]} \\ &= (Z-(r-1)) \cdot \sum_{\substack{i_1,\ldots,i_{r-1} \\ \text{distinct}}} I_{i_1} \cdots I_{i_{r-1}}. \end{aligned}$$

Since I_j is $0-1$ valued, $I_j^2 = I_j$ for all j, and we have

$$\begin{aligned} Z \sum_{\substack{i_1,\ldots,i_{r-1} \\ \text{distinct}}} I_{i_1} \cdots I_{i_{r-1}} &= \left[\sum_{i_r} I_{i_r}\right]\left[\sum_{\substack{i_1,\ldots,i_{r-1} \\ \text{distinct}}} I_{i_1} \cdots I_{i_{r-1}}\right] \\ &= (r-1) \sum_{\substack{i_1,\ldots,i_{r-1} \\ \text{distinct}}} I_{i_1} \cdots I_{i_{r-1}} + \sum_{\substack{i_1,\ldots,i_r \\ \text{distinct}}} I_{i_1} \cdots I_{i_{r-1}} I_{i_r}. \end{aligned}$$

Thus

$$Z^{[r]} = \sum_{\substack{i_1,\ldots,i_r \\ \text{distinct}}} I_{i_1} \cdots I_{i_r}.$$

This establishes the identity in (1) for r and completes the proof of Lemma 1. □

Lemma 1 can be used to obtain expressions of products of factorial powers of count vectors in a routine way. Lemma 2 will improve on this and give an alternative expression for such a product, by exploiting property (6) of $\{Y_{i,k}\}$. To state this result we will need the following notation.

Let $k_1, k_2, \ldots, k_n$ be distinct integers and let $r_1, r_2, \ldots, r_n$ be (not necessarily distinct) integers all of which are greater than or equal to 1. Let $R_0 = 0, R_m = \sum_1^n r_j, m = 1, \cdots, n$ and let $A_n = \{\lambda_l\}_{l=1}^{R_n} = \{\underbrace{k_1, \ldots, k_1}_{r_1}, \underbrace{k_2, \ldots, k_2}_{r_2}, \ldots, \underbrace{k_n, \ldots, k_n}_{r_n}\}$.

Let $\mathcal{S}_{A_n}$ be the $R_n!$ permutations of A_n, though there are only $\binom{R_n}{r_1,r_2,\cdots,r_n}$ distinct permutations. Finally, for $\pi \in \mathcal{S}_{A_n}$, let

$$S_m(\pi) \;=\; \sum_{j=1}^{m} \pi_j \text{ for } 1 \le m \le R_n. \tag{10}$$

Lemma 2. *For $n \ge 1$, let $k_1, \ldots, k_n \ge 1$ be distinct integers and $r_1, \ldots, r_n \ge 1$ be (not necessarily distinct) integers. Then,*

$$Z_{k_1}^{[r_1]} \cdots Z_{k_n}^{[r_n]} \;=\; \sum_{\pi \in \mathcal{S}_{A_n}} \sum_{1 \le i_1 < \cdots < i_{R_n}} Y_{i_1,\pi_1} Y_{i_2,\pi_2} \cdots Y_{i_{R_n},\pi_{R_n}}. \tag{11}$$

Proof. From Lemma 1 and (6), we get

$$\begin{aligned} Z_{k_1}^{[r_1]} \cdots Z_{k_n}^{[r_n]} &= \prod_{j=1}^{n} \sum_{\substack{i_{R_{j-1}+1}, \cdots, i_{R_j} \\ \text{distinct}}} Y_{i_{R_{j-1}}+1,k_j} \cdots Y_{i_{R_j},k_j} \\ &= \sum_{\substack{i_1, \cdots, i_{R_n} \\ \text{distinct}}} Y_{i_1,k_1} \cdots Y_{i_{R_1},k_1} \cdots \; \cdots Y_{i_{R_{n-1}}+1,k_n} \cdots Y_{i_{R_n},k_n} \\ &= \sum_{\pi \in \mathcal{S}_{A_n}} \sum_{1 \le i_1 < \cdots < i_{R_n}} Y_{i_1,\pi_1} Y_{i_2,\pi_2} \cdots Y_{i_{R_n},\pi_{R_n}}. \end{aligned}$$

This completes the proof of Lemma 2. □

For a vector of integers $\mathbf{k} = (k_1, k_2, \ldots)$ with $k_n \ge 1$ for all n, define $K_m = \sum_{j=1}^{m} k_j$ to be the partial sums, $k(r,s) = (k_r, k_{r+1}, \ldots, k_s)$ to be the segment from r to s. For $1 \le m \le n$ and $r \ge 1$, define

$$C(r : k(m,n)) \;=\; \sum_{r \le i_m < i_{m+1} < \cdots < i_n} Y_{i_m,k_m} Y_{i_{m+1},k_{m+1}} \cdots Y_{i_n,k_n}.$$

The following is a key lemma which gives two identities useful for the calculation of factorial moments of the count vector $(Z_{1,B}, \ldots, Z_{k,B})$.

Lemma 3. *For integers $r, n \ge 1$ and vectors $\mathbf{k}$ the following two identities hold:*

$$E[Y_{r,k_1} C(r+1; k(2,n+1))] \;=\; \prod_{m=1}^{n+1} \frac{1}{r-1+K_m+B} - \prod_{m=1}^{n+1} \frac{1}{r+K_m+B}, \tag{12}$$

and

$$E[C(r; k(1,n))] \;=\; \prod_{m=1}^{n} \frac{1}{r-1+K_m+B}. \tag{13}$$

Proof. The proof is by simultaneous induction for both (12) and (13) on n, the number of $Y_{i,k}$ factors in $C(r : k(l,m))$ where $m - l + 1 = n$. Throughout, we will rely heavily on the properties (6),(7) and (8) of $\{Y_{i,k}\}$.

We will now establish (12) for $n = 1$. Notice that

$$\begin{aligned} &E\big[Y_{r,k_1} C(r+1; k(2,2))\big] \\ &\quad = \sum_{i \ge r+1} E[Y_{r,k_1} Y_{i,k_2}] = \sum_{i \ge r+k_1} E[Y_{r,k_1} Y_{i,k_2}] \end{aligned}$$

$$
\begin{aligned}
&= E[Y_{r,k_1}Y_{r+k_1,k_2}] + \sum_{i\geq r+k_1+1} E[Y_{r,k_1}]E[Y_{i,k_2}] \\
&= \frac{1}{(r+k_1-1+B)(r+K_2-1+B)(r+K_2+B)} \\
&\quad + \frac{1}{(r+k_1-1+B)(r+k_1+B)} \sum_{i\geq r+k_1+1} \frac{1}{(i+k_2-1+B)(i+k_2+B)} \\
&= \frac{1}{(r+k_1-1+B)(r+K_2-1+B)(r+K_2+B)} \\
&\quad + \frac{1}{(r+k_1-1+B)(r+k_1+B)(r+K_2+B)} \\
&= \frac{1}{(r-1+k_1+B)(r-1+K_2+b)} - \frac{1}{(r+k_1+B)(r+K_2+B)}.
\end{aligned}
$$

This establishes (12) for $n = 1$.

Next,

$$
\begin{aligned}
E\big[C(r;k(1,1))\big] \;=\; \sum_{i_1\geq r} E[Y_{i,k_1}] \;&=\; \sum_{i\geq r} \frac{1}{(i+k_1-1+B)(i+k_1+B)} \\
&=\; \sum_{i\geq r} \left[\frac{1}{(i+k_1-1+B)} - \frac{1}{(i+k_1+B)}\right] \\
&=\; \frac{1}{r-1+k_1+B}
\end{aligned}
$$

which establishes (13) for $n = 1$.

For the induction step, let $N \geq 2$ and assume that (12) and (13) hold for $n = N-1$. We first establish (13) for $n = N$ by using the validity of (12) for $n = N-1$ as follows:

$$
\begin{aligned}
E\big[C(r;k(1,N))\big] &= E\left[\sum_{r\leq i_1<\cdots<i_N} Y_{i_1,k_1}\cdots Y_{i_N,k_N}\right] \\
&= E\left[\sum_{r\leq i} Y_{i,k_1}\left[\sum_{i+1\leq i_2<\cdots<i_N} Y_{i_2,k_2}\cdots Y_{i_N,k_N}\right]\right] \\
&= \sum_{r\leq i}\left[\prod_{m=1}^{N}\frac{1}{i+K_m-1+B} - \prod_{m=1}^{N+1}\frac{1}{i+K_m+B}\right] \\
&= \prod_{m=1}^{N}\frac{1}{r+K_m-1+B}.
\end{aligned}
$$

To finish the induction we now proceed to establish (12) for $n = N$, assuming that (12) holds for $n = N-1$ and (13) holds for $n = N$. Notice that

$$
\begin{aligned}
E[Y_{r,k_1}\mathcal{C}(r+1;k(2,N+1))] &= E[Y_{r,k_1}\mathcal{C}(r+k_1;k(2,N+1))] \\
&= E[Y_{r,k_1}Y_{r+k_1,k_2}\mathcal{C}(r+K_2;k(3,N+1))] \\
&\quad + E[Y_{r,k_1}]E[\mathcal{C}(r+k_1+1;k(2,N+1))].
\end{aligned}
$$

By conditioning on X_{r+k_1} and noting that many terms vanish when $X_{r+k_1} = 0$, the first term above simplifies as follows:

$$
\begin{aligned}
&E[Y_{r,k_1} Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))] \\
&\quad = E[Y_{r,k_1} E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))|X_r, \ldots, X_{r+k_1}]] \\
&\quad = E[Y_{r,k_1} E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))|X_{r+k_1}]] \\
&\quad = E[E[Y_{r,k_1}|X_{r+k_1}] E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))|X_{r+k_1}]] \\
&\quad = E[Y_{r,k_1}|X_{r+k_1}=1] \\
&\qquad \cdot E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))|X_{r+k_1}=1] P(X_{r+k_1}=1) \\
&\quad = E[Y_{r,k_1}|X_{r+k_1}=1] E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))].
\end{aligned}
$$

The assumption that (12) and (13) hold for $n = N-1$ yields

$$
\begin{aligned}
&E[Y_{r,k_1} \mathcal{C}(r+1; k(2,N+1))] \\
&\quad = E[Y_{r,k_1}|X_{r+k_1}=1] E[Y_{r+k_1,k_2} \mathcal{C}(r+K_2; k(3,N+1))] \\
&\qquad + E[Y_{r,k_1}] E[\mathcal{C}(r+k_1+1; k(2,N+1))] \\
&\quad = \frac{1}{r+k_1-1+B} \Big[\prod_{m=2}^{N+1} \frac{1}{r+K_m-1+B} - \prod_{m=2}^{N+1} \frac{1}{r+K_m+B} \Big] \\
&\qquad + \frac{1}{(r+k_1-1+B)(r+k_1+B)} \prod_{m=2}^{N+1} \frac{1}{r+K_m+B} \\
&\quad = \frac{1}{r+k_1-1+B} \Big[\prod_{m=2}^{N+1} \frac{1}{r+K_m-1+B} - \prod_{m=2}^{N+1} \frac{1}{r+K_m+B} \Big] \\
&\qquad + \frac{1}{r+k_1-1+B} \prod_{m=2}^{N+1} \frac{1}{r+K_m+B} - \prod_{m=1}^{N+1} \frac{1}{r+K_m+B} \\
&\quad = \prod_{m=1}^{N+1} \frac{1}{r+K_m-1+B} - \prod_{m=1}^{N+1} \frac{1}{r+K_m+B}.
\end{aligned}
$$

This establishes (12) for $n = N$ and completes the proof of the lemma. □

4. Main results and corollaries

Consider a $B-$harmonic Bernoulli sequence and the corresponding count vector $\mathbf{Z}_B$. For non-negative integers $s_1, s_2, \ldots, s_n$, define

$$\mu_B(s_1, \ldots, s_n) \;=\; E(Z_{1,B}^{[s_1]} Z_{2,B}^{[s_2]} \cdots Z_{n,B}^{[s_n]}).$$

The following theorem gives an explicit form for the factorial moments of this count vector which will be used to identify its joint distribution.

Theorem 1. *Let $\mathbf{Z}_B$ be the count vector arising from a $B-$harmonic Bernoulli sequence $\{X_i\}$. Let $k_1, \ldots, k_n$ be distinct integers and let $r_1, \ldots, r_n$ be not necessarily distinct integers, all greater than or equal to 1. Recall the notations $R_m, A_n, \mathcal{S}_{A_n}$ and $S_m(\pi)$ from just before (10). Then*

$$E\big[Z_{k_1,B}^{[r_1]} Z_{k_2,B}^{[r_2]} \cdots Z_{k_n,B}^{[r_n]}\big] \;=\; \sum_{\pi \in \mathcal{S}_{A_n}} \prod_{m=1}^{R_n} \frac{1}{S_m(\pi)+B} \tag{14}$$

Proof. From Lemmas 2 and 3, using the notation in (10),

$$\begin{aligned} E\big[Z_{k_1,B}^{[r_1]} Z_{k_2,B}^{[r_2]} \cdots Z_{k_n,B}^{[r_n]}\big] &= E\Big[\sum_{\pi\in\mathcal{S}_{A_n}} \sum_{1\le i_1<\cdots<i_{R_n}} Y_{i_1,\pi_1} Y_{i_2,\pi_2} \cdots Y_{i_{R_n},\pi_{R_n}}\Big] \\ &= E\Big[\sum_{\pi\in\mathcal{S}_{A_n}} C(1;\pi(1,R_n))\Big] \\ &= \sum_{\pi\in\mathcal{S}_{A_n}} \prod_{m=1}^{R_n} \frac{1}{S_m(\pi)+B}. \end{aligned}$$

This completes the proof of the theorem. □

The next theorem, which is the main result of this paper, gives the factorial moments of $(Z_{1,B},\cdots,Z_{N,B})$ for $B-$harmonic Bernoulli sequences and deduces the structure of the joint distribution of $\mathbf{Z}_B$.

Theorem 2. *For non-negative integers $s_1,\ldots,s_N$,*

$$\mu_B(s_1,\ldots,s_N) \;=\; \int_0^1 B v^{B-1} \prod_{j=1}^{N} \left(\frac{(1-v^j)}{j}\right)^{s_j}. \tag{15}$$

This implies that the joint distribution P_B of $\mathbf{Z}_B$ has the following structure: there is random variable V and the joint distribution Q_B of $(V,\mathbf{Z}_B)$ can be described as follows: V has a Beta$(B,1)$ distribution (which is the point mass at 0 when $B=0$) and given $V=v$, the conditional distribution $P_{B,v}$ of $(Z_{1,B},Z_{2,B},\ldots)$ is that of independent Poissons with intensities $1-v, \frac{1-v^2}{2},\ldots$ respectively.

Proof. First, let $B>0$ as the case $B=0$ is analogous or can be obtained by taking the limit $B\downarrow 0$. Second, to establish (15), we can assume that some $s_m>0$ for some m. In fact, let $(s_{k_1},\ldots,s_{k_n})$ be the vector formed from the non-zeros in $(s_1,s_2,\ldots,s_N)$, and let R_n, A_n, S_{A_n} and $S_m(\pi)$ for $\pi\in S_{A_n}$ be as defined near (10). Let also $W_0,W_1,W_2,\ldots,W_{R_n}$ be independent exponential r.v.'s with failure rates $B,\lambda_1,\ldots,\lambda_{R_n} \stackrel{def}{=} B,\underbrace{k_1,\ldots,k_1}_{r_1},\ldots,\underbrace{k_n,\ldots,k_n}_{r_n}$, respectively. Then, for any $\pi\in\mathcal{S}_{A_n}$

$$\prod_{m=1}^{R_n} \frac{\pi_m}{S_m(\pi)+B} \;=\; \prod_{m=1}^{R_n} \frac{\lambda_{\pi_m}}{S_m(\pi)+B} \;=\; P(W_{\pi_{R_n}} < W_{\pi_{R_n-1}} < \cdots < W_{\pi_1} < W_0). \tag{16}$$

From Theorem 1 and (16), we conclude

$$\begin{aligned} \Big(\prod_{j=1}^{N} j^{s_j}\Big)\cdot \mu_B(s_1,\ldots,s_N) &= \Big(\prod_{j=1}^{n} (k_j)^{s_{k_j}}\Big)\cdot E_B(Z_{k_1,B}^{[s_{k_1}]}\cdots Z_{k_n,B}^{[s_{k_n}]}) \\ &= \sum_{\pi\in S_{A_n}} \prod_{m=1}^{R_n} \frac{\pi_m}{S_m(\pi)+B} \\ &= \sum_{\pi\in S_{A_n}} P(W_{\pi_{R_n}} < \cdots < W_{\pi_1} < W_0) \\ &= P(\max(W_1,\ldots,W_{R_n}) < W_0) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty Be^{-By} \prod_{j=1}^{n} (1-e^{-k_j y})^{s_{k_j}} \, dy \\
&= \int_0^1 Bv^{B-1} \prod_{j=1}^{N} (1-v^j)^{s_j} \, dv \\
&= \left(\prod_{j=1}^{N} j^{s_j} \right) \cdot \int_0^1 Bv^{B-1} \prod_{j=1}^{N} E(Z_{j,v}^{[s_j]}) dv
\end{aligned}
$$

where, for each v, $Z_{1,v}, Z_{2,v}, \ldots$ are independent Poisson random variables with means $(1-v), (1-v^2)/2, \ldots$, respectively. This establishes the structure of P_B as desired. □

Remark 2. We now indicate an alternate argument to obtain Theorem 2. Consider the factorial moment generating function

$$\phi_B(t_1, \ldots, t_n) \stackrel{def}{=} \sum_{r_1, \ldots, r_n \geq 0} \mu_B(r_1, \ldots, r_n) \frac{{t_1}^{r_1} \ldots {t_n}^{r_n}}{r_1! \ldots r_n!}.$$

The denominator of the last factor in (14), $S_{R_n}(\pi) + B$, is the same for all values of π and equals $\sum_1^n r_j k_j + B$. Hence, we have the recurrence relation

$$\mu_B(r_1, \ldots, r_n) = \sum_1^n r_j \mu_B(r_1, \ldots, r_j - 1, \ldots, r_n)$$

which in turn leads to the partial differential equation

$$\sum_{j=1}^{n} j t_j \frac{\partial \phi_B}{\partial t_j} = (\sum_1^n t_j - B)\phi_B + B. \tag{17}$$

Also, the marginal factorial moment generating function $\phi_{j,B}(t_j)$ of $Z_{j,B}$ satisfies $jt_j \partial \phi_{j,B}(t_j)/\partial t_j = (t_j - B)\phi_{j,B}(t_j) + B$ with the boundary condition $\phi_{j,B}(0) = 1$. Its unique solution is $\phi_{j,B}(t_j) = \int_0^1 Bv^{B-1} \exp\{t_j(1-v^j)/j\} Bv^{B-1} dv$. Then, the boundary conditions for the equation in (17) are $\phi_B(0, \ldots, 0, t_j, 0, \ldots, 0) = \phi_{j,B}(t_j)$ for $1 \leq j \leq n$. It can be checked that equation (17) has a unique solution, namely

$$\phi_B(t_1, \ldots, t_n) = \int_0^1 Bv^{B-1} \exp\left\{ \sum_{j=1}^{n} \frac{t_j}{j}(1-v^j) \right\} dv,$$

which immediately gives the description of the joint distribution of $\mathbf{Z}_B$ in Theorem 2.

We now give some corollaries of the main theorems. The first gives marginal factorial moments of the count $Z_{k,B}$.

Corollary 1. *For a $B-$harmonic Bernoulli sequence,*

$$E(Z_{k,B}^{[r]}) = \frac{r!}{(k+B)(2k+B)\cdots(rk+B)}$$

Proof. From Theorem 2,

$$E(Z_{k,B}^{[r]}) = \int_0^1 Bv^{B-1} \left(\frac{1-v^k}{k} \right)^r = \frac{r!}{(k+B)(2k+B)\cdots(rk+B)}.$$

□

The second corollary computes the covariance between $Z_{k_1,B}$ and $Z_{k_2,B}$.

Corollary 2.

$$cov(Z_{k_1,B}, Z_{k_2,B}) \;=\; \frac{B}{(k_1+B)(k_2+B)(k_1+k_2+B)}.$$

Proof. From (14) in Theorem 1, we have

$$\begin{aligned} E(Z_{k_1,B})E(Z_{k_2,B}) &= \frac{1}{(k_1+B)(k_2+B)} \\ E(Z_{k_1,B}Z_{k_2,B}) &= \frac{1}{(k_1+B)(k_1+k_2+B)} + \frac{1}{(k_2+B)(k_1+k_2+B)} \\ &= \frac{1}{(k_1+B)(k_2+B)} + \frac{B}{(k_1+B)(k_2+B)(k_1+k_2+B)} \end{aligned}$$

This shows that $Z_{k_1,B}$ and $Z_{k_2,B}$ are positively correlated and

$$cov(Z_{k_1,B}, Z_{k_2,B}) \;=\; \frac{B}{(k_1+B)(k_2+B)(k_1+k_2+B)}.$$

□

The FKG or positive association property of P_B is now established.

Theorem 3. *The joint distribution P_B of $\mathbf{Z}$ possesses the FKG property.*

Proof. Let f, g be a bounded functions on R^∞ which are coordinate-wise increasing and are supported on a finite number of coordinates. We need to show that

$$\int f(\mathbf{Z})g(\mathbf{Z}))dP_B \;\geq\; \int f(\mathbf{Z})dP_B \int g(\mathbf{Z})dP_B. \tag{18}$$

It is well known that distributions on the real line and products of measures on the real line possess the FKG property [Liggett (1985)]. Since the Poisson distribution is stochastically increasing in its intensity parameter, the product measure $P_{v,B}$ (cf. Theorem 2) is stochastically decreasing in v. This means that for any bounded increasing function f, $\int f(\mathbf{z})dP_{v,B}$ is decreasing in v. Thus

$$\begin{aligned} \int f(\mathbf{Z})g(\mathbf{Z})dP_B &= \int_0^1 Bv^{B-1} \int f(\mathbf{Z})g(\mathbf{Z})dP_{v,B}\; dv \\ &\geq \int_0^1 Bv^{B-1} \int f(\mathbf{Z})dP_{v,B} \int g(\mathbf{Z})dP_{v,B}\; dv \\ &\qquad \text{since } P_{v,B} \text{ is a product measure} \\ &\geq \int_0^1 Bv^{B-1} \int f(\mathbf{Z})dP_{v,B}dv \cdot \int_0^1 Bv^{B-1} \int g(\mathbf{Z})dP_{v,B}\; dv \\ &\qquad \text{since } \int f(\mathbf{Z})dP_{v,B}, \int g(\mathbf{Z})dP_{v,B}, \text{ decreases in } v \\ &= E_B(f(\mathbf{Z}))E_B(g(\mathbf{Z})). \end{aligned}$$

This completes the proof of this theorem. □

Finally, in the introduction, we stated that the parameter B cannot be estimated from $\mathbf{Z}$. This is a consequence of the fact below.

Theorem 4. *Let M_B be the joint distribution of the $B-$harmonic Bernoulli sequence $\{X_i\}$. Then for $0 \leq B < B'$, the measures M_B and $M_{B'}$ are absolutely continuous with respect to one another.*

Proof. Since $M_B, M_{B'}$ are product measures, we compute the Kakutani dichotomy criterion

$$\prod_{k\geq 1}\left[\frac{1}{\sqrt{(k+B)(k+B')}}+\sqrt{1-\frac{1}{k+B}}\sqrt{1-\frac{1}{k+B'}}\right] = \prod_{k\geq 1}\left(1-\frac{1}{k^2}(1+o(1))\right) > 0.$$

Thus for $B \neq B'$, $M_B << M_{B'}$. This also implies that $P_B = M_B\mathbf{Z}^{-1} << P_{B'} = M_{B'}\mathbf{Z}^{-1}$. This proves this theorem. □

Acknowledgment

We thank Prof. K.B. Athreya who brought to our attention an initial version of the problem, Prof. A. Joffe who sent us [Joffe et al. (2002)], and Fred Huffer who gave the suggestion to use independent exponentials in Theorem 2.

This research was supported in part by grant NSF/DMS – 0071504.

References

[Antzoulakos and Chadjiconstantinidis (2001)] Antzoulakos, D., and Chadjiconstantinidis, S. (2001) Distributions of numbers of success runs of fixed length in Markov dependent trials. *Ann. Inst. Statist. Math.* **53** 599-619. MR1868894

[Arratia et al. (2003)] Arratia, R., Barbour, A. D., and Tavaré, S. (2003) *Logarithmic Combinatorial Structures: A Probabilistic Approach.* EMS Monographs in Mathematics, European Mathematical Society, Zürich. MR2032426

[Arratia (1992)] Arratia, R., and Tavare, S. (1992) The cycle structure of random permutations. *Ann. Probab.* **20** 1567-1591. MR1175278

[Durrett (1995)] Durrett, R. (1995) *Probability: Theory and Examples.* Duxbury, New York.

[Emery (1996)] Emery, M. (1998) *Sur un problème de Diaconis* – Unpublished manuscript.

[Joffe et al. (2002)] Joffe, A., Marchand, E., Perron, F., and Popadiuk, P. (2002) On sums of products of Bernoulli variables and random permutations. *pre-print.*

[Johnson et al. (1992)] Johnson, N. L., Kotz, S., and Kemp, A. W. (1992) *Univariate Discrete Distributions.* Second Edition, Wiley, New York. MR1224449

[Liggett (1985)] Liggett, T. M. (1985) *Interacting Particle Systems.* Springer-Verlag, New York. MR776231

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 153–163
© Institute of Mathematical Statistics, 2004

Chebyshev polynomials and G-distributed functions of F-distributed variables

Anirban DasGupta[1] **and L. Shepp**[2]

Purdue University and Rutgers University

Abstract: We address a more general version of a classic question in probability theory. Suppose $\mathbf{X} \sim \mathbf{N_p}(\mu, \mathbf{\Sigma})$. What functions of $\mathbf{X}$ also have the $N_p(\mu, \Sigma)$ distribution? For $p = 1$, we give a general result on functions that cannot have this special property. On the other hand, for the $p = 2, 3$ cases, we give a family of new nonlinear and non-analytic functions with this property by using the Chebyshev polynomials of the first, second and the third kind. As a consequence, a family of rational functions of a Cauchy-distributed variable are seen to be also Cauchy distributed. Also, with three i.i.d. $N(0, 1)$ variables, we provide a family of functions of them each of which is distributed as the symmetric stable law with exponent $\frac{1}{2}$. The article starts with a result with astronomical origin on the reciprocal of the square root of an infinite sum of nonlinear functions of normal variables being also normally distributed; this result, aside from its astronomical interest, illustrates the complexity of functions of normal variables that can also be normally distributed.

1. Introduction

It is a pleasure for both of us to be writing to honor Herman. We have known and admired Herman for as long as we can remember. This particular topic is close to Herman's heart; he has given us many cute facts over the years. Here are some to him in reciprocation.

Suppose a real random variable $X \sim N(\mu, \sigma^2)$. What functions of X are also normally distributed? In the one dimensional case, an analytic map other than the linear ones cannot also be normally distributed; in higher dimensions, this is not true. Also, it is not possible for any one-to-one map other than the linear ones to be normally distributed. Textbook examples show that in the one dimensional case nonlinear functions $U(X)$, not analytic or one-to-one, can be normally distributed if X is normally distributed; for example, if $Z \sim N(0, 1)$ and Φ denotes the $N(0, 1)$ CDF, then, trivially, $U(Z) = \Phi^{-1}(2\Phi(|Z|) - 1)$ is also distributed as $N(0, 1)$. Note that this function $U(.)$ is not one-to-one; neither is it analytic.

One of the present authors pointed out the interesting fact that if X, Y are i.i.d. $N(0, 1)$, then the nonlinear functions $U(X, Y) = \frac{2XY}{\sqrt{X^2+Y^2}}$ and $V(X, Y) = \frac{X^2-Y^2}{\sqrt{X^2+Y^2}}$ are also i.i.d. $N(0, 1)$-distributed (see Shepp (1962), Feller (1966)). These are obviously nonlinear and not one-to-one functions of X, Y. We present a collection of new pairs of functions $U(X, Y), V(X, Y)$ that are i.i.d. $N(0, 1)$-distributed. The functions $U(X, Y), V(X, Y)$ are constructed by using the sequence of Chebyshev polynomials of the first, second and the third kind and the terrain corresponding to the plots of $U(X, Y), V(X, Y)$ gets increasingly more rugged, and yet with a visual regularity, as one progresses up the hierarchy. Certain other results about

[1]Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2068, USA. e-mail: dasgupta@stat.purdue.edu

[2]Department of Statistics, Rutgers University, Piscataway, NJ 08854-8019, USA. e-mail: shepp@stat.rutgers.edu

Keywords and phrases: analytic, Cauchy, Chebyshev polynomials, normal, one-to-one, three term recursion, stable law.

AMS 2000 subject classifications: 60E05, 05E35, 85A04.

Cauchy-distributed functions of a Cauchy-distributed variable and solutions of certain Fredholm integral equations follow as corollaries to these functions U, V being i.i.d. $N(0,1)$ distributed, which we point out briefly as a matter of fact of some additional potential interest. Using the family of functions $U(X,Y), V(X,Y)$, we also provide a family of functions $f(X,Y,Z), g(X,Y,Z), h(X,Y,Z)$ such that f, g, h are i.i.d. $N(0,1)$ if X, Y, Z are i.i.d. $N(0,1)$. The article ends with a family of functions of three i.i.d. $N(0,1)$ variables, each distributed as a symmetric stable law with exponent $\frac{1}{2}$; the construction uses the Chebyshev polynomials once again.

We start with an interesting example with astronomical origin of the reciprocal of the square root of an infinite sum of dependent nonlinear functions of normally distributed variables being distributed as a normal again. The result also is relevant in the study of total signal received at a telephone base station when a fraction of the signal emitted by each wireless telephone gets lost due to various interferences. See Heath and Shepp (2003) for description of both the astronomical and the telephone signal problem. Besides the quite curious fact that it should be normally distributed at all, this result illustrates the complexity of functions of normal variables that can also be normally distributed.

2. Normal function of an infinite i.i.d. $N(0,1)$ sequence: An astronomy example

Proposition 1. *Suppose $\eta_0, \eta_1, \eta_2, \ldots$ is a sequence of i.i.d. $N(0,1)$ variables. We show the following remarkable fact: let $S_n = \sum_{k=1}^{2n} \eta_k^2$. Then*

$$N = \frac{\operatorname{sgn}(\eta_0)}{\sqrt{\sum_{n=1}^{\infty} \frac{1}{S_n^2}}} \sim N\left(0, \frac{8}{\pi}\right).$$

The problem has an astronomical origin. Consider a fixed plane and suppose stars are distributed in the plane according to a homogeneous Poisson process with intensity λ; assume λ to be 1 for convenience. Suppose now that each star emits a constant amount of radiation, say a unit amount, and that an amount inversely proportional to some power k of the star's distance from a fixed point (say the origin) reaches the point.If $k = 4$, then the total amount of light reaching the origin would equal $L = \pi^2 \sum_{n=1}^{\infty} \frac{1}{(\gamma_1+\gamma_2+\cdots+\gamma_n)^2}$, where the γ_i are i.i.d. standard exponentials, because if the ordered distances of the stars from the origin are denoted by $R_1 < R_2 < R_3 < \ldots$, then $R_n^2 \sim \frac{1}{\pi}(\gamma_1 + \gamma_2 + \cdots + \gamma_n)$, where the γ_i are i.i.d. standard exponentials. Since the sum of squares of two i.i.d. standard normals is an exponential with mean 2, it follows that L has the same distribution as $\frac{4\pi^2}{N^2}$, where N is as in Proposition 1 above. In particular, L does not have a finite mean. Earlier contributions to this problem are due to Chandrasekhar, Cox, and others; for detailed references, see Heath and Shepp (2003).

To prove the Proposition, we will show the following two facts:

(a) The Laplace transform of $\sum_{n=1}^{\infty} \frac{1}{R_n^4}$ equals $Ee^{-\lambda \sum_{n=1}^{\infty} \frac{1}{R_n^4}} = e^{-\pi^{\frac{3}{2}}\sqrt{\lambda}}$.

(b) If $\eta \sim N(0,1)$, then the Laplace transform of $\frac{1}{\eta^2}$ equals $e^{-\sqrt{2\lambda}}$.

To prove (a), consider the more general Laplace transform of the sum of the fourth powers of the reciprocals of $R \in S$ only for $0 < a < R < b$, where a, b are fixed, $\phi(\lambda, a, b) = Ee^{-\lambda \sum_{\{a<R<b\}} \frac{1}{R^4}}$. We want $\phi(\lambda, 0, \infty)$, but we can write the

"recurrence" relation:

$$\phi(\lambda, a, b) = e^{-\pi(b^2-a^2)} + \int_a^b e^{-\pi(r^2-a^2)}\phi(\lambda, r, b)e^{-\lambda r^{-4}}2\pi r\, dr$$

where the first term considers the possibility that there are no points of S in the annulus $a < r < b$ and the integral is written by summing over the location of the point in the annulus with the smallest value of $R = r$ and then using the independence properties of the Poisson random set.
Now multiply both sides by $e^{-\pi a^2}$ and differentiate on a, regarding both b and λ as fixed constants, to get

$$\big(-2\pi a\phi(\lambda, a, b) + \phi'(\lambda, a, b)\big)e^{-\pi a^2} = -2\pi a e^{-\pi a^2}\phi(\lambda, a, b)e^{-\lambda a^{-4}}.$$

Dividing by $e^{-\pi a^2}$ and solving the simple differential equation for $\phi(\lambda, a, b)$, we get,

$$\phi(\lambda, a, b) = \phi(\lambda, 0, b)e^{2\pi\int_0^a (1-e^{-\lambda u^{-4}})u du}.$$

Since $\phi(\lambda, b, b) = 1$, we find that

$$\phi(\lambda, 0, b) = e^{-2\pi\int_0^b (1-e^{-\lambda u^{-4}})u\, du}.$$

Finally let $b \to \infty$ to obtain $\phi(\lambda, 0, \infty)$ as was desired. Evaluating the integral by changing $u = t^{-\frac{1}{4}}$ and integration by parts, gives the answer stated in (a).

(b) can be proved by direct calculation, but a better way to see this is to use the fact that the hitting time, τ_1, of level one by a standard Brownian motion, $W(t), t \geq 0$, has the same distribution as η^{-2} using the reflection principle,

$$\begin{aligned} P(\tau_1 < t) &= P\big(\max W(u), u \in [0, t] > 1\big) = 2P\big(W(t) > 1\big) = P\big(\sqrt{t}|\eta| > 1\big) \\ &= P\big(\eta^{-2} < t\big). \end{aligned}$$

Finally, Wald's identity

$$Ee^{\lambda W(\tau_1) - \frac{\lambda^2}{2}\tau_1} = 1, \lambda > 0,$$

and the fact that $W(\tau_1) = 1$ gives the Laplace transform of τ_1 and hence also of η^{-2}, as

$$Ee^{-\lambda\eta^{-2}} = Ee^{-\lambda\tau_1} = e^{-\sqrt{2\lambda}}.$$

This completes the proof of Proposition 1 and illustrates the complexity of functions of normal variables that can also be normally distributed.

3. Chebyshev polynomials and normal functions

3.1. A general result

First we give a general result on large classes of functions of a random variable Z that *cannot* have the same distribution as that of Z. The result is much more general than the special case of Z being normal.

Proposition 2. *Let Z have a density that is symmetric, bounded, continuous, and everywhere strictly positive. If $f(Z) \neq \pm Z$ is either one-to-one, or has a zero derivative at some point and has a uniformly bounded derivative of some order $r \geq 2$, then $f(Z)$ cannot have the same distribution as Z.*

Proof. It is obvious that if $f(z)$ is one-to-one then Z and $f(Z)$ cannot have the same distribution under the stated conditions on the density of Z, unless $f(z) = \pm z$.

Consider now the case that $f(z)$ has a zero derivative at some point; let us take this point to be 0 for notational convenience. Let us also suppose that $|f^{(r)}(z)| \le K$ for all z, for some $K < \infty$. Suppose such a function $f(Z)$ has the same distribution as Z.

Denote $f(0) = \alpha$; then $P(|f(Z) - \alpha| \le \epsilon) = P(|Z - \alpha| \le \epsilon) \le c_1\epsilon$ for some $c_1 < \infty$, because of the boundedness assumption on the density of Z.

On the other hand, by a Taylor expansion around 0, $f(z) = \alpha + \frac{z^2}{2}f''(0) + \cdots + \frac{z^r}{r!}f^{(r)}(z^*)$, at some point between 0 and z. By the uniform boundedness condition on $f^{(r)}(z)$, from here, one has $P(|f(Z)-\alpha| \le \epsilon) \ge P(a_2|Z|^2 + a_3|Z|^3 + \cdots a_r|Z|^r \le \epsilon)$, for some fixed positive constants $a_2, a_3, \ldots, a_r$. For sufficiently small $\epsilon > 0$, this implies that $P(|f(Z) - \alpha| \le \epsilon) \ge P(M|Z|^2 \le \epsilon)$, for a suitable positive constant M.

However, $P(M|Z|^2 \le \epsilon) \ge c_2\sqrt{\epsilon}$ for some $0 < c_2 < \infty$, due to the strict positivity and continuity of the density of Z. This will contradict the first bound $P(|f(Z) - \alpha| \le \epsilon) \le c_1\epsilon$ for small ϵ, hence completing the proof. □

3.2. Normal functions of two i.i.d. $N(0,1)$ variables

Following standard notation, let $T_n(x), U_n(x)$ and $V_n(x)$ denote the nth Chebyshev polynomial of the first, second and third kind. Then for all $n \ge 1$, the pairs of functions (Z_n, W_n) in the following result are i.i.d. $N(0,1)$ distributed.

Proposition 3. *Let $X, Y \overset{i.i.d.}{\sim} N(0,1)$. For $n \ge 1$, let*

$$Z_n = YU_{n-1}\left(\frac{X}{\sqrt{X^2+Y^2}}\right), \qquad \text{and}$$
$$W_n = \sqrt{X^2+Y^2}\,T_n\left(\frac{X}{\sqrt{X^2+Y^2}}\right).$$

Then, $Z_n, W_n \overset{i.i.d.}{\sim} N(0,1)$.

There is nothing special about X, Y being i.i.d. By taking a bivariate normal vector, orthogonalizing it to a pair of i.i.d. normals, applying Proposition 3 to the i.i.d. pair, and then finally retransforming to the bivariate normal again, one similarly finds nonlinear functions of a bivariate normal that have exactly the same bivariate normal distribution as well. Here is a formal statement.

Corollary 1. *Suppose $(X_1, X_2) \sim N(0,0,1,1,\rho)$. Then, for all $n \ge 1$, the pairs of functions (Y_{1n}, Y_{2n}) defined as*

$$\begin{aligned} Y_{1n} &= X_2 U_{n-1}\left(\frac{X_1 - \rho X_2}{\sqrt{X_1^2 + (1+\rho^2)X_2^2 - 2\rho X_1X_2}}\right), \\ Y_{2n} &= \rho Y_{1n} + \sqrt{1-\rho^2}\sqrt{X_1^2 + (1+\rho^2)X_2^2 - 2\rho X_1X_2}\,T_n \\ &\quad \times \left(\frac{X_1 - \rho X_2}{\sqrt{X_1^2 + (1+\rho^2)X_2^2 - 2\rho X_1X_2}}\right), \end{aligned}$$

are also distributed as $N(0,0,1,1,\rho)$.

The first few members of the polynomials $T_n(x), U_n(x)$ are $T_1(x) = x, T_2(x) = 2x^2 - 1, T_3(x) = 4x^3 - 3x, T_4(x) = 8x^4 - 8x^2 + 1, T_5(x) = 16x^5 - 20x^3 + 5x, T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$, and $U_0(x) = 1, U_1(x) = 2x, U_2(x) = 4x^2 - 1, U_3(x) = 8x^3 - 4x, U_4(x) = 16x^4 - 12x^2 + 1, U_5(x) = 32x^5 - 32x^3 + 6x$; see, e.g, Mason and Handscomb (2003). Plugging these into the formulae for Z_n and W_n in Proposition 3, the following illustrative pairs of i.i.d. $N(0,1)$ functions of i.i.d. $N(0,1)$ variables X, Y are obtained.

Example 1. Pairs of i.i.d. $N(0,1)$ Distributed Functions when $X, Y \overset{i.i.d.}{\sim} N(0,1)$.

$$\frac{2XY}{\sqrt{X^2+Y^2}} \quad \text{and} \quad \frac{X^2-Y^2}{\sqrt{X^2+Y^2}} \quad \text{(Shepp's example)}$$

$$\big(3X^2-Y^2\big)\frac{Y}{X^2+Y^2} \quad \text{and} \quad \big(X^2-3Y^2\big)\frac{X}{X^2+Y^2}$$

$$\frac{X^4-6X^2Y^2+Y^4}{(X^2+Y^2)^{\frac{3}{2}}} \quad \text{and} \quad \frac{4XY(X^2-Y^2)}{(X^2+Y^2)^{\frac{3}{2}}}$$

$$\big(5X^4-10X^2Y^2+Y^4\big)\frac{Y}{(X^2+Y^2)^2} \quad \text{and} \quad \big(5Y^4-10X^2Y^2+X^4\big)\frac{X}{(X^2+Y^2)^2}$$

$$\frac{6X^5Y-20X^3Y^3+6XY^5}{(X^2+Y^2)^{\frac{5}{2}}} \quad \text{and} \quad \frac{X^6-15X^4Y^2+15X^2Y^4-Y^6}{(X^2+Y^2)^{\frac{5}{2}}}.$$

Remark 1. Since $Z_n(X,Y)$ and $W_n(X,Y)$ are i.i.d. $N(0,1)$ whenever X, Y are i.i.d. $N(0,1)$, one would get an i.i.d. pair of standard normals by considering the functions $Z_m(Z_n(X,Y), W_n(X,Y))$ and $W_m(Z_n(X,Y), W_n(X,Y))$. It is interesting that $Z_m(Z_n(X,Y), W_n(X,Y)) = Z_{mn}(X,Y)$ and $W_m(Z_n(X,Y), W_n(X,Y)) = W_{mn}(X,Y)$. Thus, iterations of the functions in Proposition 3 produce members of the same sequence.

Remark 2. Consider the second pair of functions in Example 1. One notices that but for a sign, the second function is obtained by plugging Y for X and X for Y in the first function. It is of course obvious that because X, Y are i.i.d., by writing Y for X and X for Y, we cannot change the distribution of the function. What is interesting is that this operation produces a function independent of the first function. This in fact occurs for all the even numbered pairs, as is formally stated in the following proposition.

Proposition 4. *For every $n \geq 0, W_{2n+1}(X,Y) = (-1)^n Z_{2n+1}(Y,X)$, and hence, for every $n \geq 0, Z_{2n+1}(X,Y)$ and $Z_{2n+1}(Y,X)$ are independently distributed.*

Progressively more rugged plots are obtained by plotting the functions $Z_n(x,y)$ and $W_n(x,y)$ as n increases; despite the greater ruggedness, the plots also get visually more appealing. A few of the plots are presented next. The plots labeled as V correspond to the functions W of Proposition 3.

Analogous to the Chebyshev polynomials of the first and second kind, those of the third kind also produce standard normal variables. However, this time there is no independent mate.

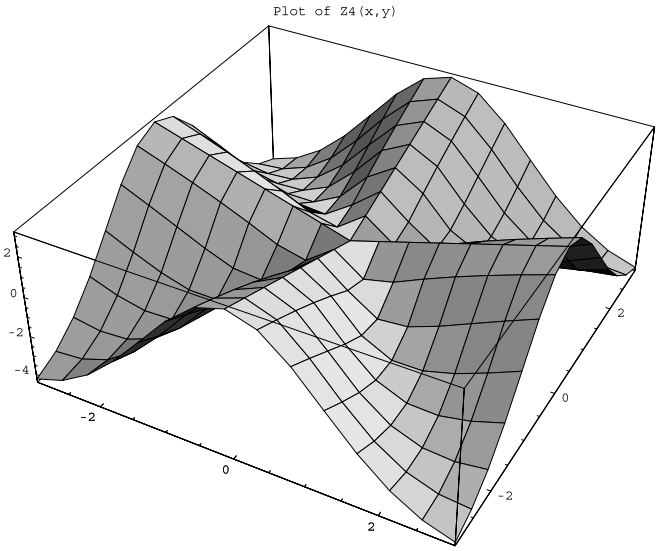
Proposition 5. *Let $X, Y \overset{i.i.d.}{\sim} N(0,1)$. For $n \geq 1$, let*

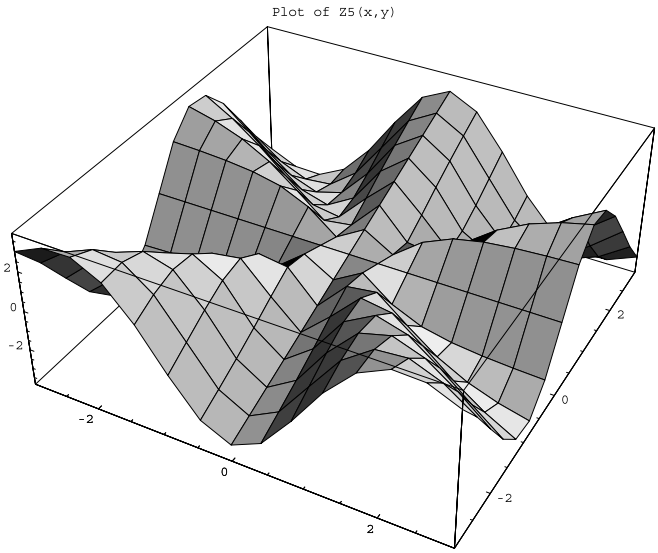
$$Q_n = \frac{\operatorname{sgn}(Y)}{\sqrt{2}}\sqrt{X^2+Y^2+X\sqrt{X^2+Y^2}}\,V_n\left(\frac{X}{\sqrt{X^2+Y^2}}\right).$$

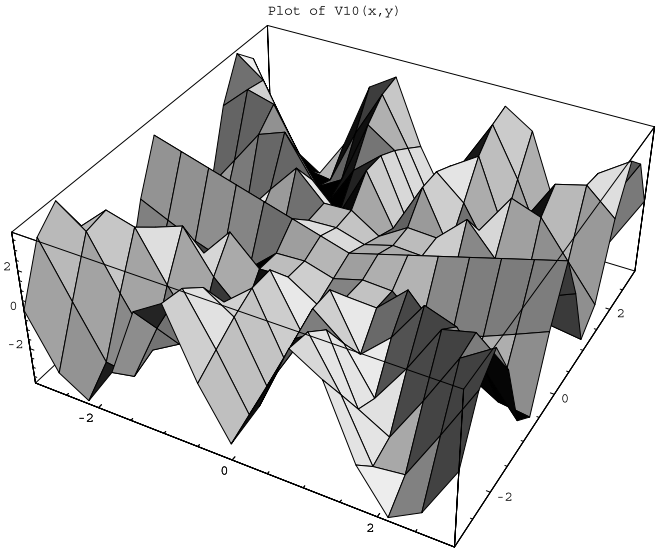
Then $Q_n \sim N(0,1)$.

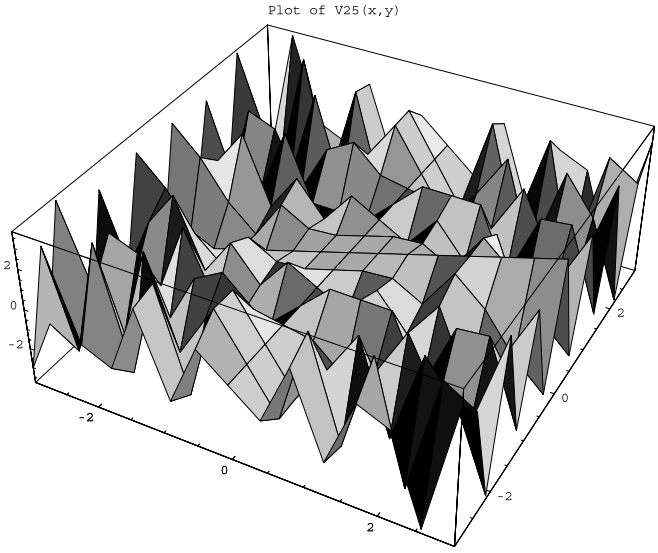
The first few polynomials $V_n(x)$ are $V_1(x) = 2x-1, V_2(x) = 4x^2-2x-1, V_3(x) = 8x^3-4x^2-4x+1, V_4(x) = 16x^4-8x^3-12x^2+4x+1$. Plugging these into the formula for Q_n, a sequence of increasingly complex standard normal functions of X, Y are obtained.

For example, using $n = 1$, if X, Y are i.i.d. $N(0,1)$, then $\frac{\operatorname{sgn}(Y)}{\sqrt{2}}(2X - \sqrt{X^2+Y^2})\sqrt{1+\frac{X}{\sqrt{X^2+Y^2}}}$ is distributed as $N(0,1)$. In comparison to the $N(0,1)$ functions Z_2, W_2 in Section 3.2, this is a more complex function with a $N(0,1)$ distribution.


Plot of Z4(x,y)
2
0
-2
-4
-2
0
2
2
0
-2


Plot of Z5(x,y)
2
0
-2
-2
0
2
2
0
-2


Plot of V10(x,y)
2
0
-2
-2
0
2
2
0
-2


Plot of V25(x,y)
2
0
-2
-2
0
2
2
0
-2

3.3. The case of three

It is interesting to construct explicitly three i.i.d. $N(0,1)$ functions $f(X,Y,Z)$, $g(X,Y,Z), h(X,Y,Z)$ of three i.i.d. $N(0,1)$ variables X,Y,Z. In this section, we present a method to explicitly construct such triplets of functions $f(X,Y,Z)$, $g(X,Y,Z), h(X,Y,Z)$ by using Chebyshev polynomials, as in the case with two of them. The functions f,g,h we construct are described below.

Proposition 6. *Let $X,Y,Z \overset{i.i.d.}{\sim} N(0,1)$. If $U(X,Y), V(X,Y)$ are i.i.d. $N(0,1)$, then $f(X,Y,Z), g(X,Y,Z), h(X,Y,Z)$ defined as*

$$\begin{aligned} f(X,Y,Z) &= U\big(V(X,Y), V\big(U(X,Y),Z\big)\big), \\ g(X,Y,Z) &= V\big(V(X,Y), V\big(U(X,Y),Z\big)\big), \\ h(X,Y,Z) &= U\big(U(X,Y),Z\big) \end{aligned}$$

are also distributed as i.i.d. $N(0,1)$.

Example 2. For $U(X,Y), V(X,Y)$, we can use the pair of i.i.d. $N(0,1)$ functions of Proposition 3. This will give a family of i.i.d. $N(0,1)$ functions f,g,h of X,Y,Z. The first two functions f,g of Proposition 6 are too complicated even when we use $U = Z_2$ and $V = W_2$ of Proposition 3. But the third function h is reasonably tidy. For example, using $U = Z_n$, and $V = W_n$ with $n = 2$, one gets the following distributed as $N(0,1)$:

$$h(X,Y,Z) = \frac{4XYZ}{\sqrt{4X^2Y^2 + Z^2(X^2+Y^2)}}.$$

4. Cauchy distributed functions, Fredholm integral equations and the stable law of exponent $\frac{1}{2}$

4.1. Cauchy distributed functions of Cauchy distributed variables

It follows from the result in Proposition 3 that if C has a $Cauchy(0,1)$ distribution, then appropriate sequences of rational functions $C\lambda_n(C)$ also have a $Cauchy(0,1)$ distribution. These results generalize the observations in Pitman and Williams (1967). This results, by consideration of characteristic functions, in the $Cauchy(0,1)$ density being solutions to a certain Fredholm integral equation of the first kind. This connection seems to be worth pointing out. First the functions $f_n(C)$ attributed to above are explicitly identified in the next result.

Proposition 7. *Let $C \sim Cauchy(0,1)$. Let $R = \frac{1}{\sqrt{1+C^2}}$ and for $k \geq 1$,*

$$\begin{aligned} f_k(C) &= \frac{1 + 2T_2(R) + 2T_4(R) + \cdots + T_{2k}(R)}{T_{2k}(R)}, \qquad \text{and} \\ g_k(C) &= \frac{2T_1(R) + 2T_3(R) + \cdots + T_{2k+1}(R)}{T_{2k+1}(R)}. \end{aligned}$$

Then $Cf_k(C)$ and $Cg_k(C)$ are also $\sim Cauchy(0,1)$.

Example 3. The functions f_k, g_k for small values of k are as follows:

$$\begin{aligned} f_1(C) &= \frac{2}{1-C^2}; \qquad g_1(C) = \frac{C^2-3}{3C^2-1}; \\ f_2(C) &= \frac{4-4C^2}{C^4-6C^2+1}; \qquad g_2(C) = \frac{C^4-10C^2+5}{5C^4-10C^2+1}; \\ f_3(C) &= \frac{6C^4-20C^2+6}{C^6-15C^4+15C^2-1}; \qquad g_3(C) = \frac{C^6-21C^4+35C^2-7}{7C^6-35C^4+21C^2-1}. \end{aligned}$$

Note that f_k, g_k are rational functions of C. Proposition 7 thus gives an infinite collection of rational functions, say $\lambda_n(C)$, such that $C\lambda_n(C) \sim C \forall n$. This implies the following result on Fredholm integral equations.

Proposition 8. *Consider the Fredholm integral equation* $\int_{-\infty}^{\infty} K(t,y)p(y)dy = g(t)$, *where* $K(t,y) = cos(ty\lambda(y))$ *and* $g(t) = e^{-|t|}$. *Then for any of the rational functions* $\lambda(y) = f_k(y), g_k(y)$ *in Proposition 7, the Cauchy*$(0,1)$ *density* $p(y) = \frac{1}{\pi(1+y^2)}$ *is a solution of the above Fredholm equation.*

4.2. The stable law with exponent $\frac{1}{2}$

Starting with three i.i.d. standard normal variables, one can construct an infinite collection of functions of them, each having a symmetric stable distribution with exponent $\frac{1}{2}$. The construction uses, as in the previous sections, the Chebyshev polynomials. It is described in the final result.

Proposition 9. *Let* X, Y, N *be i.i.d.* $N(0,1)$. *Then, for each* $n \geq 1$, $S_{1,n} = \frac{N}{Z_n(X,Y)W_n^2(X,Y)}$, *as well as* $S_{2,n} = \frac{N}{W_n(X,Y)Z_n^2(X,Y)}$ *have a symmetric stable distribution with exponent* $\frac{1}{2}$.

Example 4. Using $n = 2, 3$, the following are distributed as a symmetric stable law of exponent $\frac{1}{2}$:

$$N\frac{(X^2+Y^2)^{\frac{3}{2}}}{4X^2Y^2(X^2-Y^2)} \quad \text{and} \quad N\frac{(X^2+Y^2)^{\frac{3}{2}}}{2XY(X^2-Y^2)^2};$$

$$N\frac{(X^2+Y^2)^3}{XY^2(X^2-3Y^2)(3X^2-Y^2)^2} \quad \text{and} \quad N\frac{(X^2+Y^2)^3}{X^2Y(3X^2-Y^2)(X^2-3Y^2)^2}.$$

5. Appendix

Proof of Proposition 3. Proposition 3 is a restatement of the well known fact that if X, Y are i.i.d. $N(0,1)$, and if r, θ denote their polar coordinates, then for all $n \geq 1, r\cos n\theta$ and $r\sin n\theta$ are i.i.d. $N(0,1)$, and that the Chebyshev polynomials $T_n(x), U_n(x)$ are defined by $T_n(x) = \cos n\theta, U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}$ with $x = \cos\theta$.

Proof of Proposition 4. We need to prove that for all x, y,

$$\begin{aligned} xU_{2n}\left(\frac{y}{\sqrt{x^2+y^2}}\right) &= (-1)^n\sqrt{x^2+y^2}T_{2n+1}\left(\frac{x}{\sqrt{x^2+y^2}}\right) \\ &\Leftrightarrow \quad \forall w, \sqrt{1-w^2}U_{2n}(w) \\ &= (-1)^n T_{2n+1}\big(\sqrt{1-w^2}\big). \end{aligned}$$

Note now that

$$\frac{d}{dw}(-1)^n T_{2n+1}\big(\sqrt{1-w^2}\big) = (-1)^{n+1}\frac{w}{\sqrt{1-w^2}}(2n+1)U_{2n}\big(\sqrt{1-w^2}\big),$$

by using the identity

$$\frac{d}{dw}T_k(w) = kU_{k-1}(w).$$

On the other hand,

$$\begin{aligned} &\frac{d}{dw}\sqrt{1-w^2}U_{2n}(w) \\ &\quad = -\frac{w}{\sqrt{1-w^2}}U_{2n}(w) + \sqrt{1-w^2}\,\frac{(n+1)U_{2n-1}(w) - nU_{2n+1}(w)}{1-w^2}, \end{aligned}$$

by using the identity

$$\frac{d}{dw}U_k(w) = \frac{(k+2)U_{k-1}(w) - kU_{k+1}(w)}{2(1-w^2)};$$

see Mason and Handscomb (2003) for these derivative identities.

It is enough to show that the derivatives coincide. On some algebra, it is seen that the derivatives coincide iff $U_{2n-1}(w) - wU_{2n}(w) = (-1)^{n+1}wU_{2n}(\sqrt{1-w^2})$, which follows by induction and the three term recursion for the sequence U_n.

Proof of Proposition 5. Proposition 5, on some algebra, is a restatement of the definition of the Chebyshev polynomials of the third kind as $V_n(x) = \frac{\cos(n+\frac{1}{2})\theta}{\cos\frac{\theta}{2}}$. We omit the algebra.

Proof of Proposition 6. If X, Y, Z are i.i.d. $N(0,1)$, and $U(X,Y), V(X,Y)$ are also i.i.d. $N(0,1)$, then, obviously, $U(X,Y), V(X,Y), Z$ are i.i.d. $N(0,1)$. At the next step, use this fact with X, Y, Z replaced respectively by $U(X,Y), Z, V(X,Y)$. This results in $U(U(X,Y),Z), V(U(X,Y),Z), V(X,Y)$ being i.i.d. $N(0,1)$. Then as a final step, use this fact one more time with X, Y, Z replaced respectively by $V(X,Y), V(U(X,Y),Z), U(U(X,Y),Z)$. This completes the proof.

Proof of Proposition 7. From Proposition 3, $\frac{Z_n(X,Y)}{W_n(X,Y)} \sim Cauchy(0,1)$ for all $n \geq 1$. Thus, we need to reduce the ratio $\frac{Z_n(X,Y)}{W_n(X,Y)}$ to $Cf_k(C)$ when $n = 2k$ and to $Cg_k(C)$ when $n = 2k+1$, with C standing for the Cauchy-distributed variable $\frac{Y}{X}$.

The reduction for the two cases $n = 2k$ and $n = 2k+1$ follow, again on some algebra, on using the following three identities:

(i) $wU_{n-1}(w) = U_n(w) - T_n(w)$;

(ii) $U_{2k}(w) = T_0(w) + 2T_2(w) + \cdots + 2T_{2k}(w)$;

(iii) $U_{2k+1}(w) = 2T_1(w) + 2T_3(w) + \cdots + 2T_{2k+1}(w)$;

see Mason and Handscomb (2003) for the identities (i)–(iii). Again, we omit the algebra.

Proof of Proposition 8. Proposition 8 follows from Proposition 7 on using the facts that each f_k, g_k are even functions of C, and hence the characteristic function of $Cf_k(C)$ and $Cg_k(C)$ is the same as its Fourier cosine transform, and on using also the fact that the characteristic function of a $Cauchy(0,1)$ distributed variable is $e^{-|t|}$.

Proof of Proposition 9. Proposition 9 follows from Proposition 3 and the well known fact that for three i.i.d. standard normal variables X, Y, N, $\frac{N}{XY^2}$ is symmetric stable with exponent $\frac{1}{2}$; see, e.g., Kendall, Stuart and Ord (1987).

References

[1] Feller, W. (1966). *An Introduction to Probability Theory and its Applications, II*, John Wiley, New York. MR210154

[2] Kendall, M., Stuart, A. and Ord, J. K. (1987). *Advanced Theory of Statistics*, Vol. 1, Oxford University Press, New York. MR902361

[3] Heath, Susan and Shepp, Lawrence. *Olber's Paradox, Wireless Telephones, and Poisson Random Sets; Is the Universe Finite?*, Garden of Quanta – in honor of Hiroshi Ezawa, edited by J. Arafune, A. Arai, M. Kobayashi, K. Nakamura, T. Nakamura, I. Ojima, A. Tonomura and K. Watanabe (World Scientific Publishing Company, Pte. Ltd. Singapore, 2003). MR2045956

[4] Mason, J. C. and Handscomb, D. C.(2003). *Chebyshev Polynomials*, Chapman and Hall, CRC, New York. MR1937591

[5] Pitman, E. J. G. and Williams, E. J. (1967). Cauchy distributed functions of Cauchy sequences, *Ann. Math. Statist.*, **38**, 916–918. MR210166

[6] Shepp, L. (1962). Normal functions of normal random variables, *Siam Rev.*, **4**, 255–256.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 164–170
© Institute of Mathematical Statistics, 2004

Zeroes of infinitely differentiable characteristic functions

Herman Rubin[1] **and Thomas M. Sellke**[1]

Purdue University

Abstract: We characterize the sets where an n-dimensional, infinitely differentiable characteristic function can have its real part zero, positive, and negative, and where it can have its imaginary part zero, positive, and negative.

1. Introduction and summary

Let $f : \mathbb{R}^n \to \mathbb{C}$ be the characteristic function of a probability distribution on $\mathbb{R}^n$. Let $A^+ \subset \mathbb{R}^n$ be the set on which $\mathrm{Re}\{F(\cdot)\}$ is strictly positive, and let A^- be the set on which $\mathrm{Re}\{F(\cdot)\}$ is strictly negative. Let B^+ be the set on which $\mathrm{Im}\{f(\cdot)\}$ is strictly positive. What can we say about the sets A^+, A^- , and B^+? Since f is continuous, A^+, A^- , and B^+ are open sets. Since $f(t) = \overline{f(-t)}$ for all $t \in \mathbb{R}^n$, we have $A^+ = -A^+, A^- = -A^-$, and $B^+ \cap (-B^+) = \emptyset$. Clearly, $A^+ \cap A^- = \emptyset$. Finally, it follows from $f(0) = 1$ that $0 \in A^+$ and $0 \notin B^+$.

This paper will show that these obviously necessary conditions on the triple (A^+, A^-, B^+) are also sufficient to insure the existence of an n-dimensional characteristic function whose real part is positive precisely on A^+ and negative precisely on A^-, and whose imaginary part is positive precisely on B^+. Furthermore, this characteristic function may be taken to be infinitely differentiable.

Let $A^0 \subset \mathbb{R}^n$ be a closed set satisfying $0 \notin A^0$ and $A^0 = -A^0$. Let $B^0 \subset \mathbb{R}^n$ be a closed set containing 0 whose complement $(B^0)^c$ can be expressed as $(B^0)^c = B^+ \cup (-B^+)$, where B^+ is an open set satisfying $B^+ \cap (-B^+) = \emptyset$. It follows immediately from the main result that there exists an n-dimensional C^∞ characteristic function whose real part is zero precisely on A^0 and whose imaginary part is zero precisely on B^0. These sufficient conditions on A^0 and B^0 are obviously necessary.

Examples of one-dimensional characteristic functions with compact support are well known. However, the usual examples, and all those obtainable from the famous sufficient condition of Polya (see Theorem 6.5.3 of Chung (1974)) are not differentiable at zero, and the authors are not aware of any previously published examples of C^∞ characteristic functions with compact support.

2. Construction of the characteristic functions $g_{1,n}$ and $g_{2,n}$

For $x \in \mathbb{R}, x \neq 0$, define

$$r(x) = \frac{6}{x^2}\left(1 - \frac{\sin x}{x}\right).$$

Let $r(0) = 1$, so that r is continuous.

[1]Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA. hrubin@stat.purdue.edu; tsellke@stat.purdue.edu

Keywords and phrases: characteristic functions, zeroes.
AMS 2000 subject classifications: 60E10.

Lemma 1. *The characteristic function of the probability density* $(3/2)\{(1-|t|)^+\}^2$ *is* r.

Proof. Direct calculation. □

Lemma 2. *The function* r *is unimodal and positive.*

Proof. Since r is symmetric and since $r(0)=1$ and $\lim_{x\to\infty} r(x)=0$, it will suffice to prove that the first derivative $r'(\cdot)$ has no zeroes for $x\in(0,\infty)$. But

$$r'(x) = -\frac{6}{x^4}\big[(2+\cos x)x - 3\sin x\big],$$

so that it will suffice to prove that $w(\cdot)$ defined by

$$w(x) = (2+\cos x)x - 3\sin x$$

has no zeroes on $(0,\infty)$. It is easy to see that $w(x)$ is positive for $x\geq\pi$. To take care of $x\in(0,\pi)$, note that

$$\begin{aligned} w'(x) &= 2 - 2\cos x - x\sin x \\ w''(x) &= \sin x - x\cos x \\ w'''(x) &= x\sin x \end{aligned}$$

The third derivative $w'''(x)$ is positive for $x\in(0,\pi)$. Since $w''(0)=w'(0)=w(0)=0$, it follows that $w(x)$ is positive for $x\in(0,\pi)$, and we are done. □

Let $X_1, X_2, \ldots$ be $\ldots$ random variables with density $(3/2)\{(1-|t|)^+\}^2$. Define

$$S_1 = \sum_{k=1}^{\infty} X_k/k^2 \qquad \text{and} \qquad S_2 = \sum_{k=1}^{\infty} X_k/k^4.$$

Let h_1 be the density of S_1, and let h_2 be the density of S_2. Since $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$, the density h_1 is positive precisely on the interval $(-\pi^2/6, \pi^2/6)$. Likewise, since $\sum_{k=1}^{\infty} k^{-4} = \pi^4/90$, h_2 is positive precisely on $(-\pi^4/90, \pi^4/90)$.

It follows from Lemma 1 that the characteristic functions of S_1 and S_2 are given by

$$q_1(x) = \prod_{k=1}^{\infty} r\big(x/k^2\big) \qquad \text{and} \qquad q_2(x) = \prod_{k=1}^{\infty} r\big(x/k^4\big),$$

respectively.

By the Fourier inversion theorem (see the corollary on p. 155 of Chung (1974)),

$$h_j(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} q_j(x)\,dx,$$

for $j=1,2$. Setting $t=0$ yields

$$2\pi h_j(0) = \int_{-\infty}^{\infty} q_j(x)\,dx.$$

Thus, $\tilde{p}_j(\cdot)$ defined by

$$\tilde{p}_j(\cdot) = \frac{q_j(x)}{2\pi h_j(0)}$$

is a probability density with characteristic function given by

$$\tilde{g}_j(t) = h_j(t)/h_j(0), \qquad j=1,2.$$

Obviously, $\tilde{g}_1$ and $\tilde{g}_2$ are positive precisely on $(-\pi^2/6, \pi^2/6)$ and $(-\pi^4/90, \pi^4/90)$, respectively. Since $r(\cdot)$ is symmetric about 0 and unimodal, $\tilde{p}_1$ and $\tilde{p}_2$ are also symmetric and unimodal. From the definitions of $r(\cdot)$ and $q_j(\cdot)$ above, it is easy to see that

$$\lim_{x\to\infty} x^m \tilde{p}_j(x) = 0$$

for $j = 1, 2$ and for all $m > 0$. Thus, the densities $\tilde{p}_1$ and $\tilde{p}_2$ have all moments. It follows that $\tilde{g}_1$ and $\tilde{g}_2$ are C^∞. (See Theorem 6.4.1 of Chung (1974)). Finally, we need to show that the tails of $\tilde{p}_2$ are fatter than those of $\tilde{p}_1$ in the sense that, for each real $a > 0$,

$$\lim_{x\to\infty} \frac{\tilde{p}_1(ax)}{\tilde{p}_2(x)} = 0. \tag{2.1}$$

To do this, it will suffice to show that

$$\lim_{x\to\infty} \frac{q_1(ax)}{q_2(x)} = 0. \tag{2.2}$$

If $b, c > 0$, then obviously $\frac{r(bx)}{r(cx)} \to \frac{c^2}{b^2}$ as $x \to \infty$. Also, if $b > c > 0$, then $0 < \frac{r(bx)}{r(cx)} \leq 1$ for all $x \in \mathbb{R}$, by Lemma 2. But

$$\frac{q_1(ax)}{q_2(x)} = \prod_{k=1}^{\infty} \frac{r(ax/k^2)}{r(x/k^4)},$$

and the kth factor converges to $(a^2k^4)^{-1}$. There are only finitely many k's for which $(a^2k^4)^{-1} \geq 1$. If $(a^2k^4)^{-1} < 1$, then $0 < \frac{r(ax/k^2)}{r(x/k^4)} \leq 1$ for all x, and the limiting value $(a^2k^4)^{-1}$ can be made arbitrarily small by choosing k sufficiently large. This suffices to prove (2.2) and hence (2.1).

Define $g_1, g_2, p_1,$ and p_2 by rescaling $\tilde{g}_1, \tilde{g}_2, \tilde{p}_1,$ and $\tilde{p}_2$ as follows.

$$g_1(t) = \tilde{g}_1\big(\pi^2 t/6\big) \qquad g_2(t) = \tilde{g}_2\big(\pi^4 t/90\big)$$
$$p_1(x) = \big(6/\pi^2\big)\tilde{p}_1\big(6x/\pi^2\big) \qquad p_2(x) = \big(90/\pi^2\big)\tilde{p}_2\big(90x/\pi^4\big).$$

Our results for $\tilde{g}_1, \tilde{g}_2, \tilde{p}_1,$ and $\tilde{p}_2$ imply the results for $g_1, g_2, p_1,$ and p_2 given in the following lemma.

Lemma 3. *The functions g_1 and g_2 defined above are real-valued, nonnegative, C^∞ characteristic functions which are positive precisely on $(-1, 1)$. The corresponding probability densities p_1 and p_2 are unimodal, and the tails of p_2 are fatter then those of p_1 in the sense that, for every $a > 0$, $\lim_{x\to\infty} \frac{p_1(ax)}{p_2(x)} = 0$.*

In order to prove our main theorem, we will need an n-dimensional analog of Lemma 3. For the remainder of this paper, t and x will denote points in $\mathbb{R}^n$ with respective coordinates t_i and $x_i, i = 1, \ldots, n$.

For j=1 and 2, let $\mathbf{Y}_j$ be a random vector in $\mathbb{R}^n$ whose coordinates are i.i.d. random variables with density p_j. Then $\mathbf{Y}_j$ has density

$$\hat{p}_{j,n}(x) = \prod_{i=1}^{n} p_j(x_i)$$

and characteristic function

$$\hat{g}_{j,n}(t) = \prod_{i=1}^{n} g_j(t_i).$$

Let M be a random $n \times n$ orthogonal matrix (with the normalized Haar measure on the group of $n \times n$ orthogonal matrices as its probability distribution), and suppose M is independent of $\mathbf{Y}_j$. Then $\mathbf{Z}_j = M\mathbf{Y}_j$ is a spherically symmetric random vector in $\mathbb{R}^n$ with density

$$\tilde{p}_{j,n}(x) = \int_{S^{n-1}} \hat{p}_{j,n}\big(\|x\|u\big)d\upsilon(u),$$

where $S^{n-1} = \{t \in \mathbb{R}^n : \|t\| = 1\}$ is the unit sphere in $\mathbb{R}^n$, and υ is the rotation invariant probability measure on S^{n-1}. The characteristic function of $\mathbf{Z}_j$ is

$$\tilde{g}_{j,n}(t) = \int_{S^{n-1}} \hat{g}_{j,n}\big(\|t\|u\big)d\upsilon(u),$$

which is C^∞ and is positive precisely on $\{t \in \mathbb{R}^n : \|t\| < \sqrt{n}\}$. For j=1 and 2, let

$$g_{j,n^{(t)}} = \tilde{g}_{j,n}\big(\sqrt{n}t\,\big) \tag{2.3}$$

and

$$p_{j,n}(x) = n^{-1/2}\tilde{p}_{j,n}\big(n^{-1/2}x\big). \tag{2.4}$$

The following lemma gives us the results we need to prove the main theorem.

Lemma 4. *The functions $g_{1,n}$ and $g_{2,n}$ defined above are real-valued, nonnegative, C^∞ characteristic functions which are positive precisely on $\{t \in R^n : \|t\| < 1\}$. For each $a > 0$, there is a constant $L(a)$ such that the corresponding densities functions $p_{1,n}$ and $p_{2,n}$ satisfy*

$$p_{1,n}(ax) < L(a)p_{2,n}(x)$$

for all $x \in \mathbb{R}^n$.

Proof. Only the second assertion remains to be proved. Fix $a > 0$. It follows from Lemma 3 that there exists a number $K(a) > 0$ such that $p_1(ax_1) < K(a)p_2(x_1)$ for all $x_1 \in \mathbb{R}$. Thus

$$\hat{p}_{1,n}(ax) = \prod_{i=1}^{n} p_1(ax_i) < K^n(a)\prod_{i=1}^{n} p_2(x_i) = K^n(a)\hat{p}_{2,n}(x)$$

Furthermore,

$$\begin{aligned}\tilde{p}_{1,n}(ax) &= \int_{S^{n-1}} \hat{p}_{1,n}(a\|x\|u)\,dv(u) < K^n(a)\int_{S^{n-1}} \hat{p}_{2,n}(\|x\|u)\,dv(u)\\ &= K^n(a)\tilde{p}_{2,n}(x).\end{aligned}$$

Let $L(a) = K^n(a)$. Then it follows from (2.4) that $p_{1,n}(ax) < L(a)p_{2,n}(x)$ for all $x \in \mathbb{R}^n$. □

Remark. It is not hard to show that the spherically symmetric densities $p_{1,n}$ and $p_{2,n}$ are unimodal, and that, for each $a > 0$, they satisfy

$$\lim_{\|x\|\to\infty} \frac{p_{1,n}(ax)}{p_{2,n}(x)} = 0.$$

We will only need the facts given in Lemma 4, however.

3. The main theorem

Theorem. *Let A^+, A^-, and B^+ be open subsets of R^n satisfying $A^+ = -A^+, A^- = -A^-, B^+ \bigcap(-B^+) = \emptyset, A^+ \bigcap A^- = \emptyset, 0 \in A^+$, and $0 \notin B^+$. Then there exists an infinitely differentiable characteristic function f on $\mathbb{R}^n$ satisfying*

$$\begin{aligned} A^+ &= \{t \in \mathbb{R}^n : \mathrm{Re}(f(t)) > 0\} \\ A^- &= \{t \in \mathbb{R}^n : \mathrm{Re}(f(t)) < 0\} \\ B^+ &= \{t \in \mathbb{R}^n : \mathrm{Im}(f(t)) > 0\}. \end{aligned}$$

Proof. For $c \in \mathbb{R}^n$ and r a positive constant, let

$$B_r(c) = \{t \in \mathbb{R}^n : \|t - c\| < r\}$$

be the open ball in $\mathbb{R}^n$ with center c and radius r. We may assume without loss of generality that $B_1(0) \subset A^+$. Define

$$\tilde{A}^+ = A^+ \bigcap \{t \in \mathbb{R}^n : \|t\| > 1/2\}.$$

Since $\tilde{A}^+$ is open, it is the union of a countable set $\{B_{r_i}(c_i)\}_{i=1}^{\infty}$ of open balls. Since $\tilde{A}^+ = -\tilde{A}^+$, we have $B_{r_i}(-c_i) \subset \tilde{A}^+$ for all i. Define

$$f_i^+(t) = g_{1,n}\{(t - c_i)/r_i\} + g_{1,n}\{(t + c_i)/r_i\}.$$

By Lemma 4, f_i^+ is positive precisely on $B_{r_i}(c_i) \bigcup B_{r_i}(-c_i)$. Taking a Fourier transform yields

$$\begin{aligned} (2\pi)^{-n} \int_{\mathbb{R}^k} e^{-i(x\cdot t)} f_i^+(t)\, dt &= \{e^{-i(x\cdot c_i)} + e^{i(x\cdot c_i)}\} r_i p_{1,n}(r_i x) \\ &= 2r_i \cos(x \cdot c_i) p_{1,n}(r_i x) \end{aligned}$$

(See Theorem 7.7(c) of Rudin (1973)).

Let $\{\alpha_i\}_{i=1}^{\infty}$ be a sequence of positive constants satisfying $\alpha_i < 2^{-i-2} \times \{2r_i L(r_i)\}^{-1}$. Then

$$\left|(2\pi)^{-n} \int_{R^n} e^{-i(x\cdot t)} \sum_{i=1}^{\infty} \alpha_i f_i^+(t)\, dt\right| < \sum_{i=1}^{\infty} 2^{-i-2}\{L(r_i)\}^{-1} p_{1,n}(r_i x) < \frac{1}{4} p_{2,n}(x).$$

Furthermore, by choosing the α_i's to converge to zero sufficiently fast, we can insure that $f^+(\cdot)$ defined by

$$f^+(t) = \sum_{i=1}^{\infty} \alpha_i f_i^+(t)$$

is C^∞ and in $L^1(\mathbb{R}^n)$. Note that the real-valued, nonnegative function $f^+(\cdot)$ is nonzero precisely on A^+.

Let $\{B_{r_i'}(c_i')\}_{i=1}^{\infty}$ be a sequence of open balls whose union is A^-, and let

$$f_i^-(t) = -g_{1,n}\{(t - c_i')/r_i'\} - g_{1,n}\{(t + c_i')/r_i'\}.$$

The same argument used above shows that we can choose a sequence of positive constants $\{\beta_i\}_{i=1}^{\infty}$ such that $f^-(\cdot)$ defined by

$$f^-(t) = \sum_{i=1}^{\infty} \beta_i f_i^-(t)$$

is C^∞, in $L^1(\mathbb{R}^n)$, and satisfies

$$\left|(2\pi)^{-n}\int_{\mathbb{R}^k} e^{-(x\cdot t)}f^-(t)\,dt\right| < \frac{1}{4}p_{2,n}(x).$$

Note that the real-valued, nonpositive function $f^-(\cdot)$ is nonzero precisely on A^-.

Let $\{B_{r_i''}(c_i'')\}_{i=1}^\infty$ be a sequence of open balls whose union is B^+. Let

$$f_i^{im}(t) = i\big[g_{i,n}\big\{\big(t-c_i''\big)/r_i''\big\} - g_{1,n}\big\{\big(t+c_i''\big)/r_i''\big\}\big]$$

Then

$$\begin{aligned}(2\pi)^{-n}\int_{\mathbb{R}^k} e^{-i(x\cdot t)}f_i^{im}(t)\,dt &= \big\{e^{-i(x\cdot c_i'')} - e^{i(x\cdot c_i'')}\big\}r_i''p_{1,n}\big(r_i''x\big)\\ &= -2r_i''\sin\big(x\cdot c_i''\big)p_{1,n}\big(r_i''x\big)\end{aligned}$$

Again, we can choose a sequence of positive constants $\{\gamma_i\}_{i=1}$ so that $f^{im}(\cdot)$ defined by

$$f^{im}(t) = \sum_{i=1}^\infty \gamma_i f_i^{im}(t)$$

is C^∞, in $L^1(\mathbb{R}^n)$, and satisfies

$$\left|(2\pi)^{-n}\int_{\mathbb{R}^k} e^{-i(x\cdot t)}f^{im}(t)\,dt\right| < \frac{1}{4}p_{2,n}(x).$$

Note that the function $f^{im}(\cdot)$ is pure imaginary, and that its imaginary part is positive precisely on B^+.

Now let

$$f(t) = g_{2,n}(t) + f^+(t) + f^-(t) + f^{im}(t).$$

Clearly the real and imaginary parts of f are positive and negative on the proper sets. The function f is C^∞, and in $L^1(\mathbb{R}^n)$.

Define

$$p(x) = (2\pi)^{-n}\int_{\mathbb{R}^k} e^{-i(x\cdot t)}f(t)\,dt.$$

Since

$$\left|(2\pi)^{-n}\int_{\mathbb{R}^k} e^{-i(x\cdot t)}\big(f^+(t) + f^-(t) + f^{im}(t)\big)\,dt\right| < \frac{3}{4}p_{2,n}(x),$$

and

$$(2\pi)^{-n}\int_{\mathbb{R}^k} e^{-i(x\cdot t)}g_{2,n}(t)\,dt = p_{2,n}(x),$$

we have

$$\frac{1}{4}p_{2,n}(x) < p(x) < 2p_{2,n}(x).$$

By the Fourier inversion theorem (again, see Theorem 7.7(c) of Rudin (1973)),

$$f(t) = \int_{\mathbb{R}^k} e^{i(x\cdot t)}p(x)\,dx.$$

Also, since $f(0) = g_{2,n}(0) = 1$, we have

$$\int_{\mathbb{R}^k} p(x)\,dx = f(0) = 1.$$

Thus, $f$is the characteristic function of the probability density p, and f satisfies all the requirements of the theorem. □

Addendum

Except for slight corrections, the present paper was completed in 1984. Results very similar to the one-dimensional version of our main theorem appear in Sasvári (1985).

References

[1] Chung, K. L. (1974). *A Course in Probability Theory*, Academic Press, New York.

[2] Rudin, W. (1973). *Functional Analysis*, McGraw-Hill, New York. MR365062

[3] Sasvári, Z. (1985). Über die Nullstellenmenge von characteristischen Funktionen, *Math. Nachr.* **121**, 33–40. MR809312

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 171–179
© Institute of Mathematical Statistics, 2004

On the characteristic function of Pearson type IV distributions

Wei-Liem Loh[1]

National University of Singapore

Abstract: Using an identity of Stein (1986), this article gives an exact expression for the characteristic function of Pearson type IV distributions in terms of confluent hypergeometric functions.

1. Introduction

Pearson (1895) introduced a family of probability density functions where each member p of the family satisfies a differential equation

$$p^{(1)}(w) = -\frac{a+w}{a_2w^2+a_1w+a_0}p(w), \tag{1}$$

for some constants a, a_0, a_1 and a_2. The Pearson family is very general and it includes many of the probability distributions in common use today. For example, the beta distribution belongs to the class of Pearson type I distributions, the gamma distribution to Pearson type III distributions and the t distribution to Pearson type VII distributions.

This article focuses on the Pearson type IV distributions. These distributions have unlimited range in both directions and are unimodal. In particular, Pearson type IV distributions are characterized by members satisfying (1) with $0 < a_2 < 1$ and the equation

$$a_2w^2 + a_1w + a_0 = 0$$

having no real roots. Writing $A_0 = a_0 - a_1^2(4a_2)^{-1}$ and $A_1 = a_1(2a_2)^{-1}$, it follows from (1) that a Pearson type IV distribution has a probability density function of the form

$$p(w) = \frac{A}{[A_0 + a_2(w+A_1)^2]^{1/(2a_2)}} \exp\left[-\frac{a-A_1}{\sqrt{a_2A_0}} \arctan\left(\frac{w+A_1}{\sqrt{A_0/a_2}}\right)\right], \quad \forall w \in R,$$

where A is the normalizing constant. It is well known that Pearson type IV distributions are technically difficult to handle in practice [Stuart and Ord (1994), page 222]. Johnson, Kotz and Balakrishnan (1994), page 19, noted that working with $p(w)$ often leads to intractable mathematics, for example if one attempts to calculate its cumulative distribution function.

The main result of this article is an exact expression (see Theorem 2) for the characteristic function of a Pearson type IV distribution in terms of confluent hypergeometric functions. We note that we have been unable to find any non-asymptotic

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Republic of Singapore. e-mail: stalohwl@nus.edu.sg

Keywords and phrases: characteristic function, confluent hypergeometric function, Pearson type IV distribution, Stein's identity.

AMS 2000 subject classifications: primary 60E10; secondary 60E05, 62E15.

closed-form expression for the characteristic function of a Pearson type IV distribution in the literature.

The approach that we shall take is inspired by the results of Stein (1986) on the Pearson family of distributions. Since confluent hypergeometric functions have an extensive literature going back over two hundred years to Euler and Gauss, it is plausible that Theorem 2 may provide us with a way of understanding the behavior of Pearson type IV distributions better in a more rigorous manner.

For example, one possible use of Theorem 2 is that we can now apply Fourier analytic techniques in combination with Stein's method [see Stein (1986)] to obtain Pearson type IV approximations to the distribution of a sum of weakly dependent random variables. This work is currently in progress and hence will not be addressed here. The hope is that such a Pearson type IV approximation would have the same order of accuracy as that of an one-term Edgeworth expansion [see, for example, Feller (1971), page 539] with the (often desirable) property that the Pearson type IV approximation is a probability distribution whereas the one-term Edgeworth expansion is not.

We should also mention that besides one-term Edgeworth approximations, gamma and chi-square approximations exist in the literature [see, for example, Shorack (2000), page 383]. The latter approximations typically have the same order of accuracy as the former. However, gamma and chi-square approximations are supported on the half real line and may be qualitatively inappropriate for some applications.

Finally throughout this article, $\mathcal{I}\{.\}$ denotes the indicator function and for any function $h : R \to R$, we write $h^{(r)}$ as the rth derivative of h (if it exists) whenever $r = 1, 2, \cdots$.

2. Pearson type IV characteristic function

We shall first state an identity of Stein (1986) for Pearson type IV distributions.

Theorem 1 (Stein). *Let p be the probability density function of a Pearson type IV distribution satisfying*

$$p^{(1)}(w) = -\frac{(2\alpha_2 + 1)w + \alpha_1}{\alpha_2 w^2 + \alpha_1 w + \alpha_0} p(w), \qquad \forall w \in R, \tag{2}$$

for some constants α_0, α_1 and α_2. Then for a given bounded piecewise continuous function $h : R \to R$, the differential equation

$$(\alpha_2 w^2 + \alpha_1 w + \alpha_0) f^{(1)}(w) - w f(w) = h(w), \qquad \forall w \in R, \tag{3}$$

has a bounded continuous and piecewise continuously differentiable solution $f : R \to R$ if and only if

$$\int_{-\infty}^{\infty} h(w) p(w) dw = 0. \tag{4}$$

When (4) is satisfied, the unique bounded solution f of (3) is given by

$$\begin{aligned} f(w) &= \int_{-\infty}^{w} \frac{h(x)}{\alpha_2 x^2 + \alpha_1 x + \alpha_0} \exp\left(\int_{x}^{w} \frac{y dy}{\alpha_2 y^2 + \alpha_1 y + \alpha_0} \right) dx \\ &= -\int_{w}^{\infty} \frac{h(x)}{\alpha_2 x^2 + \alpha_1 x + \alpha_0} \exp\left(-\int_{w}^{x} \frac{y dy}{\alpha_2 y^2 + \alpha_1 y + \alpha_0} \right) dx, \qquad \forall w \in R. \end{aligned}$$

We refer the reader to Stein (1986), Chapter 6, for the proof of Theorem 1.

Let Z be a random variable having probability density function p where p satisfies (2).

Proposition 1. *Let Z be as above and ψ_Z be its characteristic function. Then ψ_Z satisfies the following homogeneous second order linear differential equation:*

$$\psi_Z^{(1)}(t) + t\alpha_0\psi_Z(t) - t\alpha_2\psi_Z^{(2)}(t) - it\alpha_1\psi_Z^{(1)}(t) = 0, \qquad \forall t \in R. \tag{5}$$

Proof. Since $\psi_Z(t) = Ee^{itZ}$, $t \in R$, we observe from Theorem 1 that

$$\begin{aligned}
& \int_{-\infty}^{\infty} [(\alpha_2 w^2 + \alpha_1 w + \alpha_0)\frac{d}{dw}(e^{itw}) - we^{itw}]p(w)dw \\
= & \int_{-\infty}^{\infty} [it(\alpha_2 w^2 + \alpha_1 w + \alpha_0)e^{itw} - we^{itw}]p(w)dw \\
= & (\alpha_2 w^2 + \alpha_1 w + \alpha_0)e^{itw}p(w)\big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} [(2\alpha_2 + 1)w + \alpha_1]e^{itw}p(w)dw \\
& - \int_{-\infty}^{\infty} (\alpha_2 w^2 + \alpha_1 w + \alpha_0)e^{itw}p^{(1)}(w)dw \\
= & 0.
\end{aligned}$$

Hence we conclude that

$$-it\alpha_2\psi_Z^{(2)}(t) + t\alpha_1\psi_Z^{(1)}(t) + it\alpha_0\psi_Z(t) + i\psi_Z^{(1)}(t) = 0, \qquad \forall t \in R.$$

This proves Proposition 1. □

Definition. Following Slater (1960), pages 2 to 5, we define the confluent hypergeometric function (with complex-valued parameters a and b) to be a power series in x of the form

$${}_1F_1(a; b; x) = \sum_{j=0}^{\infty} \frac{(a)_j x^j}{j!(b)_j},$$

where $(a)_j = a(a+1)\cdots(a+j-1)$, etc. and b is not a negative integer or 0. We further define

$$U(a; b; x) = \frac{\Gamma(1-b)}{\Gamma(1+a-b)}{}_1F_1(a; b; x) + \frac{\Gamma(b-1)}{\Gamma(a)}x^{1-b}{}_1F_1(1+a-b; 2-b; x).$$

Remark. It is well known [see for example Theorem 2.1.1 of Andrews, Askey and Roy (1999)] that the series ${}_1F_1(a; b; x)$ [and hence $U(a; b; x)$] converges absolutely for all x.

The theorem below establishes an explicit expression for $\psi_Z(t)$.

Theorem 2. *Let ψ_Z be as in Proposition 1,*

$$\begin{aligned}
\Delta &= \sqrt{4\alpha_0\alpha_2 - \alpha_1^2}, \\
r &= \frac{\sqrt{4\alpha_0\alpha_2 - \alpha_1^2}}{2\alpha_2} + \frac{i\alpha_1}{2\alpha_2}, \\
\bar{r} &= \frac{\sqrt{4\alpha_0\alpha_2 - \alpha_1^2}}{2\alpha_2} - \frac{i\alpha_1}{2\alpha_2}, \\
\nu &= 1 + \frac{1}{\alpha_2},
\end{aligned} \tag{6}$$

and $k\alpha_2 \neq 1$ *for all* $k = 1, 2, \cdots$. *Then for* $t \in R$, *we have*

$$\begin{aligned}\psi_Z(t) &= \frac{e^{-r|t|}\Gamma(\nu - r\Delta^{-1})}{\Gamma(\nu)} U(-\frac{r}{\Delta}; 1-\nu; \frac{\Delta|t|}{\alpha_2}) \mathcal{I}\{t \geq 0\} \\ &\quad + \frac{e^{-\bar{r}|t|}\Gamma(\nu - \bar{r}\Delta^{-1})}{\Gamma(\nu)} U(-\frac{\bar{r}}{\Delta}; 1-\nu; \frac{\Delta|t|}{\alpha_2}) \mathcal{I}\{t < 0\}.\end{aligned}$$

Remark. We would like to add that the confluent hypergeometric function $U(.;,;.)$ is available in a number of mathematical software packages. For example in *Mathematica* [Wolfram (1996)],

```
HypergeometricU[a, b, x]
```

is the command to evaluate $U(a; b; x)$.

Proof of Theorem 2. We observe from (5) that for all $t \in R$,

$$t\psi_Z^{(2)}(t) + (-\frac{1}{\alpha_2} + \frac{it\alpha_1}{\alpha_2})\psi_Z^{(1)}(t) - \frac{t\alpha_0}{\alpha_2}\psi_Z(t) = 0. \tag{7}$$

Step 1. Suppose that $t > 0$. We seek a solution of the above differential equation that has the form

$$\psi(t) = e^{-rt}\sum_{j=0}^{\infty} c_j t^j, \qquad \forall 0 < t < \infty,$$

for complex constants $c_0, c_1, \cdots$. Observing that

$$\begin{aligned}\psi^{(1)}(t) &= -re^{-rt}\sum_{j=0}^{\infty} c_j t^j + e^{-rt}\sum_{j=1}^{\infty} j c_j t^{j-1}, \\ \psi^{(2)}(t) &= r^2 e^{-rt}\sum_{j=0}^{\infty} c_j t^j - 2re^{-rt}\sum_{j=1}^{\infty} j c_j t^{j-1} + e^{-rt}\sum_{j=2}^{\infty} j(j-1)c_j t^{j-2},\end{aligned}$$

and substituting these expressions into the left hand side of (7), we have

$$\begin{aligned}&r^2 e^{-rt}\sum_{j=0}^{\infty} c_j t^{j+1} - 2re^{-rt}\sum_{j=1}^{\infty} j c_j t^j + e^{-rt}\sum_{j=2}^{\infty} j(j-1)c_j t^{j-1} \\ &+(-\frac{1}{\alpha_2} + \frac{it\alpha_1}{\alpha_2})(-re^{-rt}\sum_{j=0}^{\infty} c_j t^j + e^{-rt}\sum_{j=1}^{\infty} j c_j t^{j-1}) - \frac{t\alpha_0 e^{-rt}}{\alpha_2}\sum_{j=0}^{\infty} c_j t^j \qquad (8) \\ = \;& 0, \qquad \forall 0 < t < \infty.\end{aligned}$$

Equating the coefficient of t^0 in (8) to zero, we have

$$\frac{rc_0}{\alpha_2} - \frac{c_1}{\alpha_2} = 0,$$

and equating the coefficient of $t^j, j = 1, 2, \cdots$, in (8) to zero, we have

$$\begin{aligned}&r^2 c_{j-1} - 2rjc_j + j(j+1)c_{j+1} + \frac{rc_j}{\alpha_2} - \frac{(j+1)c_{j+1}}{\alpha_2} - \frac{i\alpha_1 r c_{j-1}}{\alpha_2} + \frac{i\alpha_1 j c_j}{\alpha_2} - \frac{\alpha_0 c_{j-1}}{\alpha_2} \\ &= 0.\end{aligned}$$

This implies that $c_1 = rc_0$, and in general for $j = 2, 3, \cdots$,

$$\begin{aligned} c_j &= \frac{1}{j[1-(j-1)\alpha_2]} \{c_{j-1}[r - 2r(j-1)\alpha_2 + i(j-1)\alpha_1] \\ &\quad + c_{j-2}(-\alpha_0 + r^2\alpha_2 - ir\alpha_1)\}, \end{aligned}$$

whenever $k\alpha_{n,2} \neq 1, k = 1, 2, \cdots$. We observe from (6) that r satisfies

$$r^2\alpha_2 - i\alpha_1 r - \alpha_0 = 0.$$

Since $4\alpha_0\alpha_2 > \alpha_1^2$ (from the definition of Pearson type IV distributions), we conclude that

$$\begin{aligned} c_j &= \frac{c_{j-1}}{j[1-(j-1)\alpha_2]}[r - 2r(j-1)\alpha_2 + i(j-1)\alpha_1] \\ &= c_{j-2} \prod_{k=j-1}^{j} \left\{ \frac{r - 2r(k-1)\alpha_2 + i(k-1)\alpha_1}{k[1-(k-1)\alpha_2]} \right\} \\ &= c_0 \prod_{k=1}^{j} \left\{ \frac{r - 2r(k-1)\alpha_2 + i(k-1)\alpha_1}{k[1-(k-1)\alpha_2]} \right\}, \quad \forall j = 1, 2, \cdots, \end{aligned}$$

and hence for $t \geq 0$,

$$\begin{aligned} \psi(t) &= c_0 e^{-r|t|} \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \prod_{k=1}^{j} \left\{ \frac{r - 2r(k-1)\alpha_2 + i(k-1)\alpha_1}{1-(k-1)\alpha_2} \right\} \\ &= c_0 e^{-r|t|} \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \prod_{k=1}^{j} \left\{ \frac{(k-1)\Delta - r}{\alpha_2(k-1) - 1} \right\} \\ &= c_0 e^{-r|t|} {}_1F_1(-\frac{r}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2}). \end{aligned} \tag{9}$$

Step 2. Suppose that $t < 0$. Writing $\xi = -t$ and $u_Z(\xi) = \psi_Z(t)$, we have

$$\psi_Z^{(1)}(t) = \frac{du_Z(\xi)}{d\xi}\frac{d\xi}{dt} = -u_Z^{(1)}(\xi),$$

and

$$\psi_Z^{(2)}(t) = \frac{d}{d\xi}(-\frac{du_Z(\xi)}{d\xi})\frac{d\xi}{dt} = u_Z^{(2)}(\xi).$$

Consequently, (5) now takes the form

$$\xi u_Z^{(2)}(\xi) + (-\frac{1}{\alpha_2} - \frac{i\xi\alpha_1}{\alpha_2})u_Z^{(1)}(\xi) - \frac{\xi\alpha_0}{\alpha_2}u_Z(\xi) = 0, \quad \forall \xi > 0. \tag{10}$$

We seek a solution of the above differential equation that has the form

$$u(\xi) = e^{-\bar{r}\xi} \sum_{j=0}^{\infty} d_j \xi^j, \quad \forall 0 < \xi < \infty,$$

for complex constants $d_0, d_1, \cdots$. Arguing as in Step 1, we observe that for $t = -\xi < 0$,

$$\begin{aligned} u(\xi) &= d_0 e^{-\bar{r}|t|} \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \prod_{k=1}^{j} \left\{ \frac{(k-1)\Delta - \bar{r}}{\alpha_2(k-1) - 1} \right\} \\ &= d_0 e^{-\bar{r}|t|} {}_1F_1(-\frac{\bar{r}}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2}). \end{aligned} \tag{11}$$

Since a solution of (7) is continuous at $t = 0$, we have $c_0 = d_0$. Thus we conclude from (9) and (11) that a solution of (7) is

$$\psi(t) = e^{-r|t|}{}_1F_1(-\frac{r}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2})\mathcal{I}\{t \geq 0\} + e^{-\bar{r}|t|}{}_1F_1(-\frac{\bar{r}}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2})\mathcal{I}\{t < 0\}. \tag{12}$$

Step 3. Suppose that $t > 0$. We seek a solution of (7) that has the form

$$\tilde{\psi}(t) = e^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j}, \qquad \forall 0 < t < \infty,$$

for complex constants $c_0, c_1, \cdots$. Observing that

$$\begin{aligned}
\tilde{\psi}^{(1)}(t) &= -re^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j} + e^{-rt}\sum_{j=0}^{\infty}(\nu+j)c_j t^{\nu+j-1}, \\
\tilde{\psi}^{(2)}(t) &= r^2e^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j} - 2re^{-rt}\sum_{j=0}^{\infty}(\nu+j)c_j t^{\nu+j-1} \\
&\quad + e^{-rt}\sum_{j=0}^{\infty}(\nu+j)(\nu+j-1)c_j t^{\nu+j-2},
\end{aligned}$$

and substituting these expressions into the left hand side of (7), we have

$$\begin{aligned}
&r^2e^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j+1} - 2re^{-rt}\sum_{j=0}^{\infty}(\nu+j)c_j t^{\nu+j} + e^{-rt}\sum_{j=0}^{\infty}(\nu+j)(\nu+j-1)c_j t^{\nu+j-1} \\
&\quad - \frac{1}{\alpha_2}[-re^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j} + e^{-rt}\sum_{j=0}^{\infty}(\nu+j)c_j t^{\nu+j-1}] \\
&\quad + \frac{i\alpha_1}{\alpha_2}[-re^{-rt}\sum_{j=0}^{\infty} c_j t^{\nu+j+1} + e^{-rt}\sum_{j=0}^{\infty}(\nu+j)c_j t^{\nu+j}] - \frac{\alpha_0 e^{-rt}}{\alpha_1}\sum_{j=0}^{\infty} c_j t^{\nu+j+1} \\
&= 0, \quad \forall 0 < t < \infty.
\end{aligned} \tag{13}$$

Equating the coefficient of t^ν in (13) to zero, we have

$$-2r\nu c_0 + (\nu+1)\nu c_1 + \frac{rc_0}{\alpha_2} - \frac{(\nu+1)c_1}{\alpha_2} + \frac{i\alpha_1\nu c_0}{\alpha_2} = 0,$$

and equating the coefficient of $t^{\nu+j-1}$, $j = 2, 3, \cdots$, in (13) to zero, we have

$$\begin{aligned}
&r^2c_{j-2} - 2r(\nu+j-1)c_{j-1} + (\nu+j)(\nu+j-1)c_j + \frac{rc_{j-1}}{\alpha_2} - \frac{(\nu+j)c_j}{\alpha_2} \\
&+\frac{i\alpha_1}{\alpha_2}[-rc_{j-2} + (\nu+j-1)c_{j-1}] - \frac{\alpha_0 c_{j-2}}{\alpha_2} = 0.
\end{aligned}$$

This gives

$$c_1 = \frac{2r\alpha_2 + r - i\alpha_1\nu}{\alpha_2(1+\nu)}c_0,$$

and in general for $j = 2, 3, \cdots$,

$$\begin{aligned} c_j &= \frac{1}{j(\nu+j)\alpha_2}\{[2(\nu+j-1)r\alpha_2 - r - i\alpha_1(\nu+j-1)]c_{j-1} \\ &\quad -(\alpha_2 r^2 - i\alpha_1 r - \alpha_0)c_{j-2}\} \\ &= c_0 \prod_{k=1}^{j} \frac{2(\nu+k-1)r\alpha_2 - r - i\alpha_1(\nu+k-1)}{k(\nu+k)\alpha_2}. \end{aligned}$$

Hence for $t \geq 0$, we have

$$\begin{aligned} \tilde{\psi}(t) &= c_0 e^{-rt} \sum_{j=0}^{\infty} t^{\nu+j} \prod_{k=1}^{j} \frac{2(\nu+k-1)r\alpha_2 - r - i\alpha_1(\nu+k-1)}{k(\nu+k)\alpha_2} \\ &= c_0 t^{\nu} e^{-rt} {}_1F_1(\nu - \frac{r}{\Delta}; \nu+1; \frac{\Delta t}{\alpha_2}). \end{aligned} \tag{14}$$

Step 4. Suppose that $t < 0$. Writing $\xi = -t$ and $u_Z(\xi) = \psi_Z(t)$, we seek a solution of (10) that has the form

$$\tilde{u}(\xi) = e^{-\bar{r}\xi} \sum_{j=0}^{\infty} d_j \xi^{\nu+j}, \qquad \forall 0 < \xi < \infty,$$

for complex constants $d_0, d_1, \cdots$. Arguing as in Step 3, we observe that for $t = -\xi < 0$,

$$\begin{aligned} \tilde{u}(\xi) &= d_0 |t|^{\nu} e^{-\bar{r}|t|} \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \prod_{k=1}^{j} \frac{(\nu+k-1)\Delta - \bar{r}}{(\nu+k)\alpha_2} \\ &= d_0 |t|^{\nu} e^{-\bar{r}|t|} {}_1F_1(\nu - \frac{\bar{r}}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2}). \end{aligned} \tag{15}$$

Since a solution of (7) is continuous at $t = 0$, we have $c_0 = d_0$. Thus we conclude from (14) and (15) that a solution of (7) is

$$\begin{aligned} \psi(t) &= |t|^{\nu} e^{-r|t|} {}_1F_1(\nu - \frac{r}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2}) \mathcal{I}\{t \geq 0\} \\ &\quad + |t|^{\nu} e^{-\bar{r}|t|} {}_1F_1(\nu - \frac{\bar{r}}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2}) \mathcal{I}\{t < 0\}. \end{aligned} \tag{16}$$

As the solutions in (12) and (16) are independent, the general solution of (7) is given by

$$\begin{aligned} \psi(t) &= \left\{ A e^{-r|t|} {}_1F_1(-\frac{r}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2}) \right. \\ &\quad \left. + B|t|^{\nu} e^{-r|t|} {}_1F_1(\nu - \frac{r}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2}) \right\} \mathcal{I}\{t \geq 0\} \\ &\quad + \left\{ \tilde{A} e^{-\bar{r}|t|} {}_1F_1(-\frac{\bar{r}}{\Delta}; -\frac{1}{\alpha_2}; \frac{\Delta|t|}{\alpha_2}) \right. \\ &\quad \left. + \tilde{B}|t|^{\nu} e^{-\bar{r}|t|} {}_1F_1(\nu - \frac{\bar{r}}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2}) \right\} \mathcal{I}\{t < 0\}, \end{aligned}$$

where $A, \tilde{A}, B$ and $\tilde{B}$ are arbitrary constants. Consequently since $\psi_Z(0) = 1$, we have $A = \tilde{A} = 1$ and

$$\begin{aligned}\psi_Z(t) &= \left\{e^{-r|t|}{}_1F_1(-\frac{r}{\Delta}; 1-\nu; \frac{\Delta|t|}{\alpha_2}) \right.\\ &\quad \left. +B|t|^\nu e^{-r|t|}{}_1F_1(\nu - \frac{r}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2})\right\}\mathcal{I}\{t \geq 0\}\\ &\quad + \left\{e^{-\bar{r}|t|}{}_1F_1(-\frac{\bar{r}}{\Delta}; 1-\nu; \frac{\Delta|t|}{\alpha_2}) \right.\\ &\quad \left. +\tilde{B}|t|^\nu e^{-\bar{r}|t|}{}_1F_1(\nu - \frac{\bar{r}}{\Delta}; \nu+1; \frac{\Delta|t|}{\alpha_2})\right\}\mathcal{I}\{t < 0\}, \qquad (17)\end{aligned}$$

for some constants B and $\tilde{B}$.

Step 5. To complete the proof of Theorem 2, it suffices to determine the constants B and $\tilde{B}$ in (17). We observe from Slater (1960), page 60, that for $x \to \infty$,

$${}_1F_1(a; b; x) = x^{a-b}e^x\frac{\Gamma(b)}{\Gamma(a)}(1 + O(|x|^{-1}).$$

Hence it follows from (17) that as $t \to \infty$,

$$\begin{aligned}\psi_Z(t) &= e^{-rt}e^{\Delta t/\alpha_2}\left\{(\frac{\Delta t}{\alpha_2})^{\nu-1-r\Delta^{-1}}\frac{\Gamma(-\alpha_2^{-1})}{\Gamma(-r\Delta^{-1})}\right.\\ &\quad \left. +Bt^\nu(\frac{\Delta t}{\alpha_2})^{-1-r\Delta^{-1}}\frac{\Gamma(\nu+1)}{\Gamma(\nu - r\Delta^{-1})}\right\}(1 + o(1))\\ &= e^{-rt}e^{\Delta t/\alpha_2}t^\nu(\frac{\Delta t}{\alpha_2})^{-1-r\Delta^{-1}}\left\{(\frac{\Delta}{\alpha_2})^\nu\frac{\Gamma(-\alpha_2^{-1})}{\Gamma(-r\Delta^{-1})}\right.\\ &\quad \left. + B\frac{\Gamma(\nu+1)}{\Gamma(\nu - r\Delta^{-1})}\right\}(1 + o(1)).\end{aligned}$$

Since $\lim_{t\to\infty}\psi_{Z_n}(t) = 0$, we have

$$B = -(\frac{\Delta}{\alpha_2})^\nu\frac{\Gamma(-\alpha_2^{-1})\Gamma(\nu - r\Delta^{-1})}{\Gamma(\nu+1)\Gamma(-r\Delta^{-1})}. \qquad (18)$$

Similarly as $t \to -\infty$,

$$\begin{aligned}\psi_Z(t) &= e^{-\bar{r}|t|}e^{\Delta|t|/\alpha_2}\left\{(\frac{\Delta|t|}{\alpha_2})^{\nu-1-\bar{r}\Delta^{-1}}\frac{\Gamma(-\alpha_2^{-1})}{\Gamma(-\bar{r}\Delta^{-1})}\right.\\ &\quad \left. +\tilde{B}|t|^\nu(\frac{\Delta|t|}{\alpha_2})^{-1-\bar{r}\Delta^{-1}}\frac{\Gamma(\nu+1)}{\Gamma(\nu - \bar{r}\Delta^{-1})}\right\}(1 + o(1))\\ &= e^{-\bar{r}t}e^{\Delta|t|/\alpha_2}|t|^\nu(\frac{\Delta|t|}{\alpha_2})^{-1-\bar{r}\Delta^{-1}}\left\{(\frac{\Delta}{\alpha_2})^\nu\frac{\Gamma(-\alpha_2^{-1})}{\Gamma(-\bar{r}\Delta^{-1})}\right.\\ &\quad \left. +\tilde{B}\frac{\Gamma(\nu+1)}{\Gamma(\nu - \bar{r}\Delta^{-1})}\right\}(1 + o(1)).\end{aligned}$$

Since $\lim_{t\to-\infty}\psi_Z(t) = 0$, we have

$$\tilde{B} = -(\frac{\Delta}{\alpha_2})^\nu\frac{\Gamma(-\alpha_2^{-1})\Gamma(\nu - \bar{r}\Delta^{-1})}{\Gamma(\nu+1)\Gamma(-\bar{r}\Delta^{-1})}. \qquad (19)$$

Theorem 2 now follows from (17), (18), (19), the definition of $U(.;.;.)$ and Euler's reflection formula, namely

$$\Gamma(x)\Gamma(1-x) = \frac{x}{\sin(\pi x)},$$

[see, for example, Theorem 1.2.1 of Andrews, Askey and Roy (1999)]. □

Acknowledgments

I would like to thank Professor Anirban DasGupta for his suggestions and comments on this article.

References

[1] Andrews, G. E., Askey, R. and Roy, R. (1999). *Special Functions.* Cambridge Univ. Press, Cambridge.

[2] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edition. Wiley, New York.

[3] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd edition. Wiley, New York. MR1299979

[4] Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variations in homogeneous material. *Phil. Trans. Roy. Soc. Lond. Ser. A* **186** 343-414.

[5] Shorack, G. R. (2000). *Probability for Statisticians.* Springer, New York.

[6] Slater, L. J. (1960). *Confluent Hypergeometric Functions.* Cambridge Univ. Press, Cambridge.

[7] Stein, C. M. (1986). *Approximate Computation of Expectations.* IMS Lecture Notes-Monograph Series, Vol. 7. Hayward, California.

[8] Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 1, 6th edition. Edward Arnold, London.

[9] Wolfram, S. (1996). *The Mathematica Book*, 3rd edition. Cambridge Univ. Press, New York.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 180–206
© Institute of Mathematical Statistics, 2004

Characterizations, Sub and resampling, and goodness of fit

L. Brown[1], Anirban DasGupta[2], John Marden[3] and Dimitris Politis[4]

University of Pennsylvania, Purdue University, University of Illinois at Urbana-Champaign, University of California, San Diego

Abstract: We present a general proposal for testing for goodness of fit, based on resampling and subsampling methods, and illustrate it with graphical and analytical tests for the problems of testing for univariate or multivariate normality. The proposal shows promising, and in some cases dramatic, success in detecting nonnormality. Compared to common competitors, such as a Q-Q plot or a likelihood ratio test against a specified alternative, our proposal seems to be the most useful when the sample size is small, such as 10 or 12, or even very small, such as 6! We also show how our proposal provides tangible information about the nature of the true cdf from which one is sampling. Thus, our proposal also has data analytic value. Although only the normality problem is addressed here, the scope of application of the general proposal should be much broader.

1. Introduction

The purpose of this article is to present a general proposal, based on re or subsampling, for goodness of fit tests and apply it to the problem of testing for univariate or multivariate normality of iid data. Based on the evidence we have accumulated, the proposal seems to have unexpected success. It comes out especially well, relative to its common competitors, when the sample size is small, or even very small. The common tests, graphical or analytical, do not have much credibility for very small sample sizes. For example, a Q-Q plot with a sample of size 6 would be hardly credible; neither would be an analytical test, such as the Shapiro-Wilk, the Anderson-Darling or the Kolmogorov-Smirnov test with estimated parameters (Shapiro and Wilk (1965), Anderson and Darling (1952,1954), Stephens (1976), Babu and Rao (2004)). But, somewhat mysteriously, the tests based on our proposal seem to have impressive detection power even with such small sample sizes. Furthermore, the proposal is general, and so its scope of application is broader than just the normality problem. However, in this article, we choose to investigate only the normality problem in detail, it being the obvious first application one would want to try. Although we have not conducted a complete technical analysis, we still hope that we have presented here a useful set of ideas with broad applicability.

The basic idea is to use a suitably chosen characterization result for the null hypothesis and combine it with the bootstrap or subsampling to produce a goodness

[1]Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Hunstman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA. e-mail: lbrown@wharton.upenn.edu

[2]Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2068, USA. e-mail: dasgupta@stat.purdue.edu

[3]Department of Statistics, University of Illinois at Urbana-Champaign, 116B Illini Hall, 725 S. Wright St., Champaign, IL 61820, USA. e-mail: jimarden@uiuc.edu

[4]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA. e-mail: dpolitis@ucsd.edu

Keywords and phrases: bootstrap, characterization, consistency, goodness of fit, normal, multivariate normal, power, Q-Q plot, scatterplot, subsampling.

AMS 2000 subject classifications: 62G09, 62E10.

of fit test. The idea has been mentioned previously. But it has not been investigated in the way or at length, as we do it here (see McDonald and Katti (1974), Mudholkar, McDermott and Srivastava (1992), Mudholkar, Marchetti and Lin (2002) and D'Agostino and Stephens (1986)). To illustrate the basic idea, it is well known that if $X_1, X_2, \ldots, X_n$ are iid samples from some cdf F on the real line with a finite variance, then F is a normal distribution if and only if the sample mean $\overline{X}$ and the sample variance s^2 are independent, and distributed respectively, as a normal and a (scaled) chisquare. Therefore, using standard notation, with G_m denoting the cdf of a chisquare distribution with m degrees of freedom, the random variables $U_n = \Phi(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma})$ and $V_n = G_{n-1}(\frac{(n-1)s^2}{\sigma^2})$ would be independent $U[0,1]$ random variables. Proxies of U_n, V_n can be computed, in the usual way, by using either a resample (such as the ordinary bootstrap), or a subsample, with some subsample size b. These proxies, namely the pairs, $w_i^* = (U_i^*, V_i^*)$ can then be plotted in the unit square to visually assess evidence of any structured or patterned deviation from a random uniform like scattering. They can also be used to construct formal tests, in addition to graphical tests. The use of the univariate normality problem, and of $\overline{X}$ and s^2 are both artifacts. Other statistics can be used, and in fact we do so (interquartile range/s and s, for instance). We also investigate the multivariate normality problem, which remains to date, a notoriously difficult problem, especially for small sample sizes, the case we most emphasize in this article.

We begin with a quantification of the statistical folklore that Q-Q plots tend to look linear in the central part of the plot for many types of nonnormal data. We present these results on the Q-Q plot for two main reasons. The precise quantifications we give would be surprising to many people; in addition, these results provide a background for why complementary graphical tests, such as the ones we offer, can be useful.

The resampling based graphical tests are presented and analyzed next. A charming property of our resampling based test is that it does not stop at simply detecting nonnormality. It gives substantially more information about the nature of the true cdf from which one is sampling, if it is not a normal cdf. We show how a skillful analysis of the graphical test would produce such useful information by looking at key features of the plots, for instance, empty corners, or a pronounced trend. In this sense, our proposal also has the flavor of being a useful data analytic tool.

Subsampling based tests are presented at the end. But we do not analyze them with as much detail as the resampling based tests. The main reason is limitation of space. But comparison of the resampling based tests and the test based on subsampling reveals quite interesting phenomena. For example, when a structured deviation from a uniform like scattering is seen, the structures are *different* for the re and subsampling based tests. Thus, we seem to have the situation that we do not need to necessarily choose one or the other. The resampling and subsampling based tests complement each other. They can both be used, as alternatives or complements, to common tests, and especially when the sample sizes are small, or even very small.

To summarize, the principal contributions and the salient features of this article are the following:

1. We suggest a flexible general proposal for testing goodness of fit to parametric families based on characterizations of the family;

2. We illustrate the method for the problems of testing univariate and multivariate normality;

3. The method is based on re or subsampling, and tests based on the two methods nicely complement each other;

4. Graphical tests form the core of our proposal, and they are especially useful for small sample sizes due to lack of credible graphical tests when the sample size is small;

5. We give companion formal tests to our graphical tests with some power studies; but the graphical test is more effective in our assessment;

6. We provide a theoretical background for why new graphical tests should be welcome in the area by providing some precise quantifications for just how misleading Q-Q plots can be. The exact results should be surprising to many.

7. We indicate scope of additional applications by discussing three interesting problems.

2. Why Q-Q plots can mislead

The principal contribution of our article is a proposal for new resampling based graphical tests for goodness of fit. Since Q-Q plots are of wide and universal use for that purpose, it would be helpful to explain why we think that alternative graphical tests would be useful, and perhaps even needed. Towards this end, we first provide a few technical results and some numerics to illustrate how Q-Q plots can be misleading. It has been part of the general knowledge and folklore that Q-Q plots can be misleading; but the results below give some precise explanation for and quantification of such misleading behavior of Q-Q plots.

Q-Q plots can mislead because of two reasons. They look approximately linear in the central part for many types of nonnormal data, and because of the common standard we apply to ourselves (and teach students) that we should not overreact to wiggles in the Q-Q plot and what counts is an overall visual impression of linearity. The following results explain why that standard is a dangerous one. First some notation is introduced.

The exact definition of the Q-Q plot varies a little from source to source. For the numerical illustrations, we will define a Q-Q plot as a plot of the pairs $(z_{(i-1/2)/n}, X_{(i)})$, where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the standard normal distribution and $X_{(i)}$ is the ith sample order statistic (at other places, $z_{(i-1/2)/n}$ is replaced by $z_{(i+1/2)/(n+1)}, z_{(i+1/2)/(n+3/4)}$, etc. Due to the asymptotic nature of our results, these distinctions do not affect the statements of the results). For notational simplicity, we will simply write z_i for $z_{(i-1/2)/n}$. The natural index for visual linearity of the Q-Q plot is the coefficient of correlation

$$r_n = \frac{\sum_{i=1}^n z_i(X_{(i)} - \overline{X})}{\sqrt{\sum_{i=1}^n z_i^2 \sum_{i=1}^n (X_{(i)} - \overline{X})^2}} = \frac{\sum_{i=1}^n z_i X_{(i)}}{\sqrt{\sum_{i=1}^n z_i^2 \sum_{i=1}^n (X_{(i)} - \overline{X})^2}}.$$

As we mentioned above, the central part of a Q-Q plot tends to look approximately linear for many types of nonnormal data. This necessitates another index for linearity of the central part in a Q-Q plot. Thus, for $0 < \alpha < 0.5$, we define the *trimmed correlation*

$$r_\alpha = r_{n,\alpha} = \frac{\sum_{i=k+1}^{n-k} z_i X_{(i)}}{\sqrt{\sum_{i=k+1}^{n-k} z_i^2 \sum_{i=k+1}^{n-k} (X_{(i)} - \overline{X_k})^2}},$$

where $k = [n\alpha]$, and $\overline{X_k}$ is the corresponding trimmed mean. In other words, r_α is the correlation in the Q-Q plot when $100\alpha\%$ of the points are deleted from each tail of the plot. r_α typically is larger in magnitude than r_n, as we shall see below.

We will assume that the true underlying CDF F is continuous, although a number of our results do not require that assumption.

2.1. Almost sure limits of r_n and r_α

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be iid observations from a CDF F with finite variance σ^2. Then*

$$r_n \to \rho(F) = \frac{\int_0^1 F^{-1}(x)\Phi^{-1}(x)\,dx}{\sigma}$$

with probability 1.

Proof. Multiply the numerator as well as each term within the square-root sign in the denominator by n. The term $\frac{1}{n}\sum_{i=1}^n z_i^2$ converges to $\int_0^1(\Phi^{-1}(x))^2\,dx$, being a Riemann sum for that integral. The second term $\frac{1}{n}\sum_{i=1}^n(X_{(i)} - \overline{X})^2$ converges a.s. to σ^2 by the usual strong law. Since $\int_0^1(\Phi^{-1}(x))^2\,dx = 1$, on division by n, the denominator in r_n converges a.s. to σ.

The numerator needs a little work. Using the same notation as in Serfling (1980) (pp. 277–279), define the double sequence $t_{ni} = (i - 1/2)/n$ and $J(t) = \Phi^{-1}(t)$. Thus J is everywhere continuous and satisfies for every $r > 0$ and in particular for $r = 2$, the growth condition $|J(t)| \le M[t(1-t)]^{1/r-1+\delta}$ for some $\delta > 0$. Trivially, $\max_{1\le i\le n}|t_{ni} - i/n| \to 0$. Finally, there exists a positive constant a such that $a.\min_{1\le i\le n}\{i/n, 1 - i/n\} \le t_{ni} \le 1 - a.\min_{1\le i\le n}\{i/n, 1 - i/n\}$. Specifically, this holds with $a = 1/2$. It follows from Example A and Example A^* in pp. 277–279 in Serfling (1980) that on division by n, the numerator of r_n converges a.s. to $\int_0^1 F^{-1}(x)\Phi^{-1}(x)\,dx$, establishing the statement of Theorem 1.

The almost sure limit of the truncated correlation r_α is stated next; we omit its proof as it is very similar to the proof of Theorem 1. □

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ be iid observations from a CDF F. Let $0 < \alpha < 0.5$, and*

$$\mu_\alpha = \frac{\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x\,dF(x)}{1 - 2\alpha}.$$

Then, with probability 1,

$$r_\alpha \to \rho_\alpha(F) = \frac{\int_\alpha^{1-\alpha} F^{-1}(x)\Phi^{-1}(x)\,dx}{\sqrt{\int_\alpha^{1-\alpha}(\Phi^{-1}(x))^2\,dx \cdot \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}(x - \mu_\alpha)^2\,dF(x)}}.$$

Theorem 1 and 2 are used in the following Table to explain why Q-Q plots show an overall visual linearity for many types of nonnormal data, and especially so in the central part of the plot.

Discussion of Table 1

We see from Table 1 that for each distribution that we tried, the trimmed correlation is larger than the untrimmed one. We also see that as little as 5% trimming from each tail produces a correlation at least as large as .95, even for the extremely skewed Exponential case. For symmetric populations, 5% trimming produces a nearly perfectly linear Q-Q plot, asymptotically. Theorem 1, Theorem 2,

Table 1: Limiting correlation in Q-Q plots.

F	No Trimming	5% trimming
Uniform	.9772	.9949
Double Exp.	.9811	.9941
Logistic	.9663	.9995
t(3)	.9008	.9984
t(5)	.9832	.9991
Tukey distribution	.9706	.9997
		(defined as .9N(0,1) + .1N(0,9))
chisquare(5)	.9577	.9826
Exponential	.9032	.9536

and Table 1 vindicate our common empirical experience that the central part of a Q-Q plot is very likely to look linear for all types of data: light tailed, medium tailed, heavy tailed, symmetric, skewed. Information about nonnormality from a Q-Q plot can only come from the tails and the somewhat pervasive practice of concentrating on the overall linearity and ignoring the wiggles at the tails renders the Q-Q plot substantially useless in detecting nonnormality. Certainly we are not suggesting, and it is not true, that everyone uses the Q-Q plot by concentrating on the central part. Still, these results suggest that alternative or complementary graphical tests can be useful, especially for small sample sizes. A part of our efforts in the rest of this article address that.

3. Resampling based tests for univariate normality

3.1. Test based on $\overline{X}$ and s^2

Let $X_1, X_2, \ldots, X_n$ be iid observations from a $N(\mu, \sigma^2)$ distribution. A well known characterization of the family of normal distributions is that the sample mean $\overline{X}$ and the sample variance s^2 are independently distributed (see Kagan,Linnik and Rao (1973); a good generalization is Parthasarathy (1976). The generalizations due to him can be used for other resampling based tests of normality). If one can test their independence using the sample data, it would in principle provide a means of testing for the normality of the underlying population. But of course to test the independence, we will have to have some idea of the joint distribution of $\overline{X}$ and s^2, and this cannot be done using just one set of sample observations in the standard statistical paradigm. Here is where resampling can be useful.

Thus, for some $B > 1$, let $X^*_{i1}, X^*_{i2}, \ldots, X^*_{in}, i = 1, 2, \ldots, B$ be a sample from the empirical CDF of the original sample values $X_1, X_2, \ldots, X_n$. Define,

$$\overline{X}^*_i = \frac{1}{n}\sum_{j=1}^{n} X^*_{ij}, \qquad \text{and} \qquad s^{2*}_i = \frac{1}{n-1}\sum_{j=1}^{n}\left(X^*_{ij} - \overline{X}^*_i\right)^2.$$

Let Φ denote the standard normal CDF and G_m the CDF of the *chisquare* distribution with m degrees of freedom. Under the null hypothesis of normality, the statistics

$$U_n = \Phi\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right) \qquad \text{and} \qquad V_n = G_{n-1}\left(\frac{(n-1)s^2}{\sigma^2}\right)$$

are independently distributed as $U[0, 1]$.

Motivated by this, define: for $i = 1, 2, \ldots, B$,

$$u^*_i = \Phi\left(\frac{\sqrt{n}(\overline{X}^*_i - \overline{X})}{s}\right) \qquad \text{and} \qquad v^*_i = G_{n-1}\left(\frac{(n-1)s^{2*}_i}{s^2}\right).$$

Let $w_i^* = (u_i^*, v_i^*), i = 1, 2, \ldots, B$. If the null hypothesis is true, the w_i^* should be roughly uniformly scattered in the unit square $[0,1] \times [0,1]$. This is the graphical test we propose in this section. A subsampling based test using the same idea will be described in a subsequent section. We will present evidence that this resampling based graphical test is quite effective, and relatively speaking, is more useful for *small* sample sizes. This is because for small n, it is hard to think of other procedures that will have much credibility. For example, if $n = 6$, a case that we present here, it is not very credible to draw a Q-Q plot. Our resampling based test would be more credible for such small sample sizes.

The following consistency theorem shows that our method will correctly identify the joint distribution of (U_n, V_n), asymptotically. Although we use the test in small samples, the consistency theorem still provides some necessary theoretical foundation for our method.

Theorem 3. *Using standard notation,*

$$\sup_{0 \le u \le 1, 0 \le v \le 1} \left| P_*\big(U^* \le u, V^* \le v\big) - P_F\big(U_n \le u, V_n \le v\big) \right| \to 0$$

in probability, provided F has four moments, where F denotes the true CDF from which $X_1, X_2, \ldots, X_n$ are iid observations.

Proof. We observe that the ordinary bootstrap is consistent for the joint distribution of $(\overline{X}, s^2)$ if F has four moments. Theorem 3 follows from this and the uniform *delta* theorem for the bootstrap (see van der Vaart (1998)). □

Under the null hypothesis, (U_n, V_n) are uniformly distributed in the unit square for each n, and hence also asymptotically. We next describe the joint asymptotic distribution of (U_n, V_n) under a general F with four moments. It will follow that our test is not consistent against a specific alternative F if and only if F has the same first four moments as some $N(\mu, \sigma^2)$ distribution. From the point of view of common statistical practice, this is not a major drawback. To have a test consistent against all alternatives, we will have to use more than $\overline{X}$ and s^2.

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ be iid observations from a CDF F with four finite moments. Let μ_3, μ_4 denote the third and the fourth central moment of F, and $\kappa = \frac{\mu_4}{\sigma^4}$. Then,*

$(U_n, V_n) \Rightarrow H$, where H has the density

$$\begin{aligned} h(u,v) \;=\; & \sqrt{\frac{2}{\kappa - 1}} \frac{1}{\sqrt{1 - \frac{\mu_3^2}{(\kappa-1)\sigma^6}}} \exp\Big\{ -\frac{1}{2(\mu_3^2 - (\kappa-1)\sigma^6)} \\ & \times \Big[2\sqrt{2}\mu_3 \sigma^3 \Phi^{-1}(u)\Phi^{-1}(v) + (\kappa - 3)\sigma^6 \big(\Phi^{-1}(v)\big)^2 \\ & - \mu_3^2 \big(\big(\Phi^{-1}(u)\big)^2 + \big(\Phi^{-1}(v)\big)^2 \big) \Big] \Big\}. \end{aligned} \tag{1}$$

Proof. Let

$$Z_{1n} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}, \qquad Z_{2n} = \frac{\sqrt{n}(s^2 - \sigma^2)}{\sqrt{\mu_4 - \sigma^4}}.$$

Then, it is well known that $(Z_{1n}, Z_{2n}) \Rightarrow (Z_1, Z_2) \sim N(0, 0, \Sigma)$, where $\Sigma = ((\sigma_{ij}))$, with $\sigma_{11} = 1, \sigma_{12} = \frac{\mu_3}{\sigma^3\sqrt{\kappa - 1}}$, and $\sigma_{22} = 1$.

Hence, from the definitions of U_n, V_n, it follows that we only need the joint asymptotic distribution of $(\Phi(Z_{1n}), \Phi(\sqrt{\frac{\kappa-1}{2}}Z_{2n}))$.By the continuity theorem for weak convergence, therefore, $(U_n, V_n) \Rightarrow (\Phi(Z_1), \Phi(\sqrt{\frac{\kappa-1}{2}}Z_2))$. Thus, we need to derive the joint density of $(\Phi(Z_1), \Phi(\sqrt{\frac{\kappa-1}{2}}Z_2))$, which will be our $h(u, v)$.

Let $f(x, y)$ denote the bivariate normal density of (Z_1, Z_2), i.e., let

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)}.$$

Then,

$$\begin{aligned} H(u,v) &= P\left(\Phi(Z_1) \le u, \Phi\left(\sqrt{\frac{\kappa-1}{2}}Z_2\right) \le v\right) \\ &= P\left(Z_1 \le \Phi^{-1}(u), Z_2 \le \sqrt{\frac{2}{\kappa-1}}\Phi^{-1}(v)\right) \\ &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\sqrt{\frac{2}{\kappa-1}}\Phi^{-1}(v)} f(x,y)\, dy\, dx. \end{aligned}$$

The joint density $h(u, v)$ is obtained by obtaining the mixed partial derivative $\frac{\partial^2}{\partial v \partial u} H(u, v)$. Direct differentiation using the chain rule gives

$$h(u,v) = \sqrt{\frac{2}{\kappa-1}} \frac{1}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))} f\left(\Phi^{-1}(u), \sqrt{\frac{2}{\kappa-1}}\Phi^{-1}(v)\right),$$

on some algebra.

From here, the stated formula for $h(u, v)$ follows on some further algebra, which we omit. □

3.2. Learning from the plots

It is clear from the expression for $h(u, v)$ that if the third central moment μ_3 is zero, then U, V are independent; moreover, U is marginally uniform. Thus, intuitively, we may expect that our proposal would have less success for distinguishing normal data from other symmetric data, and more success in detecting nonnormality when the population is skewed. This is in fact true, as we shall later see in our simulations of the test. It would be useful to see the plots of the density $h(u, v)$ for some trial nonnormal distributions, and try to synchronize them with actual simulations of the bootstrapped pairs w_i^*. Such a synchronization would help us learn something about the nature of the true population as opposed to just concluding nonnormality. In this, we have had reasonable success, as we shall again see in our simulations. We remark that this is one reason that knowing the formula in Theorem 4 for the asymptotic density $h(u, v)$ is useful; other uses of knowing the asymptotic density are discussed below.

It is informative to look at a few other summary quantities of the asymptotic density $h(u, v)$ that we can try to synchronize with our plots of the w_i^*. We have in mind summaries that would indicate if we are likely to see an upward or downward trend in the plot under a given specific F, and if we might expect noticeable departures from a uniform scattering such as empty corners. The next two results shed some light on those questions.

Theorem 5. *Let $(U, V) \sim h(u, v)$. Then, $\rho := Corr(U, V)$ has the following values for the corresponding choices of F:*

$\rho \approx .69$ *if* $F =$ Exponential;
$\rho \approx .56$ *if* $F =$ Chisquare(5);
$\rho \approx .44$ *if* $F =$ Beta$(2, 6)$;
$\rho \approx .50$ *if* $F =$ Beta$(2, 10)$;
$\rho \approx .53$ *if* $F =$ Poisson(1);
$\rho \approx .28$ *if* $F =$ Poisson(5).

The values of ρ stated above follow by using the formula for $h(u, v)$ and doing the requisite expectation calculations by a two dimensional numerical integration.

A discussion of the utility of knowing the asymptotic correlations will follow the next theorem.

Theorem 6. *Let* $p_{11} = P(U \le .2, V \le .2)$, $p_{12} = P(U \le .2, V \ge .8)$, $p_{13} = P(U \ge .8, V \le .2)$ *and* $p_{14} = P(U \ge .8, V \ge .8)$.

Then, $p_{11} = p_{12} = p_{13} = p_{14} = .04$ *if* $F =$ Normal;
$p_{11} = .024$, $p_{12} = .064$, $p_{13} = .0255$, $p_{14} = .068$ *if* $F =$ Double Exponential;
$p_{11} = .023$, $p_{12} = .067$, $p_{13} = .024$, $p_{14} = .071$ *if* $F =$ t(5);
$p_{11} = .01$, $p_{12} = .02$, $p_{13} = .01$, $p_{14} = .02$ *if* $F =$ Uniform;
$p_{11} = .04$, $p_{12} = .008$, $p_{13} = .004$, $p_{14} = .148$ *if* $F =$ Exponential;
$p_{11} = .04$, $p_{12} = .012$, $p_{13} = .006$, $p_{14} = .097$ *if* $F =$ Beta$(2, 6)$;
$p_{11} = .045$, $p_{12} = .01$, $p_{13} = .005$, $p_{14} = .117$ *if* $F =$ Beta$(2, 10)$.

Proof. Again, the values stated in the Theorem are obtained by using the formula for $h(u, v)$ and doing the required numerical integrations. □

3.3. Synchronization of theorems and plots

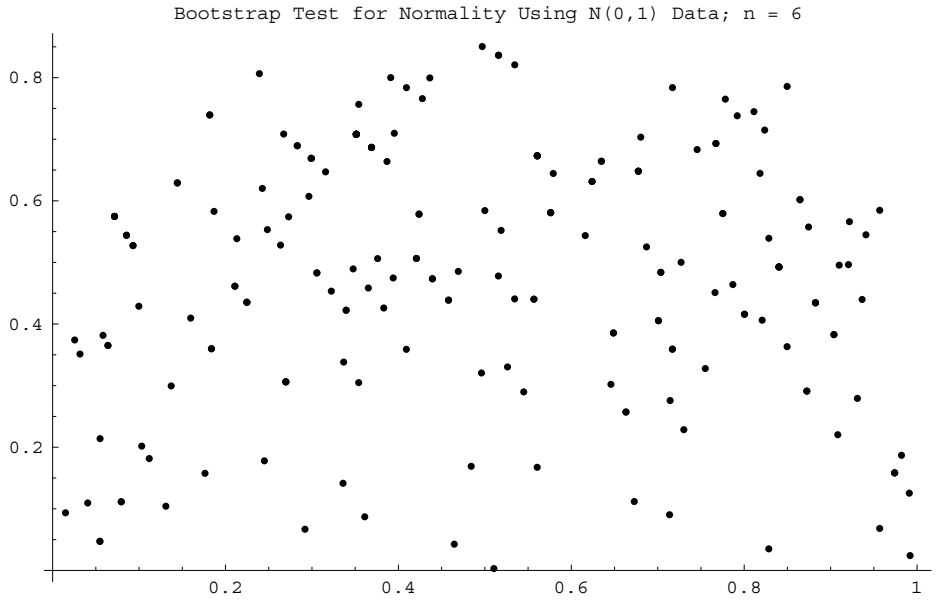
Together, Theorem 5 and Theorem 6 have the potential of giving useful information about the nature of the true CDF F from which one is sampling, by inspecting the cloud of the w_i^* and comparing certain features of the cloud with the general pattern of the numbers quoted in Theorems 5 and 6. Here are some main points.

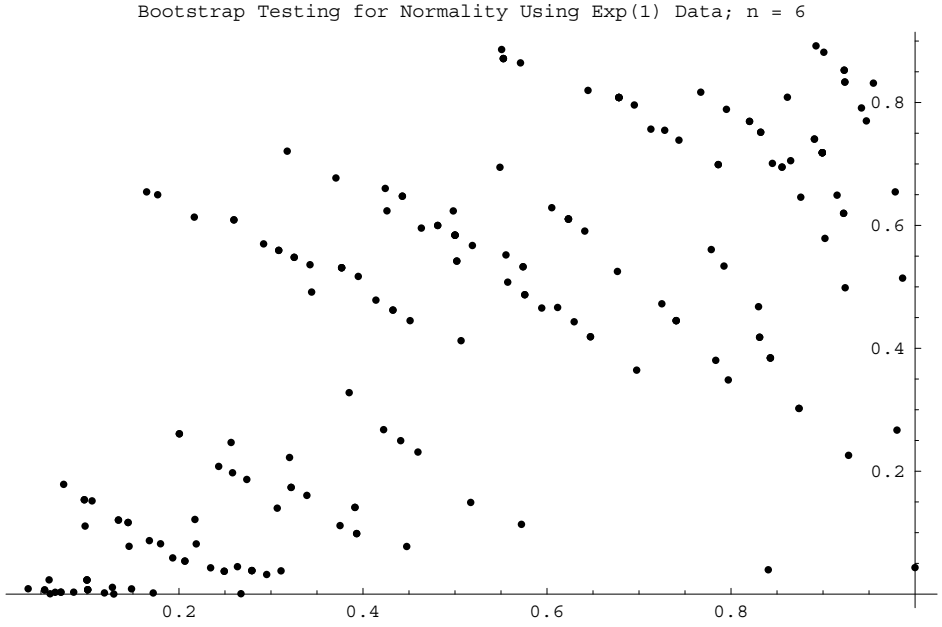
1. A pronounced upward trend in the w_i^* cloud would indicate a right skewed population (such as Exponential or a small degree of freedom chisquare or a right skewed Beta, etc.), while a mild upward trend may be indicative of a population slightly right skewed, such as a Poisson with a moderately large mean.

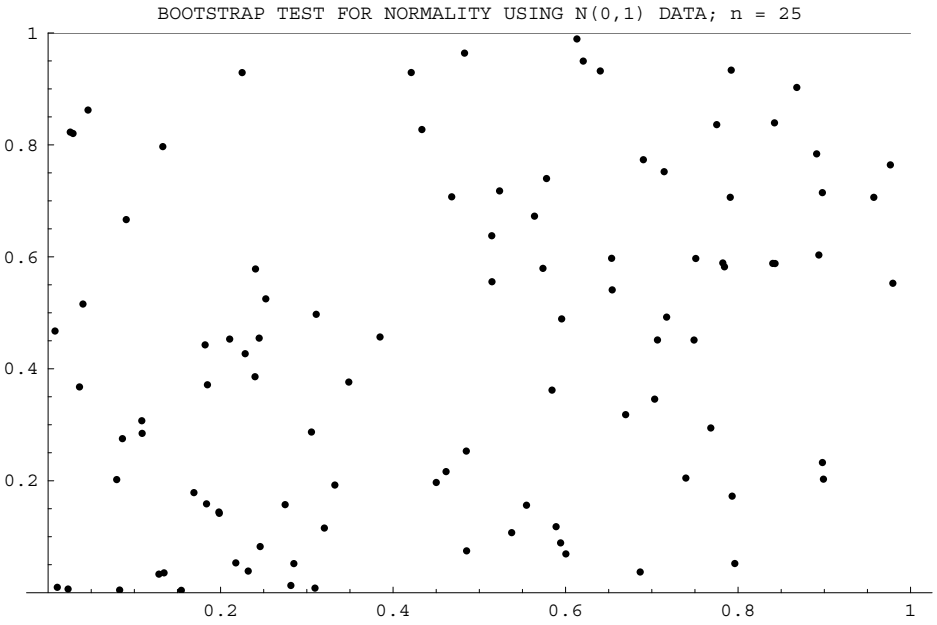
2. To make a finer distinction, Theorem 6 can be useful. $p_{11}, p_{12}, p_{13}, p_{14}$ respectively measure the density of the points in the lower left, upper left, lower right, and the upper right corner of the w_i^* cloud. From Theorem 6 we learn that for right skewed populations, the upper left and the lower right corners should be rather empty, while the upper right corner should be relatively much more crowded. This is rather interesting, and consistent with the correlation information provided by Theorem 5 too.

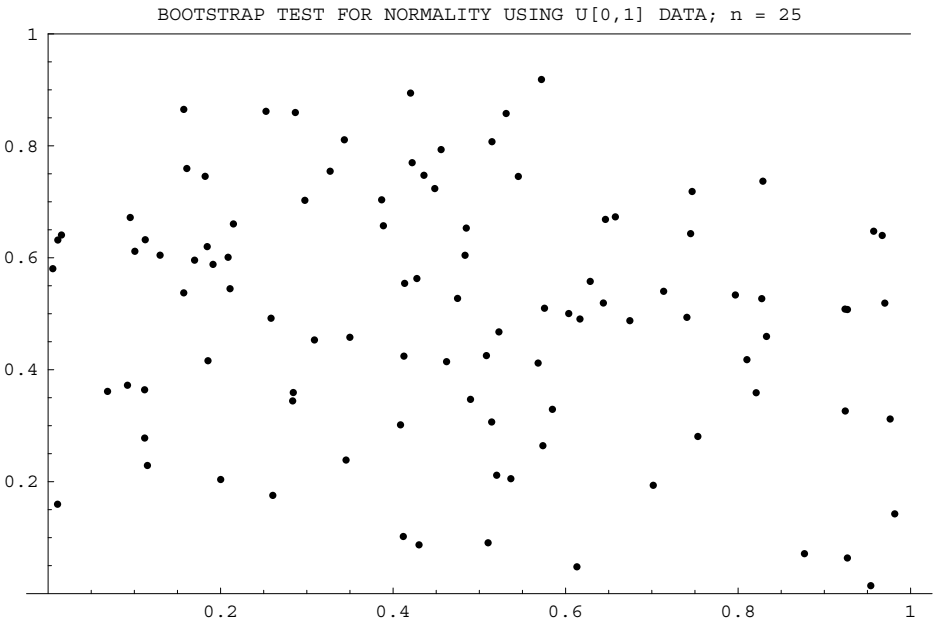
3. In contrast, for symmetric heavy tailed populations, the two upper corners should be relatively more crowded compared to the two lower corners, as we can see from the numbers obtained in Theorem 6 for Double Exponential and t(5) distributions. For uniform data, all four corners should be about equally dense, with a general sparsity of points in all four corners. In our opinion, these conclusions that one can draw from Theorems 5 and 6 together about the nature of the true CDF are potentially quite useful.

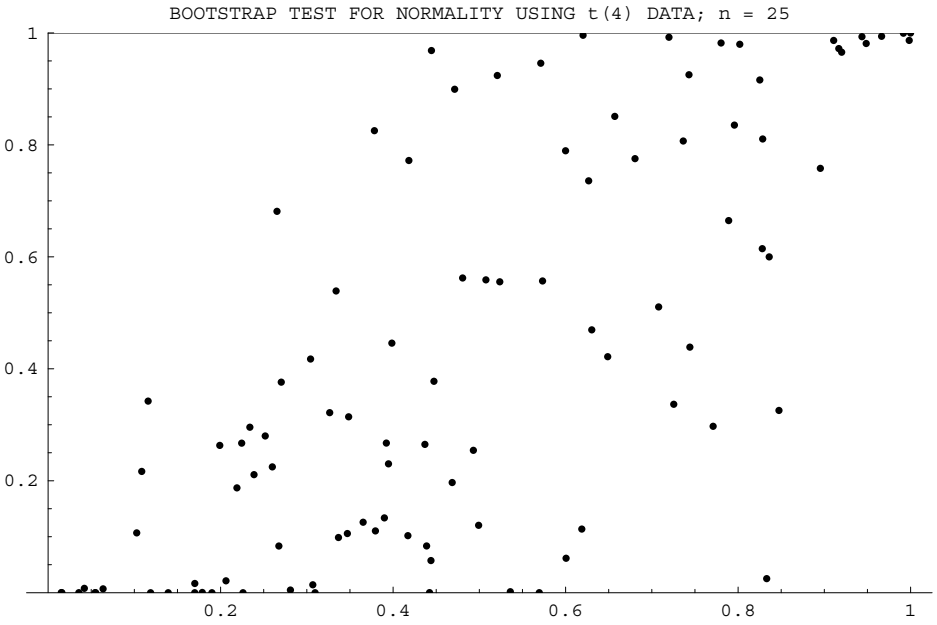
We next present a selection of scatterplots corresponding to our test above. Due to reasons of space, we are unable to present all the plots we have. The plots we present characterize what we saw in our plots typically; the resample size B varies

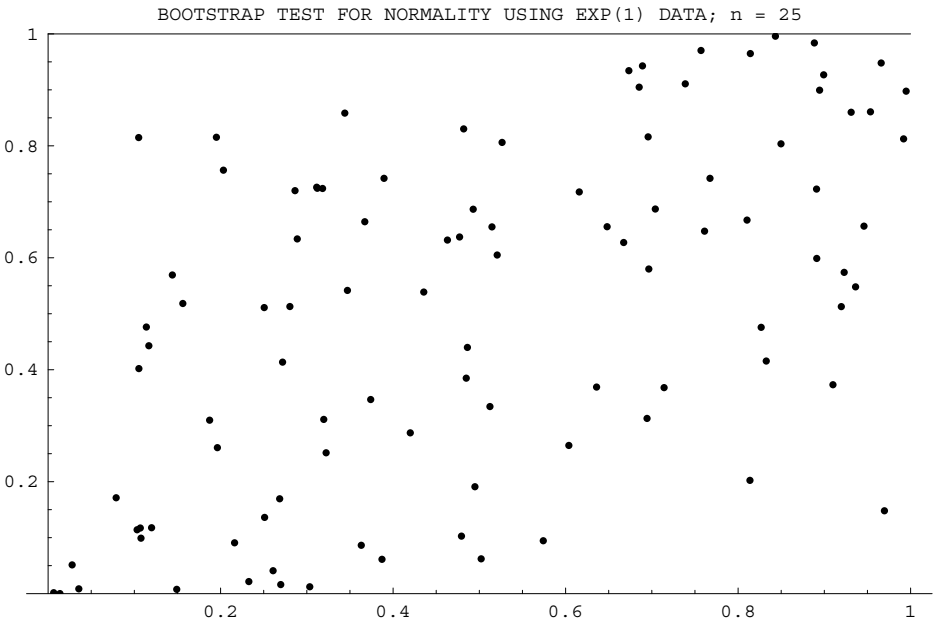

Bootstrap Test for Normality Using N(0,1) Data; n = 6
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Bootstrap Testing for Normality Using Exp(1) Data; n = 6
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8


BOOTSTRAP TEST FOR NORMALITY USING N(0,1) DATA; n = 25
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


BOOTSTRAP TEST FOR NORMALITY USING U[0,1] DATA; n = 25
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


BOOTSTRAP TEST FOR NORMALITY USING t(4) DATA; n = 25
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


BOOTSTRAP TEST FOR NORMALITY USING EXP(1) DATA; n = 25
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1

between 100 and 200 in the plots. The main conclusions we draw from our plots are summarized in the following discussion.

The most dramatic aspect of these plots is the transparent structure in the plots for the right skewed Exponential case for the extremely small sample size of $n = 6$. We also see satisfactory agreement as regards the density of points at the corners with the statements in Theorem 6. Note the relatively empty upper left and lower right corners in the Exponential plot, as Theorem 6 predicts, and the general sparsity of points in all the corners in the uniform case, also as Theorem 6 predicts. The plot for the t case shows mixed success; the very empty upper left corner is not predicted by Theorem 6. However, the plot itself looks very nonuniform in the unit square, and in that sense the t(4) plot can be regarded as a success. To summarize, certain predictions of Theorems 5 and 6 manifest reasonably in these plots, which is reassuring.

The three dimensional plots of the asymptotic density function $h(u, v)$ are also presented next for the uniform, t(5), and the Exponential case, for completeness and better understanding.

3.4. Comparative power and a formal test

While graphical tests have a simple appeal and are preferred by some, a formal test is more objective. We will offer some in this subsection; however, for the kinds of small sample sizes we are emphasizing, the chi-square approximation is not good. The correct percentiles needed for an accurate application of the formal test would require numerical evaluation. In the power table reported below, that was done.

The formal test

The test is a standard chisquare test. Partition the unit square into subrectangles $[a_i, b_j]$, where $a_i = b_i = .2i$, and let in a collection of B points, O_{ij} be the observed number of pairs w^* in the subrectangle $[a_i, b_j]$. The expected number of points in each subrectangle is $.04B$. Thus, the test is as follows:

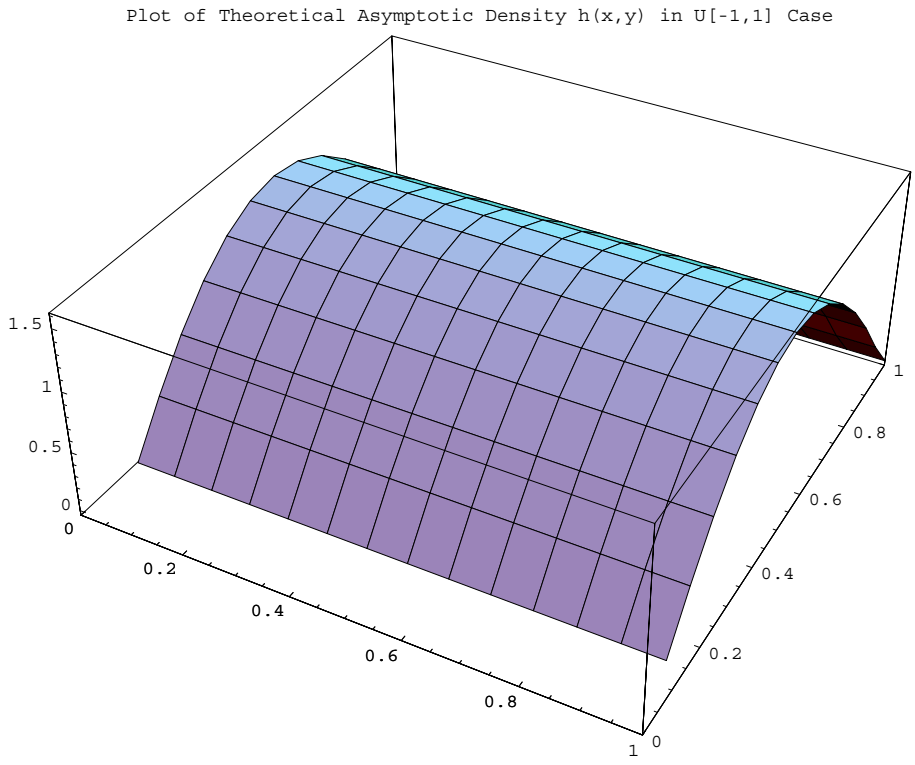
Calculate $\chi^2 = \sum \frac{(O_{ij} - .04B)^2}{.04B}$ and find the P-value $P(\chi^2(24) > \chi^2)$.

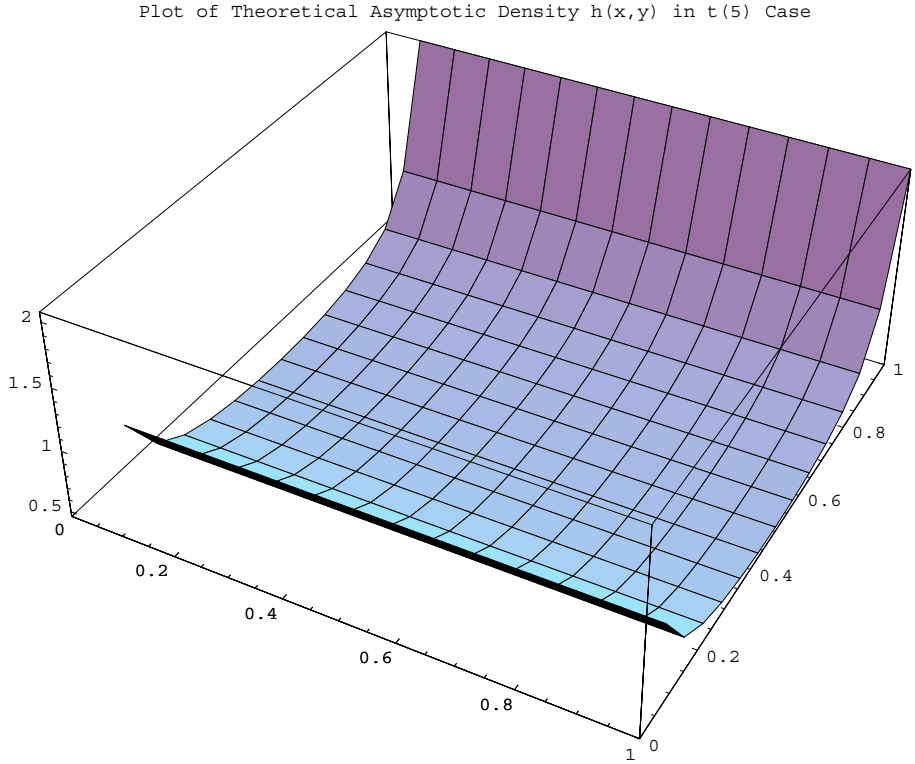
How does the test perform? One way to address the issue is to see whether a test statistic based on the plot has reasonable power. It is clear that the plot-based tests cannot be more powerful than the best test (for a given alternative), but maybe they can be competitive.

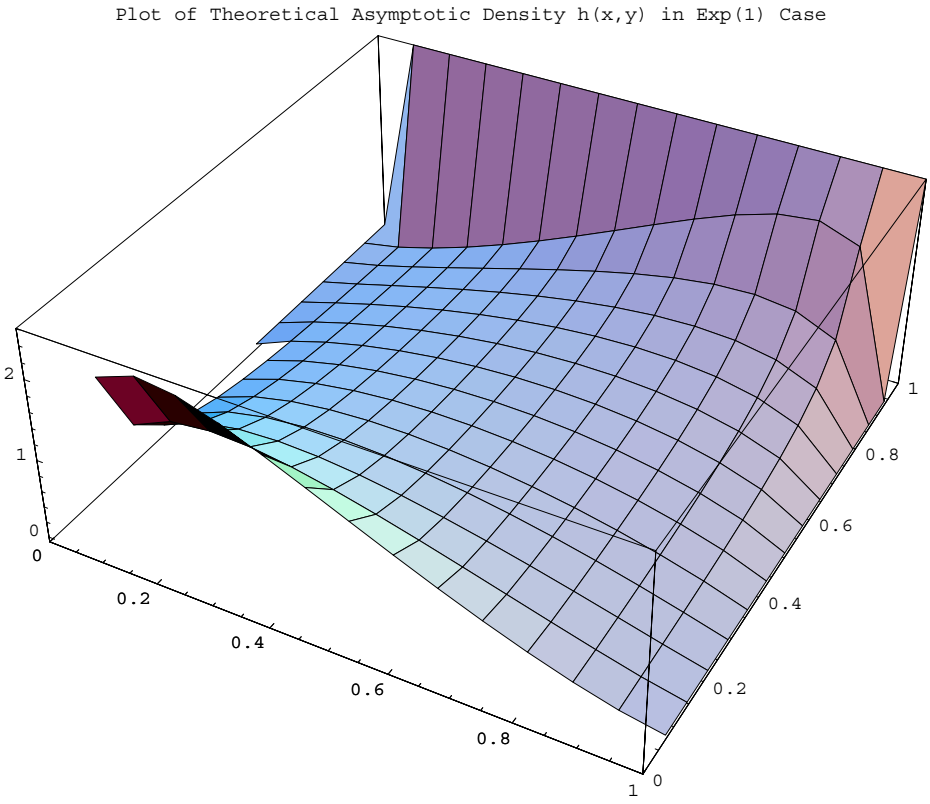
We take the best test to be the likelihood ratio test for testing the alternative versus the normal, using the location-scale family for each distribution. The plot-based tests include the χ^2 test in the paper, two based on the $MAD(v_i^*)$ (median absolute deviation of the v_i^*'s), one which rejects for large values and one for small values, and two based on $Correlation(u_i^*, v_i^*)$. Note the likelihood ratio test can *only* be used when there is a specified alternative, but the plot-based tests are *omnibus*. Thus, what counts is whether the plot-based tests show some all round good performance.

The tables below have the estimated powers (for $\alpha = 0.05$) for various alternatives, for $n = 6$ and 25.

$n = 6$	χ^2	MAD(>)	MAD(<)	Corr(>)	Corr(<)	LRT
Normal	0.050	0.050	0.050	0.050	0.050	0.050
Exponential	0.176	0.075	0.064	0.293	0.006	0.344
Uniform	0.048	0.033	0.105	0.041	0.044	0.118
t_2	0.185	0.079	0.036	0.146	0.138	0.197
t_5	0.070	0.059	0.043	0.064	0.067	0.089


Plot of Theoretical Asymptotic Density h(x,y) in U[-1,1] Case
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1


Plot of Theoretical Asymptotic Density h(x,y) in t(5) Case
2
1.5
1
0.5
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1



$n = 25$	χ^2	MAD(>)	MAD(<)	Corr(>)	Corr(<)	LRT
Normal	0.050	0.050	0.050	0.050	0.050	0.050
Exponential	0.821	0.469	0.022	0.930	0.000	0.989
Uniform	0.164	0.000	0.506	0.045	0.038	0.690
t_2	0.553	0.635	0.003	0.261	0.264	0.721
t_5	0.179	0.208	0.011	0.104	0.121	0.289

The powers for $n = 6$ are naturally fairly low, but we can see that for each distribution, there is a plot-based test that comes reasonably close to the LRT. For the Exponential, the correlation (>) test does very well. For the uniform, the best test rejects for small values of *MAD*. For the *t*'s, rejecting for large values of *MAD* works reasonably well, and the χ^2 and two correlation tests do fine. These results are consistent with the plots in the paper, i.e., for skewed distributions there is a positive correlation between the u_i^*'s and v_i^*'s, and for symmetric distributions, the differences are revealed in the spread of the v_i^*'s . On balance, the Corr(>) test for suspected right skewed cases and the χ^2 test for heavy-tailed symmetric cases seem to be good plot-based formal tests. However, further numerical power studies will be necessary to confirm these recommendations.

3.5. Another pair of statistics

One of the strengths of our approach is that the pair of statistics that can be used to define U_n, V_n is flexible, and therefore different tests can be used to test for normality. We now describe an alternative test based on another pair of statistics.

It too shows impressive power in our simulations in detecting right skewed data for quite small sample sizes.

Let $X_1, X_2, \ldots, X_n$ be the sample values and let Q, s denote respectively the interquartile range and the standard deviation of the data. From Basu's theorem (Basu (1955)), $\frac{Q}{s}$ and s are independent if $X_1, X_2, \ldots, X_n$ are samples from any normal distribution. The exact distribution of $\frac{Q}{s}$ in finite samples is cumbersome. So in forming the quantile transformations, we use the asymptotic distribution of $\frac{Q}{s}$. This is, admittedly, a compromise. But at the end, the test we propose still works very well at least for right skewed alternatives. So the compromise is not a serious drawback at least in some applications, and one has no good alternative to using the asymptotic distribution of $\frac{Q}{s}$. The asymptotic distribution of $\frac{Q}{s}$ for any population F with four moments is explicitly worked out in DasGupta and Haff (2003). In particular, they give the following results for the normal, Exponential and the Beta(2,10) case, the three cases we present here as illustration of the power of this test.

(a) $\sqrt{n}(\frac{IQR}{s} - 1.349) \Rightarrow N(0, 1.566)$ if $F =$ normal;

(b) $\sqrt{n}(\frac{IQR}{s} - 1.099) \Rightarrow N(0, 3.060)$ if $F =$ Exponential.

(c) $\sqrt{n}(\frac{IQR}{s} - 1.345) \Rightarrow N(0, 1.933)$ if $F =$ Beta(2,10).

Hence, as in Subsection 3.1, define:

$u_i^* = \Phi(\frac{\sqrt{n}}{\tau}(\frac{Q_i^*}{s_i^*} - \frac{Q}{s}))$, and $v_i^* = G_{n-1}((n-1)\frac{s_i^{2*}}{s^2})$ and $w_i^* = (u_i^*, v_i^*)$; note that τ^2 is the appropriate variance of the limiting normal distribution of $\frac{IQR}{s}$ as we indicate above. As in Subsection 3.1, we then plot the pairs w_i^* and check for an approximately uniform scattering, *particularly lack of any striking structure.*

The plots below are for the normal, Exponential and Beta(2,10) case; the last two were chosen because we are particularly interested in establishing the efficacy of our procedures for picking up skewed alternatives. It is clear from the plots that for the skewed cases, even at a small sample size $n = 12$, they show striking visual structure, far removed from an approximately uniform scattering. In contrast, the plot for the normal data look much more uniform.

Exactly as in Subsection 3.1, there are analogs of Theorem 3 and Theorem 4 for this case too; however, we will not present them.

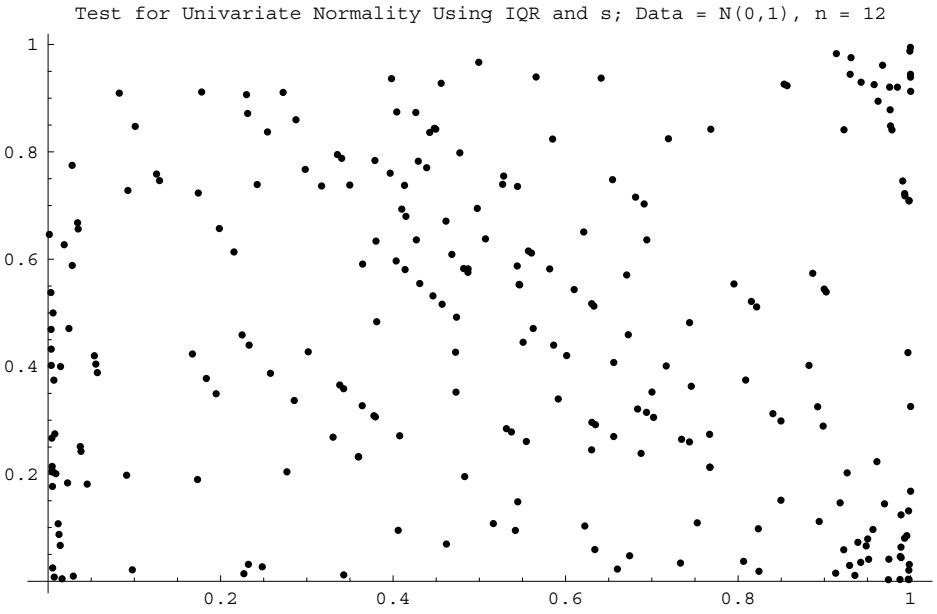
We now address the multivariate case briefly.

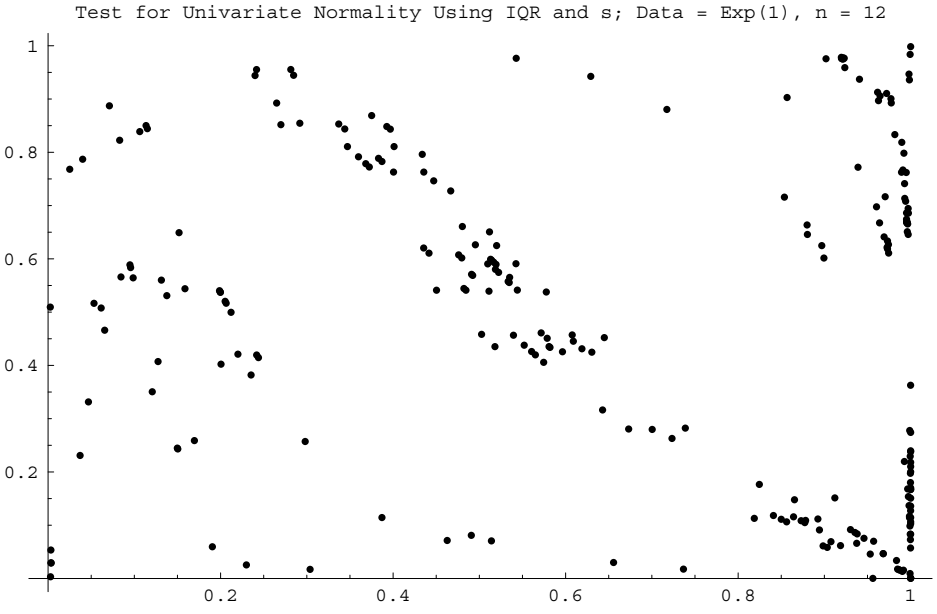
4. Resampling based tests for multivariate normality

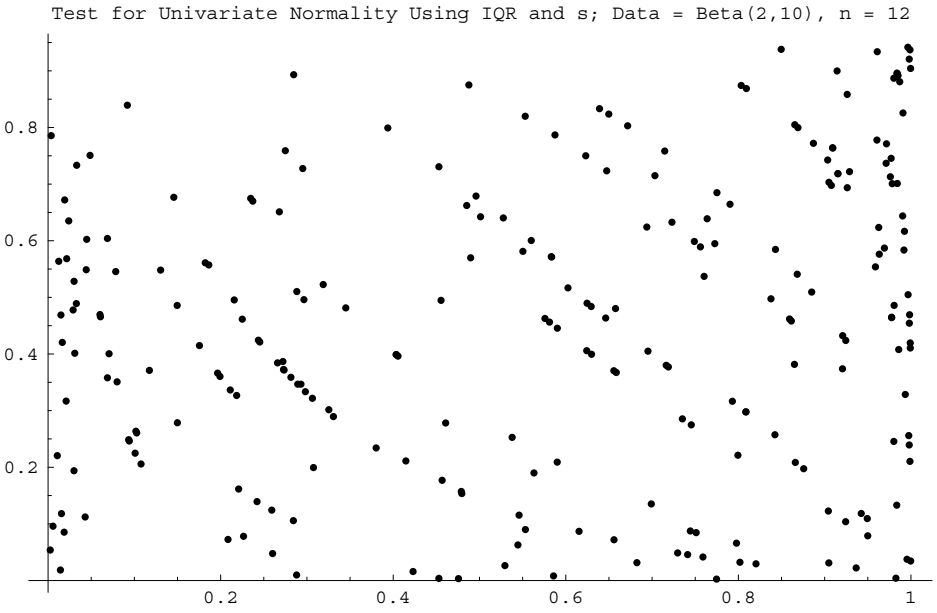
As in the univariate case, our proposed test uses the independence of the sample mean vector and the sample variance-covariance matrix. A difficult issue is the selection of two statistics, one a function of the mean vector and the other a function of the covariance matrix, that are to be used, as in the univariate case, for obtaining the w_i^* via use of the quantile transformation. We use the statistics $c'\overline{X}$, and either $\mathrm{tr}(\Sigma^{-1}S)$, or $\frac{|S|}{|\Sigma|}$. Our choice is exclusively guided by the fact that for these cases, the distributions of the statistics in finite samples are known. Other choices can (and should) be explored, but the technicalities would be substantially more complex.

Test 1. Suppose $X_1, X_2, \ldots, X_n$ are iid p-variate multivariate normal observations, distributed as $N_p(\mu, \Sigma)$. Then, for a given vector c, $c'\overline{X} \sim N_p(c'\mu\ ,\ \frac{1}{n}c'\Sigma c)$, and $\mathrm{tr}(\Sigma^{-1}S) \sim$ *chisquare*$(p(n-1))$. Thus, using the same notation as in Section 3.1,

$$U_n = \Phi\left(\frac{\sqrt{n}(c'\overline{X} - c'\mu)}{\sqrt{c'\Sigma c}}\right) \qquad \text{and} \qquad V_n = G_{p(n-1)}\big(\mathrm{tr}\big(\Sigma^{-1}S\big)\big)$$


Test for Univariate Normality Using IQR and s; Data = N(0,1), n = 12
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Test for Univariate Normality Using IQR and s; Data = Exp(1), n = 12
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1



are independently $U[0,1]$ distributed. For $i = 1, 2, \ldots, B$, define

$$u_i^* = \Phi\left(\frac{\sqrt{n}(c'\overline{X_i}^* - c'\overline{X})}{\sqrt{c'Sc}}\right) \qquad \text{and} \qquad v_i^* = G_{p(n-1)}\big(\mathrm{tr}\big(S^{-1}S_i^*\big)\big),$$

where $\overline{X_i}^*, S_i^*$ are the mean vector and the covariance matrix of the ith bootstrap sample, and $\overline{X}, S$ are the mean vector and the covariance matrix for the original data. As before, we plot the pairs $w_i^* = (u_i^*, v_i^*), i = 1, 2, \ldots, B$ and check for an approximately uniform scattering.

Test 2. Instead of $\mathrm{tr}(\Sigma^{-1}S)$, consider the statistic $\frac{|S|}{|\Sigma|} \sim \prod_{i=1}^{p} \chi^2(n-i)$, where the *chisquare* variables are independently distributed.

For the special case $p = 2$, the distribution can be reduced to that of $\frac{\chi^4(2n-4)}{4}$ (see Anderson (1984)). Hence, U_n (as defined in Test 1 above), and

$$V_n = G_{2n-4}\left(2\frac{|S|^{\frac{1}{2}}}{|\Sigma|^{\frac{1}{2}}}\right)$$

are independently $U[0,1]$ distributed. Define now u_i^* as in Test 1 above, but

$$v_i^* = G_{2n-4}\left(2\frac{|S_i^*|^{\frac{1}{2}}}{|S|^{\frac{1}{2}}}\right),$$

and plot the pairs $w_i^* = (u_i^*, v_i^*)$ to check for an approximately uniform scattering.

The CDF of $\frac{|S|}{|\Sigma|}$ can be written in a reasonably amenable form also for the case $p = 3$ by using the Hypergeometric functions, but we will not describe the three dimensional case here.

As in the univariate case, we will see that Tests 1 and 2 can be quite effective and especially for small samples they are relatively more useful than alternative

tests used in the literature. For example, the common graphical test for bivariate normality that plots the *Mahalanobis* D^2 values against *chisquare* percentiles (see Johnson and Wichern (1992))would not have very much credibility at sample sizes such as $n = 10$ (a sample size we will try).

Corresponding to Theorem 3 , we have a similar consistency theorem.

Theorem 7. $\sup_{0\le u\le 1, 0\le v\le 1} |P_*(U^* \le u,\, V^* \le v) - P_F(U_n \le u, V_n \le v)| \to 0$ *in probability, provided the true CDF* F *has four moments* (*in the usual sense for a multivariate CDF*).

The nonull asymptotics (i.e., the analog of Theorem 4) are *much* harder to write down analytically. We have a notationally messy version for the bivariate case. However, we will not present it due to the notational complexity.

The plots of the pairs w_i^* corresponding to both Test 1 and Test 2 are important to examine from the point of view of applications. The plots corresponding to the first test are presented next. The plots corresponding to the second test look very similar and are omitted here.

The plots again show the impressive power of the tests to detect skewness, as is clear from the Bivariate Gamma plot (we adopt the definition of Bivariate Gamma as $(X, Y) = (U + W, V + W)$, where U, V, W are independent Gammas with the same scale parameter; see Li (2003) for certain recent applications of such representations.) The normal plot looks reasonably devoid of any structure or drastic nonuniformity. Considering that testing for bivariate normality continues to remain a very hard problem for such small sample sizes, our proposals appear to show good potential for being useful and definitely competitive. The ideas we present need to be examined in more detail, however.

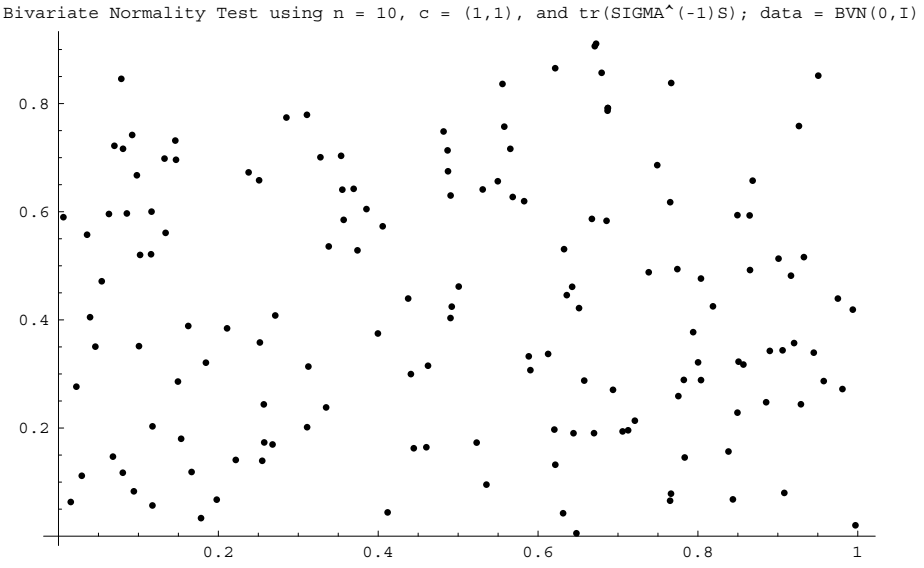
5. Subsampling based tests

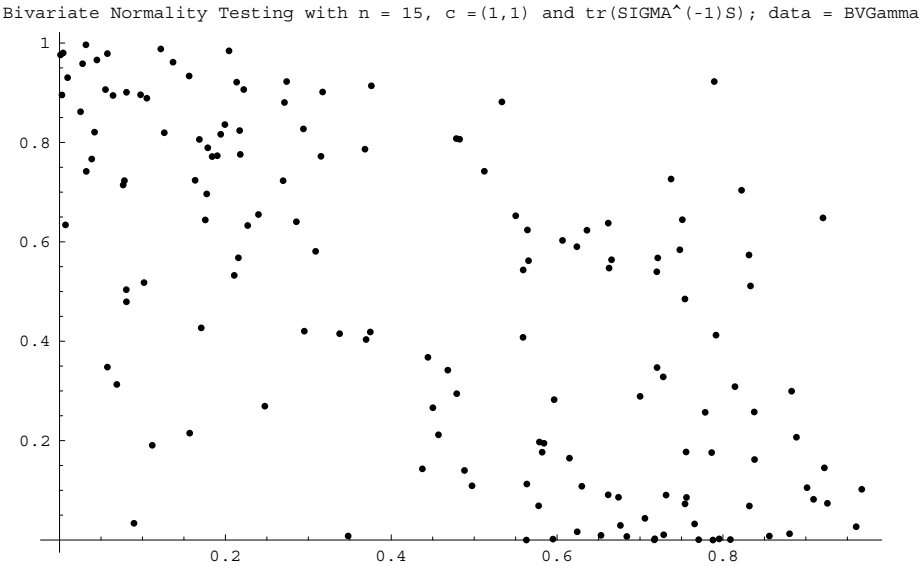
An alternative to the resampling based tests of the preceding sections is to use subsampling. From a purely theoretical point of view, there is no reason to prefer subsampling in this problem. Resampling and subsampling will both produce uniformly consistent distribution estimators, but neither will produce a test that is consistent against all alternatives. However, as a matter of practicality, it might be useful to use each method as a complement to the other. In fact, our subsampling based plots below show that there is probably some truth in that. In this section we will present a brief description of subsampling based tests. A more complete presentation of the ideas in this section will be presented elsewhere.

5.1. Consistency

We return to the univariate case and again focus on the independence of the sample mean and sample variance; however, in this section, we will consider the subsampling methodology—see e.g., Politis, Romano and Wolf (1999). Denote by $\mathcal{B}_{b,1}, \ldots, \mathcal{B}_{b,Q}$ the $Q = \binom{n}{b}$ subsamples of size b that can be extracted from the sample $X_1, \ldots, X_n$. The subsamples are ordered in an arbitrary fashion except that, for convenience, the first $q = [n/b]$ subsamples will be taken to be the non-overlapping stretches, i.e., $\mathcal{B}_{b,1} = (X_1, \ldots, X_b)$, $\mathcal{B}_{b,2} = (X_{b+1}, \ldots, X_{2b}), \ldots,$ $\mathcal{B}_{b,q} = (X_{(q-1)b+1}, \ldots, X_{qb})$. In the above, b is an integer in $(1, n)$ and $[\cdot]$ denotes integer part.

Let $\bar{X}_{b,i}$ and $s^2_{b,i}$ denote the sample mean and sample variance as calculated from subsample $\mathcal{B}_{b,i}$ alone. Similarly, let $U_{b,i} = \Phi(\frac{\sqrt{b}\ (\bar{X}_{b,i} - \mu)}{\sigma})$, and $V_{b,i} =$


Bivariate Normality Test using n = 10, c = (1,1), and tr(SIGMA^(-1)S); data = BVN(0,I)
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Bivariate Normality Testing with n = 15, c =(1,1) and tr(SIGMA^(-1)S); data = BVGamma
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8

$G_{b-1}(\frac{\sqrt{b-1}\ s^2_{b,i}}{\sigma^2})$. Thus, if b were n, these would just be U_n and V_n as defined in subsection 3.1. Note that $U_{b,i}$ and $V_{b,i}$ are not proper statistics since μ and σ are unknown; our proxies for $U_{b,i}$ and $V_{b,i}$ will be $\hat{U}_{b,i} = \Phi(\frac{\sqrt{b}\ (\bar{X}_{b,i}-\bar{X})}{s})$ and $\hat{V}_{b,i} = G_{b-1}(\frac{\sqrt{b-1}\ s^2_{b,i}}{s^2})$ respectively.

Let $H_b(x,y) = P(U_{b,1} \le x, V_{b,1} \le y)$. Recall that, under normality, H_b is uniform on the unit square. However, using subsampling we can consistently estimate H_b (or its limit H given in Theorem 4) whether normality holds or not. As in Politis et al. (1999), we define the subsampling distribution estimator by

$$\hat{L}_b(x,y) = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}\{\hat{U}_{b,i} \le x, \hat{V}_{b,i} \le y\}. \tag{2}$$

Then the following consistency result ensues.

Theorem 8. *Assume the conditions of Theorem* 4. *Then*

(i) *For any fixed integer* $b > 1$, *we have* $\hat{L}_b(x,y) \xrightarrow{P} H_b(x,y)$ *as* $n \to \infty$ *for all points* (x,y) *of continuity of* H_b.

(ii) *If* $\min(b, n/b) \to \infty$, *then* $\sup_{x,y} |\hat{L}_b(x,y) - H(x,y)| \xrightarrow{P} 0$.

Proof. (i) Let (x,y) be a point of continuity of H_b, and define

$$L_b(x,y) = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}\{U_{b,i} \le x, V_{b,i} \le y\}. \tag{3}$$

Note that by an argument similar to that in the proof of Theorem 2.2.1 in Politis, Romano and Wolf (1999), we have that

$$\hat{L}_b(x,y) - L_b(x,y) \to 0$$

on a set whose probability tends to one. Thus it suffices to show that $L_b(x,y) \xrightarrow{P} H_b(x,y)$. But note that $EL_b(x,y) = H_b(x,y)$; hence, it suffices to show that $Var(L_b(x,y)) = o(1)$.

Let

$$\tilde{L}_b(x,y) = \frac{1}{q}\sum_{i=1}^{q} \mathbf{1}\{U_{b,i} \le x, V_{b,i} \le y\}.$$

By a Cauchy–Schwartz argument, it can be shown that $Var(L_b(x,y)) \le Var(\tilde{L}_b(x,y))$; in other words, extra averaging will not increase the variance.

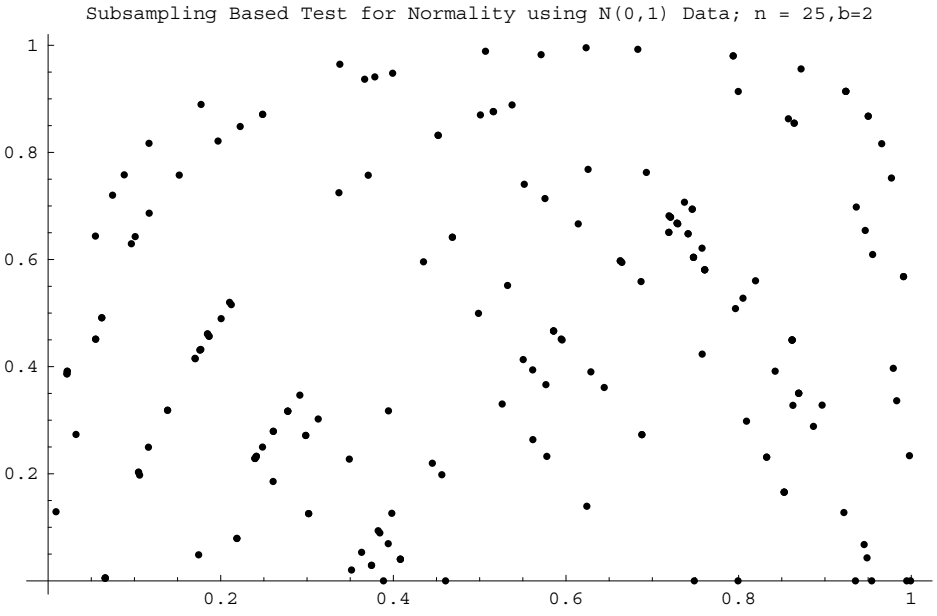
But $Var(\tilde{L}_b(x,y)) = O(1/q) = O(b/n)$ since $\tilde{L}_b(x,y)$ is an average of q i.i.d. random variables. Hence $Var(L_b(x,y)) = O(b/n) = o(1)$ and part (i) is proven. Part (ii) follows by a similar argument; the uniform convergence follows from the continuity of H given in Theorem 4 and a version of Polya's theorem for random cdfs. □

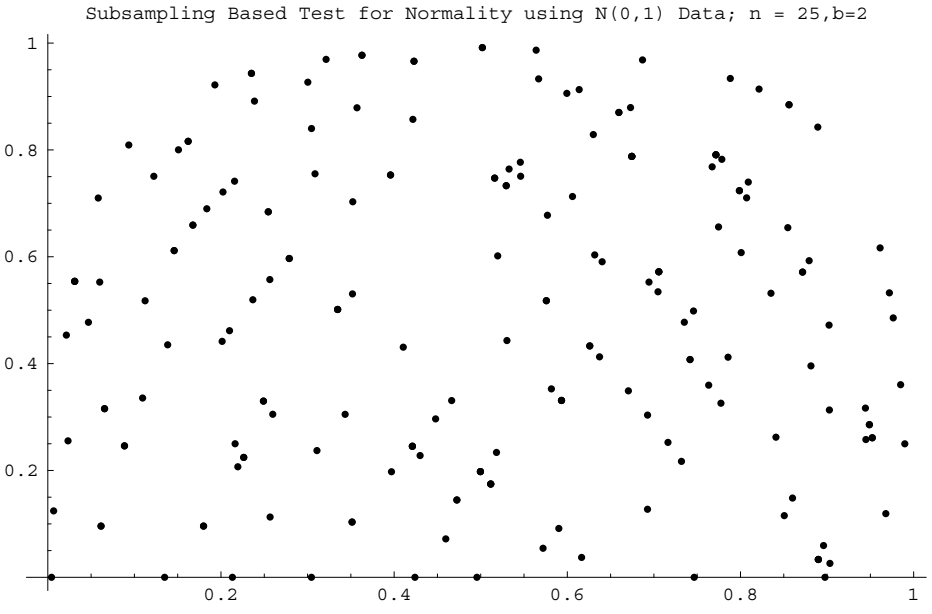
5.2. Subsampling based scatterplots

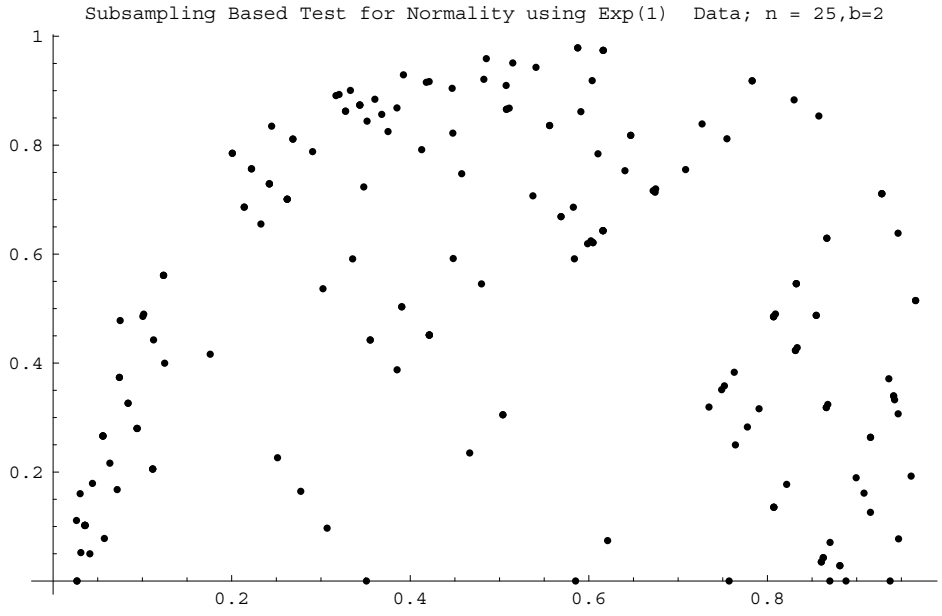
Theorem 8 suggests looking at a scatterplot of the pairs $\hat{w}_{b,i} = (\hat{U}_{b,i}, \hat{V}_{b,i})$ to detect non-normality since (under normality) the points should look uniformly scattered over the unit square, in a fashion analogous to the pairs $\hat{w}^*_i$ in Sections 3 and 4.

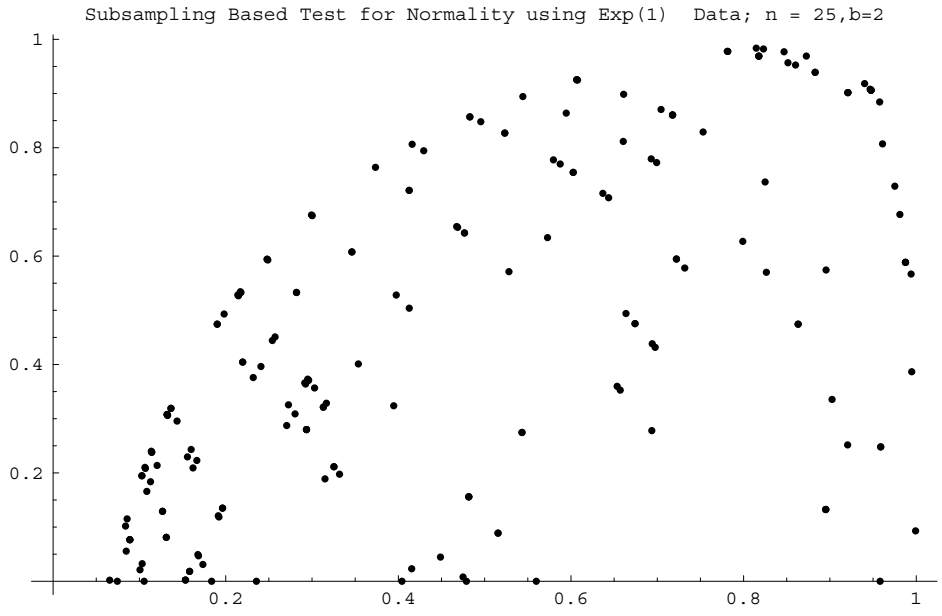
Below, we present a few of these scatterplots and then discuss the plots. The subsample size b in the plots is taken to be 2.

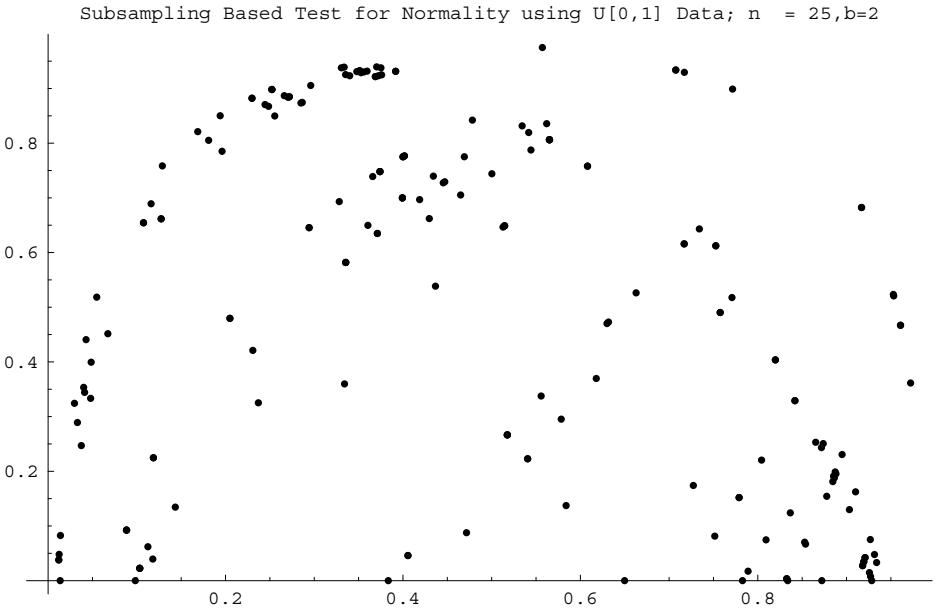
For each distribution, two separate plots are presented to illustrate the quite dramatic nonuniform structure for the nonnormal cases.

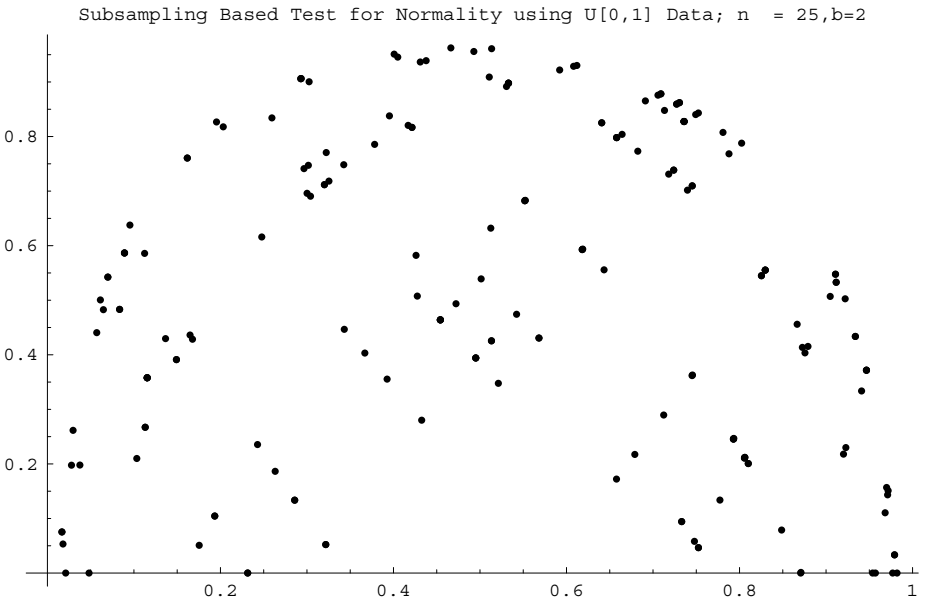

Subsampling Based Test for Normality using N(0,1) Data; n = 25,b=2
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Subsampling Based Test for Normality using N(0,1) Data; n = 25,b=2
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Subsampling Based Test for Normality using Exp(1) Data; n = 25,b=2
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8


Subsampling Based Test for Normality using Exp(1) Data; n = 25,b=2
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1


Subsampling Based Test for Normality using U[0,1] Data; n = 25,b=2
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8


Subsampling Based Test for Normality using U[0,1] Data; n = 25,b=2
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1

5.3. Discussion of the plots

Again, we are forced to present a limited number of plots due to space considerations. The plots corresponding to the Exponential and the uniform case show obvious *nonuniform* structure; they also show significant amounts of empty space. In fact, compared to the corresponding scatterplots for uniform data for the bootstrap based test in Section 3.3, the structured deviation from a uniform scattering is more evident in these plots. Subsampling seems to be working rather well in detecting nonnormality in the way we propose here! But there is also a problem. The problem seems to be that even for normal data, the scatterplots exhibit structured patterns, much in the same way for uniform data, but to *a lesser extent*. Additional theoretical justification for these very special patterns in the plots is needed.

We do not address other issues such as choice of the subsample size due to space considerations and for our focus in this article on just the resampling part.

6. Scope of other applications

The main merits of our proposal in this article are that they give a user something of credibility to use in small samples, and that the proposal has scope for broad applications. To apply our proposal in a given problem, one only has to look for an effective characterization result for the null hypothesis. If there are many characterizations available, presumably one can choose which one to use. We give a very brief discussion of potential other problems where our proposal may be useful. We plan to present these ideas in the problems stated below in detail in a future article.

1. Testing for sphericity

Suppose $X_1, X_2, \ldots, X_n$ are iid p-vectors and we want to test the hypothesis H_0: the common distribution of the X_i is spherically symmetric. For simplicity of explanation here, consider only the case $p = 2$. Literature on this problem includes Baringhaus (1991), Koltchinskii and Li (1998) and Beran (1979).

Transforming each X to its polar coordinates r, θ, under H_0, r and θ are independent. Thus, we can test H_0 by testing for independence of r and θ. The data we will use is a sample of n pairs of values $(r_i, \theta_i), i = 1, 2, \ldots, n$. Although the testing can be done directly from these pairs without recourse to resampling or subsampling, for *small* n, re or subsampling tests may be useful, as we witnessed in the preceding sections in this article.

There are several choices on how we can proceed. A simple correlation based test can be used. Specifically, denoting D_i as the difference of the ranks of the r_i and θ_i (respectively among all the r_i and all the θ_i), we can use the well known Spearman coefficient:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}.$$

For small n, we may instead bootstrap the (r_i, θ_i) pairs and form a scatterplot of the bootstrapped pairs *for each bootstrap replication*. The availability of replicated scatterplots gives one an advantage in assessing if any noticeable correlation between r and θ seems to be present. This would be an easy, although simple, visual method. At a slightly more sophisticated level, we can bootstrap the r_S statistic and compare percentiles of the bootstrap distribution to the theoretical percentiles under H_0 of the r_S statistic. We are suggesting that we break ties just by halving the ranks. For small n, the theoretical percentiles are available exactly; otherwise, we can use

the percentiles from the central limit theorem for r_S as (hopefully not too bad) approximations.

We should mention that other choices exist. An obvious one is Hoeffding's D-statistic for independence. Under H_0, $nD_n + \frac{1}{36}$ has a known (nonnormal) limit distribution. Although an exact formula for its CDF appears to be unknown, from the known formula for its characteristic function (see Hoeffding (1948)), we can pin down any specified percentile of the limit distribution. In addition, for small n, the exact distribution of D_n under H_0 is available too. We can thus find either the exact or approximate percentiles of the sampling distribution of $nD_n + \frac{1}{36}$, and compare percentiles of the bootstrap distribution to them. If we prefer a plot based test, we can construct a Q-Q plot of bootstrap percentiles against the theoretical percentiles under H_0 and interpret the plot in the standard manner a Q-Q plot is used.

2. Testing for Poissonity

This is an important problem for practitioners and has quite a bit of literature, e.g., Brown and Zhao (2002), and Gurtler and Henze (2000). Both articles give references to classic literature. If $X_1, X_2, \ldots, X_n$ are iid from a $Poisson(\lambda)$ distribution, then obviously $\sum_{i=1}^{n} X_i$ is also Poisson-distributed, and therefore every cumulant of the sampling distribution of $\sum_{i=1}^{n} X_i$ is $n\lambda$. We can consider testing that a set of specified cumulants are equal by using re or subsampling methods. Or, we can consider a fixed cumulant, say the third for example, and inspect if the cumulant estimated from a bootstrap distribution behaves like a linear function of n passing through the origin. For example, if the original sample size is $n = 15$, we can estimate a given order cumulant of $\sum_{i=1}^{m} X_i$ for each $m = 1, 2, \ldots, 15$, and visually assess if the estimated values fall roughly on a straight line passing through the origin as m runs through 1 to 15. The graphical test can then be repeated for a cumulant of another order and the slopes of the lines compared for approximate equality too. Using cumulants of different orders would make the test more powerful, and we recommend it.

The cumulants can be estimated from the bootstrap distribution either by differentiating the empirical cumulant generating function $\log(\sum_s e^{ts} P_*(S_n^* = s))$ or by estimating instead the moments and then using the known relations between cumulants and moments (see, e.g., Shiryaev (1980)).

3. Testing for exponentiality

Testing for exponentiality has a huge literature and is of great interest in many areas of application. We simply recommend Doksum and Yandell (1984) as a review of the classic literature on the problem. A large number of characterization results for the family of Exponential distributions are known in the literature. Essentially any of them, or a combination, can be used to test for exponentiality. We do not have reliable information at this time on which characterizations translate into better tests. We mention here only one as illustration of how this can be done.

One possibility is to use the spacings based characterization that $(n - i + 1)R_i$ are iid $Exponential(\lambda)$ where λ is the mean of the population under H_0, and R_i are the successive spacings. There are a number of ways that our general method can be used. Here are a few. A simple plot based test can select two values of i, for example $i = [n/2]$, and $[n/2] + 1$, so that the ordinary bootstrap instead of a m-out of-n bootstrap can be used, and check the pairs for independence. For example, a scatterplot of the bootstrapped pairs can be constructed. Or, one can standardize

the bootstrapped values by $\overline{X}$, so that we will then have pairs of approximately iid *Exponential*(1) values. Then we can use the quantile transformation on them and check these for uniformity in the unit square as in Section 3. Or, just as we described in the section on testing for sphericity, we can use the Hoeffding D-statistic in conjunction with the bootstrap with the selected pairs of $(n-i+1)R_i$. One can then use two other values of i to increase the diagnostic power of the test. There are ways to use all of the $(n-i+1)R_i$ simultaneously as well, but we do not give the details here.

Acknowledgement

Peter Bickel mentioned to one of the authors that uniform data look like normal on a Q-Q plot and suggested a study. Len Haff and David Moore made helpful comments. J. K. Ghosh, Bimal Sinha and Malay Ghosh made comments on the results in Section 2. The work was partially funded by NSF grant DMS 00-71757.

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes, Ann. Math. Stat. 23 193–212. MR50238

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit, JASA 49 765–769. MR69459

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York. MR771294

Babu, G. J. and Rao, C. R. (2004). Goodness of fit tests when parameters are estimated, Sankhyá, 66, To appear. MR2015221

Basu, D. (1955). On statistics independent of a complete and sufficient statistic, Sankhyá, 377–380. MR74745

Baringhaus, L. (1991). Testing for spherical symmetry of a multivariate distribution, Ann. Stat. 19(2) 899–917. MR1105851

Beran, R. (1979). Testing for ellipsoidal symmetry of a multivariate density, Ann. Stat. 7(1) 150–162. MR515690

Brown, L. and Zhao, L. (2002). A test for the Poisson distribution, Sankhyá, A Special Issue in Memory of D. Basu, A. DasGupta Eds., 64, 3, Part I, 611–625. MR1985402

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness of Fit Techniques*, Marcel Dekker Inc., New York. MR874534

DasGupta, A. and Haff, L. R. (2003). Asymptotic values and expansions for correlations between different measures of spread, Invited article for Special issue in Memory of S.S. Gupta, Jour. Stat. Planning and Inf.

Doksum, K. and Yandell, B. (1984). Tests for Exponentiality, In Handbook of Statistics, 4, P.R. Krishnaiah and P.K. Sen Eds., North-Holland, Amsterdam, 579–612. MR831730

Gurtler, N. and Henze, N. (2000). Recent and classical goodness of fit tests for the Poisson distribution, Jour. Stat. Planning and Inf., 90(2) 207–225. MR1795597

Hoeffding, W. (1948). A nonparametric test of independence, Ann. Math. Stat., 19, 546–557. MR29139

Johnson, R. and Wichern, D. (1992). *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, New Jersey. MR1168210

Kagan, A. M., Linnik, Yu. V. and Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics*, John Wiley, New York. MR346969

Koltchinskii, V. and Li, L. (1998). Testing for spherical symmetry of a multivariate distribution, Jour. Mult. Analysis, 65(2) 228–244. MR1625889

Li, Xuefeng (2003). *Infinitely Divisible Time Series Models*, Ph.D. Thesis, University of Pennsylvania, Department of Statistics.

McDonald, K. L. and Katti, S. K. (1974). Test for normality using a characterization, in: *Proc. of Internat. Conf. on Characterizations of Stat. Distributions with Applications*, pp. 91–104.

Mudholkar, G. S., McDermott, M. and Srivastava, D. (1992). A test of p-variate normality, Biometrika **79**(4) 850–854. MR1209484

Mudholkar, G. S., Marchetti, C. E. and Lin, C. T. (2002). Independence characterizations and testing normality against restricted skewness—kurtosis alternatives, Jour. Stat. Planning and Inf., **104**(2) 485–501. MR1906268

Parthasarathy, K. R. (1976). Characterisation of the normal law through the local independence of certain statistics, Sankhyá, Ser. A, **38**(2) 174–178. MR461747

Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*, Springer-Verlag, New York. MR1707286

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York. MR595165

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality: Complete samples, Biometrika **52** 591–611. MR205384

Shiryaev, A. (1980). Probability, Springer, New York. MR1368405

Stephens, M. A. (1976). Asymptotic results for goodness of fit statistics with unknown parameters, Ann. Stat. **4** 357–369. MR397984

van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press, New York. MR1652247

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 207–217
© Institute of Mathematical Statistics, 2004

Notes on the bias-variance trade-off phenomenon

Jeesen Chen[1]

University of Cincinnati

Abstract: The main inequality (Theorem 1) here involves the Hellinger distance of a statistical model of an observation X, which imposes bounds on the mean of any estimator in terms of its variance. We use this inequality to explain some of the bias-variance trade-off phenomena studied in Doss and Sethuraman (1989) and Liu and Brown (1993). We provide some quantified results about how the reduction of bias would increase the variance of an estimator.

1. Introduction

In certain estimation problems the following "bias-variance trade-off" phenomenon might occur: the price of reducing the bias of an etimator T is the dramatic increase of its variance. For problems exhibiting this property, one shouldn't apply the bias reducing procedures blindly. Furthermore, any estimator having good mean square error performance should be biased, and there is a balance between the bias function and the variance function. It is desirable to study the scope of this phenomenon and how the variance and the bias of an estimator affect each other.

Doss and Sethuraman (1989) seem to have been the first to demonstrate the existence of the long suspected bias-variance trade-off phenomenon. However, this result requires stringent conditions, such as the nonexistence of unbiased estimators for the problem and the square integrability of relative densities for the statistical model, thus severely restricting its applicability.

Liu and Brown (1993) broadened the scope of, and brought a new element, the singular/regular property of an estimation problem, into the study of the trade-off phenomenon. Here the focus is on a special aspect of the trade-off phenomenon, the "nonexistence of informative (i.e. bounded variances) unbiased estimators" property, and its connection with the singular/regular property is studied. For singular estimation problems, the bias-variance trade-off phenomenon is an essential component since the "nonexistence of informative unbiased estimators" property always holds (see Theorem 1 of Liu and Brown (1993)). For regular estimation problems, however, the connection is not clear. On one hand, due to the effect of a singular point as a limiting point, the "nonexistence of informative unbiased estimators" property does occur in some regular estimation problems, even though those problems may be quadratic-mean-differentiable with Fisher information totally bounded away from zero. (See Example 2 of Liu and Brown (1993)). On the other hand, there are many known regular estimation problems having informative unbiased estimators. Therefore, focusing on the singular/regular property alone can't completely describe the scope of bias-variance trade-off phenomenon.

[1]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025, USA. e-mail: swilly@math.uc.edu

Keywords and phrases: Hellinger distance, Hellinger modulus, variance-mean relationship, bias-variance trade-off phenomenon, singular problems.

AMS 2000 subject classifications: primary 62F11; secondary 62F12, 62G05, 62A99.

It is intriguing to consider how the results of Liu and Brown (1993) may be perceived. The impression may be that Theorem 1 of Liu and Brown (1993), the "nonexistence of informative unbiased estimators" for a singular estimation problem, seems compatible with the well-known Rao-Cramér inequality. This inequality, under suitable regularity conditions, provides a lower bound of variances for unbiased estimators in terms of the reciprocal of the Fisher information number. For a singular point (or, a point with zero Fisher information number), the lower bound of variances for unbiased estimators becomes infinite; hence it is impossible to have an informative unbiased estimator (if the regularity conditions of Rao-Cramér inequality hold). With this impression, one might be surprised to see Example 4 of Liu and Brown (1993) which exhibits an unbiased estimator with finite variance at a singular point. This seems to contradict the Rao-Cramér inequality or Theorem 1 of Liu and Brown (1993). Of course, there is no contradiction here: first, Example 4 of Liu and Brown (1993) violates the required regularity conditions for the Rao-Cramér inequality; second, Theorem 1 of Liu and Brown (1993) only prevents the possibility of an unbiased estimator having a uniform finite upper bound for variances in any Hellinger neighborhood of a singular point, and not the possibility of an unbiased estimator with finite variance at *a* singular *point*. Nevertheless, the possible confusion indicates the need to find a framework in which we can put all the perception here into a more coherent view. One suggestion is to use an "appropriate variation" of the Rao-Cramér inequality to understand the bias-variance trade-off phenomenon. This modification of the Rao-Cramér inequality would place restrictions regarding the variances of unbiased estimators on the supremum of variances in any Hellinger neighborhood of a point, instead of restricting the variance of the point only. (We believe our results in this paper validate the above suggestion.)

Low (1995), in the context of the functional estimation of finite and infinite normal populations, studies possible bias-variance trade-off by solving explicitly constraint optimization problems: imposing a constraint on either the variance or the square of the bias, then finding the procedure which minimizes the supremum of the unconstrained performance measure. This approach, due to mathematical difficulties involved, seems very difficult to carry out for general estimation problems. However, the investigation of the "bias-variance trade-off" phenomenon in the framework of the study of quantitative restrictions between bias and variance is interesting.

In this paper, we observe that the "nonexistence of informative unbiased estimators" phenomenon and the "bias-variance trade-off" phenomenon exemplify the mutual restrictions between mean functions and variance functions of estimators. These restrictions are described in our main inequality, Theorem 1. We are able to use this inequality to study, for finite sample cases, the "bias-variance trade-off" phenomenon and the "nonexistence of informative unbiased estimators" phenomenon for singular as well as regular estimation problems. A simple application of Theorem 1, Corollary 1, induces a sufficient condition for the "nonexistence of informative unbiased estimators" phenomenon. Corollary 1 is applicable to singular problems, (e.g. it implies Theorem 1 of Liu and Brown (1993)), as well as to regular problems (e.g. Example 2 of Section 4). Additional applications, such as Theorem 2 and Theorem 3, shed further light on the trade-off phenomenon by giving some quantified results. These results not only imply (and extend) Theorem 1 and Theorem 3 of Liu and Brown (1993), they also provide a general lower bound for constraint minimax performance. (See Corollary 3 and related comments.) We may summarize the idea conveyed by these results as: if the estimator we consider has variance less than the smallest possible variances for any unbiased estimators,

then the range of the bias function is at least comparable to a fixed proportion of the range of the parameter function to be estimated.

We address the influence of a singular point as a limiting (parameter) point in Theorem 4. Although this is not a direct consequence of Theorem 1, the format of Theorem 1 facilitates results like Theorem 4.

We state our results in Section 2 and prove them in Section 3. In Section 4 we explain the meaning of Example 2 and Example 4 of Liu and Brown (1993) in our approach to the "bias-variance trade-off" phenomenon. We also argue that examples like Example 4 of Liu and Brown (1993) validate our version of the "mean-variance restriction," in which the restrictions imposing on the bias function of an estimator by its variance function are on the difference of biases at two points instead of the bias function at a point. Example 1 of Section 4, which has been considered by Low (1995) (and maybe others also), shows that our lower bound for minimax performance, Corollary 3, is sharp. The last example, Example 2, shows that the "nonexistence of informative unbiased estimator" phenomenon may occur even if the parameter space does not have any limiting point (with respect to Hellinger distance.)

2. Statements of results

We shall consider the following estimation problem. Let X be a random variable, which takes values in a measure space (Ω, μ), with distribution from a family of probability measures $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$. Furthermore, it is assumed that every P_θ in $\mathcal{F}$ is dominated by the measure μ, and if $P_{\theta_1} = P_{\theta_2}$, then $\theta_1 = \theta_2$. For $\theta \in \Theta$, we denote the Radon-Nikodym derivative of P_θ with respect to the $\sigma-$measure μ as $f_\theta = dP_\theta/d\mu$. For $\theta_1, \theta_2 \in \Theta$, let

$$\rho(\theta_1, \theta_2) := \left\{ \int_\Omega \left[f_{\theta_1}(x)^{1/2} - f_{\theta_2}(x)^{1/2} \right]^2 \mu(dx) \right\}^{1/2} \tag{2.1}$$

denote the Hellinger distance between θ_1 and θ_2, on Θ, induced by the statistical model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$. Suppose $(V, \|\cdot\|)$ is a pseudo-normed linear space, and $q : \Theta \mapsto V$ is a function. We shall estimate $q(\theta)$ based on an observation X. The estimators $T : \Omega \mapsto \mathbf{V}$ we consider are well-behaved functions (satisfying the required measurability conditions) so that, for $\theta \in \Theta$

$$\psi_T(\theta) := \int_\Omega f_\theta(x) T(x) \mu(dx) \tag{2.2}$$

is meaningful and belongs to $\mathbf{V}$, and $v_T^*(\theta) := \int_\Omega f_\theta(x) \|T(x)\|^2 \mu(dx)$ is meaningful. We also adopt the following notaitons:

$$\beta_T(\theta) := E_\theta\big(T(x) - q(\theta)\big), \tag{2.3}$$

the bias function of T;

$$\gamma_T(\theta) := \left\{ E_\theta \|T(X) - q(\theta)\|^2 \right\}^{1/2} \tag{2.4}$$

the mean square risk function of T; and, for $\Theta_0 \subset \Theta$,

$$M_T(\Theta_0) := \sup \left\{ E_\theta \|(T(X) - q(\theta)\|^2 : \theta \in \Theta_0 \right\}. \tag{2.5}$$

The starting point of our study is the following inequality:

Theorem 1. *For $\theta, \theta_0 \in \Theta$, if $\rho(\theta, \theta_0) > 0$, then*

$$\begin{aligned}
&\big[\gamma_T(\theta) + \gamma_T(\theta_0)\big]\rho(\theta, \theta_0) \\
&\quad \geq \left\| \big(\beta_T(\theta) - \beta_T(\theta_0)\big) + \left(1 - \frac{1}{2}\rho^2(\theta, \theta_0)\right)\big(q(\theta) - q(\theta_0)\big) \right\| \\
&\quad \geq \left| \big\|\beta_T(\theta) - \beta_T(\theta_0)\big\| - \left(1 - \frac{1}{2}\rho^2(\theta, \theta_0)\right)\big\|q(\theta) - q(\theta_0)\big\| \right|. \qquad (2.6)
\end{aligned}$$

An easy consequence of (2.6) is:

Corollary 1. *Suppose Θ_1 is a non-empty subset of $\Theta - \{\theta_0\}$, then*

$$\begin{aligned}
&2 \sup\big\{\gamma_T(\theta) : \theta \in \Theta_1 \cup \{\theta_0\}\big\} \\
&\qquad + \sup\left\{ \frac{\|\beta_T(\theta) - \beta_T(\theta_0)\|}{\rho(\theta, \theta_0)} : \theta \in \Theta_1 \right\} \\
&\qquad \geq \sup\left\{ \left[1 - \frac{1}{2}\rho^2(\theta, \theta_0)\right] \frac{\|q(\theta) - q(\theta_0)\|}{\rho(\theta, \theta_0)} : \theta \in \Theta_1 \right\}. \qquad (2.7)
\end{aligned}$$

Let us denote the value of the right-hand side of (2.7) as $Q_q(\theta_0; \Theta_1)$. We point out that the quantity $Q_q(\theta_0; \Theta_1)$ does not depend on the estimator T. It is easy to see that $Q_q(\theta_0; \Theta_1) = \infty$ is a sufficient condition for the "nonexistence of informative unbiased estimators" phenomenon. There are two ways to make $Q_q(\theta_0; \Theta_1) = \infty$: either $\inf_{\theta \in \Theta} \rho(\theta, \theta_0) > 0$ with $\sup_{\theta \in \Theta_1}[1 - \frac{1}{2}\rho^2(\theta, \theta_0)]\|q(\theta) - q(\theta_0)\| = \infty$ or $\inf_{\theta \in \Theta_1} \rho(\theta, \theta_0) = 0$ with $\limsup_{\rho(\theta, \theta_0) \to 0, \theta \in \Theta_1} \frac{\|q(\theta) - q(\theta_0)\|}{\rho(\theta, \theta_0)} = \infty$. See Example 2 of Section 4 for the first case and Examples 1 and 3 of Liu and Brown (1993) for the second case.

In the following, we focus on the case that θ_0 is a limit point of Θ_1 with respect to $\rho-$distance. Note that we may replace $Q_q(\theta_0; \Theta_1)$ in the right-hand side of (2.7) by an easily computable lower bound $\limsup_{\rho(\theta, \theta_0) \to 0, \theta \in \Theta_1} \frac{\|q(\theta) - q(\theta_0)\|}{\rho(\theta, \theta_0)}$.

For the convenience of our discussion let us introduce:

Definition 1 (Hellinger Information). Suppose $\Theta_1 \subset \Theta$ and θ_0 is a non-isolated point of Θ_1 with respect to $\rho-$metric on Θ. The Hellinger Information of θ_0 about the $q(\cdot)-$estimation problem and the (sub-)parameter space Θ_1 is defined as

$$J_q(\theta_0; \Theta_1) := 4 \left[\limsup_{\rho(\theta, \theta_0) \to 0+, \theta \in \Theta_1} \frac{\|q(\theta) - q(\theta_0)\|}{\rho(\theta, \theta_0)} \right]^{-2}. \qquad (2.8)$$

For the development of this notation and its relationship to Fisher Information, see Chen (1995). We mention here that this notation is related to "sensitivity" proposed by Pitman (1978). Also, it is equivalent to the "Geometric Information" in Donoho and Liu (1987), and, in terms of Hellinger modulus (see Liu and Brown (1993) (2.9) and (2.2)), it is $(\lim_{\varepsilon \to 0+} \frac{b(\varepsilon)}{\varepsilon})^{-2}$. When $J_q(\theta_0; \Theta) = 0$ (resp. > 0), we say that the $q(\cdot)-$estimation problem is singular (resp. regular) at point θ_0.

With the notation of Hellinger Information, an easy corollary of Theorem 1 is:

Corollary 2. *Suppose θ_0 is an accumulation point of $\Theta_0 \subset \Theta$. Then, for $J = J_q(\theta_0; \Theta_0)$*

$$2\,[M_T(\Theta_0)]^{1/2} + \sup\left\{ \frac{\|\beta_T(\theta) - \beta_T(\theta_0)\|}{\rho(\theta, \theta_0)} : \theta \in \Theta_0, \theta \neq \theta_0 \right\} \geq \frac{2}{\sqrt{J}}, \qquad (2.9)$$

or, equivalently,

$$\sup\left\{\frac{\|\beta_T(\theta)-\beta_T(\theta_0)\|}{\rho(\theta,\theta_0)} : \theta\in\Theta_0,\ \theta\neq\theta_0\right\} \geq \frac{2}{\sqrt{J}}\Big[1-\big(M_T(\Theta_0)J\big)^{1/2}\Big]. \quad (2.10)$$

A trivial implication of (2.10) is: if $M_T(\Theta_0) < 1/J_q(\theta_0;\Theta_0)$, then T is not unbiased on Θ_0. Moreover, (2.10) puts a restriction on the bias function $\beta_T(\theta)$ of T. We shall state this restriction more explicitly in the next theorem.

Theorem 2. *Suppose θ_0 is an accumulation point of $\Theta_0\subset\Theta$, and M is a positive number such that $M < [J_q(\theta_0;\Theta_0)]^{-1}$. Let $d_M := 1-(MJ_q(\theta_0;\Theta_0))^{1/2}$. Suppose that T is an estimator with $E_\theta\|T(X)-q(\theta)\|^2\leq M$ for all $\theta\in\Theta_0$. Then, for any $\lambda>0$, there exist $\theta_\lambda\in\Theta_0$, not dependent on T, such that $0<\rho(\theta_\lambda,\theta_0)\leq(2\lambda d_M)^{1/2}$, $\|q(\theta_\lambda)-q(\theta_0)\|>0$, and*

$$\big\|\beta_T(\theta_\lambda)-\beta_T(\theta_0)\big\| \geq (1-\lambda)d_M\cdot\big\|q(\theta_\lambda)-q(\theta_0)\big\|. \quad (2.11)$$

Applying Theorem 2, it is easy to obtain a lower bound for constrained minimax performance.

Corollary 3. *Let θ_0 be an accumulation point of Θ, $J=J_q(\theta_0;\Theta)>0$. Let M and τ be positive numbers, and*

$$B(M;\tau) := \inf_T \sup_\theta \big\{\|\beta_T(\theta)-\beta_T(\theta_0)\|^2\big\}$$

where θ is over $\|q(\theta)-q(\theta_0)\|\leq\tau$ and T is over $E_\theta\|T(X)-q(\theta)\|^2\leq M$. Then

$$B(M;\tau)\geq\Big\{\big[1-(MJ)^{1/2}\big]\wedge 0\Big\}^2\tau^2. \quad (2.12)$$

In the restriction normal mean case (see Example 1), the lower bound (2.12) is sharp.

Now, let us turn to the case in which θ_0 is a singular point, i.e., $J_q(\theta_0;\Theta_0)=0$. From (2.9) or (2.10), we have either $M_T(\Theta_0)=\infty$ or $\sup\{\frac{\|\beta_T(\theta)-\beta_T(\theta_0)\|}{\rho(\theta,\theta_0)} : \theta\in\Theta_0,\ \theta\neq\theta_0\}=\infty$. This implies the non-existence of an informative unbiased estimator for such Θ_0. Therefore, Theorem 1 of Liu and Brown (1993) is a weaker version of Corollary 2.

From Theorem 2 (or Corollary 3), it is easy to see that there exists no sequence of asymptotically unbiased estimators (based on the same finite number of observations) that would have uniformly bounded variance in any small Hellinger neighborhood of a singular point θ_0. Hence, Theorem 2 above implies Theorem 3 of Liu and Brown (1993). For singular estimation problems, those estimators achieving good mean square error performance must balance bias and variance, and (2.11) gives a quantitative result about its bias function $\beta_T(\theta)$. Furthermore, we are able to describe the "rate" of $\|\beta_T(\theta)-\beta_T(\theta_0)\|$ as follows.

Theorem 3. *Suppose $J_q(\theta_0;\Theta_0)=0$. Let $\Theta_1=\{\theta_1,\theta_2,\ldots\}\subset\Theta_0-\{\theta_0\}$ be a slow sequence of θ_0 in the sense that $\lim_{j\to\infty}\rho(\theta_j,\theta_0)=0$ and $\lim_{j\to\infty}\frac{\|q(\theta_j)-q(\theta_0)\|}{\rho(\theta_j,\theta_0)}=\infty$. If T is an estimator with $\sup\{E_{\theta_j}\|T(X)-q(\theta_j)\|^2 : j=0,1,2,\ldots\}<\infty$, then*

$$\lim_{j\to\infty}\frac{\|\beta_T(\theta_j)-\beta_T(\theta_0)\|}{\|q(\theta_j)-q(\theta_0)\|}=1. \quad (2.13)$$

One of the important observations of Liu and Brown (1993) is that the bias-variance trade-off phenomenon might occur on a set Θ_1 due to the effect of a singular point θ_0 as a limit point of Θ_1. The next result states it more explicitly.

Theorem 4. *Suppose* $\mathbf{V}$ *is a subspace of* $d-$*dimensional Euclidean space* $\mathbf{R}^d$ *with the usual Euclidean norm* $\|\cdot\|$. *Let* θ_0 *be a singular point,* $\Theta_1 = \{\theta_1, \ldots\} \subset \Theta - \{\theta_0\}$ *be a slow sequence of* θ_0 *and* T *be an unbiased estimator on* Θ_1. *Then,* $\sup\{E_\theta \|T(X) - q(\theta)\|^2 : \theta \in \Theta_1\} = \infty$.

3. Proofs

Theorem 1 is a simple application of the following inequality.

Lemma 1. *For points* $\eta_1, \eta_2 \in \mathbf{V}; \theta_1, \theta_2 \in \Theta$ *with* $\rho(\theta_1, \theta_2) > 0$, *we have*

$$\begin{aligned}
&\Big\{E_{\theta_1}\big\|T(X) - \eta_1\big\|^2\Big\}^{1/2} + \Big\{E_{\theta_2}\big\|T(X) - \eta_2\big\|^2\Big\}^{1/2} \\
&\qquad \geq \left\|\psi_T(\theta_1) - \psi_T(\theta_2) - \frac{1}{2}\rho^2(\theta_1, \theta_2)(\eta_1 - \eta_2)\right\| \Big/ \rho(\theta_1, \theta_2). \qquad (3.1)
\end{aligned}$$

Proof. Without loss of generality, we assume that $E_{\theta_i}\|T(X) - \eta_i\|^2 < \infty$ for $i = 1, 2$. Define $\alpha_i(x) = f_{\theta_i}(x)^{1/2}(T(x) - \eta_i)$ for $i = 1, 2$, and $\beta(x) = f_{\theta_1}(x)^{1/2} - f_{\theta_2}(x)^{1/2}$. Then

$$\begin{aligned}
&\int_\Omega \beta(x)\big[\alpha_1(x) + \alpha_2(x)\big]\mu(dx) \\
&\quad = \int_\Omega \Big[f_{\theta_1}(x)\big(T(x) - \eta_1\big) - f_{\theta_2}(x)\big(T(x) - \eta_2\big) \\
&\qquad\qquad + \big[f_{\theta_1}(x) f_{\theta_2}(x)\big]^{1/2}(\eta_1 - \eta_2)\Big]\mu(dx) \\
&\quad = E_{\theta_1}\big(T(X) - \eta_1\big) - E_{\theta_2}\big(T(X) - \eta_2\big) \\
&\qquad + \int_\Omega \big[f_{\theta_1}(x) f_{\theta_2}(x)\big]^{1/2}\mu(dx)(\eta_1 - \eta_2) \\
&\quad = \big(\psi_T(\theta_1) - \eta_1\big) - \big(\psi_T(\theta_2) - \eta_2\big) + \left[1 - \frac{1}{2}\rho^2(\theta_1, \theta_2)\right](\eta_1 - \eta_2) \\
&\quad = \big(\psi_T(\theta_1) - \psi_T(\theta_2)\big) - \frac{1}{2}\rho^2(\theta_1, \theta_2)(\eta_1 - \eta_2). \qquad (3.2)
\end{aligned}$$

On the other hand, by the triangle inequality and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\left\|\int_\Omega \beta(x)\big[\alpha_1(x) + \alpha_2(x)\big]\mu(dx)\right\| \\
&\qquad \leq \sum_{i=1}^{2}\left\|\int_\Omega \beta(x)\alpha_i(x)\mu(dx)\right\| \\
&\qquad \leq \sum_{i=1}^{2}\int_\Omega |\beta(x)|\big\|\alpha_i(x)\big\|\mu(dx)
\end{aligned}$$

$$\leq \sum_{i=1}^{2} \left[\int_{\Omega} \beta^2(x)\mu(dx)\right]^{1/2} \cdot \left[\int_{\Omega} \left\|\alpha_i(x)\right\|^2 \mu(dx)\right]^{1/2}$$

$$= \rho(\theta_1,\theta_2) \sum_{i=1}^{2} \left[E_{\theta_i}\left\|T(X)-\eta_i\right\|^2\right]^{1/2}. \tag{3.3}$$

Combining (3.2) and (3.3), we obtain (3.1). □

Proof of Theorem 1. Applying Lemma 1, we have

$$\begin{aligned}
&\big[\gamma_T(\theta)+\gamma_T(\theta_0)\big]\rho(\theta,\theta_0) \\
&\quad \geq \left\|\psi_T(\theta)-\psi_T(\theta_0)-\frac{1}{2}\rho^2(\theta,\theta_0)\big(q(\theta)-q(\theta_0)\big)\right\| \\
&\quad = \left\|\beta_T(\theta)-\beta_T(\theta_0)+\left(1-\frac{1}{2}\rho^2(\theta,\theta_0)\right)\big(q(\theta)-q(\theta_0)\big)\right\|, \qquad (3.4)
\end{aligned}$$

this proves the first inequality of (2.6).

Applying the triangle inequality and the fact that $1-\frac{1}{2}\rho^2(\theta,\theta_0)\geq 0$, we obtain the second inequality of (2.6). □

Proof of Corollary 1. Notice that (2.6) implies

$$\begin{aligned}
&2\max\big(\gamma_T(\theta),\gamma_T(\theta_0)\big)+\frac{\|\beta_T(\theta)-\beta_T(\theta_0)\|}{\rho(\theta,\theta_0)} \\
&\qquad \geq \left[1-\frac{1}{2}\rho^2(\theta,\theta_0)\right]\frac{\|q(\theta)-q(\theta_0)\|}{\rho(\theta,\theta_0)}. \qquad (3.5)
\end{aligned}$$

Letting θ vary over Θ_1 in inequality (3.5), we obtain (2.7). □

Proof of Corollary 2. It is easy to prove that

$$\begin{aligned}
Q_q(\theta_0;\Theta_0) &\geq \limsup_{\rho(\theta,\theta_0)\to 0,\theta\in\Theta_0} \frac{\|q(\theta)-q(\theta_0)\|}{\rho(\theta,\theta_0)} \\
&= \left[\frac{1}{4}J_q(\theta_0;\Theta_0)\right]^{-1/2}. \qquad (3.6)
\end{aligned}$$

This, together with Corollary 1, proves (2.9). □

Proof of Theorem 2. We use J to replace $J_q(\theta_0;\Theta_0)$ in this proof.

Applying Theorem 1 and the condition $\gamma_T(\theta)+\gamma_T(\theta_0)\leq 2M^{1/2}$, we have, for all $\theta\in\Theta_0$, that

$$\frac{\|\beta_T(\theta)-\beta_T(\theta_0)\|}{\rho(\theta,\theta_0)} \geq \left[1-\frac{1}{2}\rho^2(\theta,\theta_0)\right]\frac{\|q(\theta)-q(\theta_0)\|}{\rho(\theta,\theta_0)} - 2M^{1/2} \tag{3.7}$$

and

$$\begin{aligned}
&\frac{\|\beta_T(\theta)-\beta_T(\theta_0)\|}{\|q(\theta)-q(\theta_0)\|} \\
&\qquad \geq 1-\frac{1}{2}\rho^2(\theta,\theta_0)-2M^{1/2}\cdot\left[\frac{\|q(\theta)-q(\theta_0)\|}{\rho(\theta,\theta_0)}\right]^{-1}. \qquad (3.8)
\end{aligned}$$

For $\varepsilon > 0$, let $\Theta_0(\varepsilon) := \{\theta : q(\theta) \neq q(\theta_0), 0 < \rho(\theta, \theta_0) \leq \varepsilon\} \cap \Theta_0$. By (3.8), for $\varepsilon = (2\lambda d_M)^{1/2}$, we have

$$\begin{aligned} \sup_{\theta \in \Theta_0(\varepsilon)} \frac{\|\beta_T(\theta) - \beta_T(\theta_0)\|}{\|q(\theta) - q(\theta_0)\|} &\geq 1 - \frac{1}{2}\varepsilon^2 - 2M^{1/2} \cdot \left(\frac{1}{4}J\right)^{1/2} \\ &= (1-\lambda)d_M. \end{aligned} \tag{3.9}$$

This proves Theorem 2. □

Proof of Corollary 3. Let $d_M := 1 - (MJ)^{1/2}$. For the case $d_M \leq 0$, we use the trivial inequality $B(M;\tau) \geq 0$ and for the case $d_M > 0$, we use Theorem 2 to obtain $B(M;\tau) \geq (d_M\tau)^2$. This proves (2.12). □

Proof of Theorem 3. Let M be a positive number such that $E_{\theta_j}\|T(X) - q(\theta_j)\|^2 \leq M^2$ for $j = 0, 1, 2, \ldots$. Then, by (2.6),

$$2M \geq \frac{\|q(\theta_j) - q(\theta_0)\|}{\rho(\theta_j, \theta_0)} \left| \frac{\|\beta_T(\theta_j) - \beta_T(\theta_0)\|}{\|q(\theta_j) - q(\theta_0)\|} - \left(1 - \frac{1}{2}\rho^2(\theta_j, \theta_0)\right)\right|. \tag{3.10}$$

Hence,

$$\begin{aligned} &1 + \rho^2(\theta_j, \theta_0)/2 - 2M\big[\big\|q(\theta_j) - q(\theta_0)\big\|/\rho(\theta_j, \theta_0)\big]^{-1} \\ &\quad \leq \big\|\beta(\theta_j) - \beta(\theta_0)\big\|/\big\|q(\theta_j) - q(\theta_0)\big\| \\ &\quad \leq 1 + \rho^2(\theta_j, \theta_0)/2 + 2M\big[\big\|q(\theta_j) - q(\theta_0)\big\|/\rho(\theta_j, \theta_0)\big]^{-1}. \end{aligned} \tag{3.11}$$

Let $j \to \infty$, we have the desired (2.13). □

In order to prove Theorem 4, we need the following lemma.

Lemma 2. *Suppose* $\mathbf{V}$ *is a subspace of* d*-dimensional Euclidean space* $\mathbf{R}^d$ *with the usual Euclidean norm* $\|\cdot\|$. *Let* $\Theta_1 = \{\theta_1, \ldots\} \subset \Theta - \{\theta_0\}$ *be a sequence with limit point* θ_0 *and* $\lim_{j\to\infty} q(\theta_j) = q(\theta_0)$. *Then, for any estimator* T,

$$E_{\theta_0}\big\|T(X) - q(\theta_0)\big\|^2 \leq M_T(\Theta_1). \tag{3.12}$$

Proof. (3.12) is automatically true if $M_T(\Theta_1) = \infty$. Let us consider the case that $M_T(\Theta_1) < \infty$. Since $\lim_{j\to\infty} \rho(\theta_j, \theta_0) = 0$ as $j \to \infty$, the distribution of T under $\theta = \theta_j$ converges to the distribution of T under $\theta = \theta_0$. Let us write

$$\begin{aligned} T &= (T_1, T_2, \ldots, T_d), \\ q(\theta) &= \big(q_1(\theta), q_2(\theta), \ldots, q_d(\theta)\big), \\ \psi_T(\theta) &= E_\theta T(X) = \big(\psi_1(\theta), \psi_2(\theta), \ldots, \psi_d(\theta)\big), \qquad \text{and} \\ \nu_T(\theta) &= \big(\mathrm{var}_\theta(T_1), \mathrm{var}_\theta(T_2), \ldots, \mathrm{var}_\theta(T_d)\big). \end{aligned}$$

Notice that

$$\begin{aligned} &E_\theta\big\|T(X) - q(\theta)\big\|^2 \\ &\quad = E_\theta\big\|T(X) - \psi_T(\theta)\big\|^2 + \big\|\psi_T(\theta) - q(\theta)\big\|^2 \\ &\quad = \sum_{i=1}^d \mathrm{var}_\theta(T_i) + \sum_{i=1}^d \big(\psi_i(\theta) - q_i(\theta)\big)^2, \end{aligned} \tag{3.13}$$

and, since $M_T(\Theta_1) < \infty$,

$$\lim_{j\to\infty} \psi_i(\theta_j) = \psi_i(\theta_0) \qquad \text{for } i = 1, 2, \ldots, d. \tag{3.14}$$

By Problem 4.4.9, page 150 of Bickel and Doksum (1977), we have

$$\liminf_j \operatorname{var}_{\theta_j}(T_i) \geq \operatorname{var}_{\theta_0}(T_i) \qquad \text{for } i = 1, 2, \ldots, d. \tag{3.15}$$

With the assumption $\lim_{j\to\infty} q(\theta_j) = q(\theta_0)$, and (3.13) $\sim$ (3.15), we have

$$\liminf_j E_{\theta_j} \big\|T(X) - q(\theta_j)\big\|^2 \geq E_{\theta_0} \big\|T(X) - q(\theta_0)\big\|^2. \tag{3.16}$$

This proves (3.12). □

Proof of Theorem 4. First, if $\lim_{j\to\infty} q(\theta_j) = \infty$, then it is easy to prove that $\sup\{E_\theta \|T(X) - q(\theta)\|^2 : \theta \in \Theta_1\} = \infty$. Next, if $\lim_{j\to\infty} q(\theta_j)$ exists, we simply change the definition of $q(\theta_0)$ to be equal to $\lim_{j\to\infty} q(\theta_j)$. Under this new definition of q, θ_0 is still a singular point and Θ_1 is still a slow sequence of θ_0. If $M_T(\Theta_1) < \infty$, then (3.12) implies $M_T(\Theta_1 \cup \{\theta_0\}) = M_T(\Theta_1) < \infty$ and (2.9) implies $2[M_T(\theta_1 \cup \{\theta_0\})] = \infty$, a contradiction. This proves $M_T(\Theta_1) = \infty$. □

4. Comments and examples

Example 2 of Liu and Brown (1993) shows that the "nonexistence of informative unbiased estimators" phenomenon might occur in a quadratic-mean-differentiable (QMD) problem with Fisher Information totally bounded away from zero. This statement is true if we replace the term "Fisher Information" by "Hellinger Information" since it is well-known that Fisher Information and Hellinger Information are equal in QMD problems. Due to the fact that the Hellinger Information number $J(\theta)$ is not necessarily continuous with respect to the Hellinger distance $\rho(\theta; \theta_0)$, the condition that Fisher Information (or Hellinger Information) be totally bounded away from zero does not exclude the possibility of a singular point as a limiting point. If such a singular limiting point exists, by Theorem 4, the "nonexistence of informative unbiased estimators" phenomenon could occur.

Example 4 of Liu and Brown (1994) exhibits an unbiased estimator with finite variance at a singular point. The spirit of this example does not contradict the impression left by the "mean-variance restriction" described in Theorem 1 or Corollary 1. Obviously, one can modify an estimator so as to obtain an unbiased estimator at any predescribed point. The requirement that an estimator have finite variance at a predescribed point does not pose any conflict because the "mean-variance restriction" (Theorem 1) places a lower bound on the *sum* of variances at two points, instead of on variances at each point. Further, one could even view this example as a validation of the form of "mean-variance restriction" (Theorem 1), in which the restriction imposed by sums of variances (or, rather, sums of root mean-square risks) is on the difference of the bias function $(\beta_T(\theta) - \beta_T(\theta_0))$ and not on the bias function $(\beta_T(\theta_0))$ itself.

The following example shows that in the bounded normal case, the lower bound of Corollary 3 is sharp. This example has been considered by Low (1995).

Example 1. If $X \sim N(\theta, \sigma^2)$ and $q(\theta) = \theta$, then $J = J_q(\theta_0; \Theta) = \frac{1}{\sigma^2}$ for any open interval Θ which contains θ_0. By (2.9),

$$B(M; \tau) \geq \left\{\left[1 - M^{1/2} \cdot \sigma^{-1}\right] \wedge 0\right\}^2 \tau^2. \tag{4.1}$$

Let T_M be the affine procedure studied in Low (1995), (2.4),

$$T_M(X) = \big(M^{1/2} \cdot \sigma^{-1} \wedge 1\big)(X - \theta_0) + \theta_0. \tag{4.2}$$

It is easy to show that $E_\theta \|T_M(X) - \theta\|^2 \le M \wedge \sigma^2$ and that

$$\sup\big\{\big\|\beta_{T_M}(\theta) - \beta_{T_M}(\theta_0)\big\|^2 : |\theta - \theta_0| \le \tau\big\} = \big\{\big[1 - M^{1/2} \cdot \sigma^{-1}\big] \wedge 0\big\}^2 \tau^2.$$

This, together with (2.12) proves

$$B(M;\tau) = \big\{\big[1 - M^{1/2} \cdot \sigma^{-1}\big] \wedge 0\big\}^2 \tau^2. \tag{4.3}$$

If we compare $B(M;\tau)$ with $\beta^2(\nu,\sigma,\tau)$ in (2.1) and (2.3) of Low (1995), we find that $B(M;\tau) = \beta^2(M,\sigma,\tau)$ in the above Example 2. It is interesting to point out that Low's argument to obtain a lower bound on $\beta^2(\nu,\sigma,\tau)$ is an application of the Rao-Cramér Inequality. This approach, if extended to a general case, would require conditions to guarantee the differentiability of the bias function of T. Our method, which is based on Theorem 1, does not require the differentiability of the bias function of T.

Finally, let us exhibit an example of the "nonexistence of informative unbiased estimator" phenomenon for discrete Θ without any limiting point with respect to ρ-distance.

Example 2. Let $X \sim$ Poisson (θ) with $\theta \in N = \{1,2,3,\ldots\}$, and $r > 1$. Suppose we want to estimate $q(\theta) = e^{r\theta}$. The square of Hellinger distance is

$$\begin{aligned} \rho^2(\theta,1) &= 2 - 2\sum_{x=0}^{\infty} \left\{ \frac{e^{-\theta} \cdot \theta^x}{x!} \cdot \frac{e^{-1} \cdot 1^x}{x!} \right\}^{1/2} \\ &= 2 - 2\exp\left\{ -\frac{1}{2}\left(\sqrt{\theta} - 1\right)^2 \right\}. \end{aligned}$$

It is easy to verify that $[1 - \frac{1}{2}\rho^2(\theta,1)]\|q(\theta) - q(1)\|/\rho(\theta,1) \to \infty$ for $\theta \to \infty$. According to Corollary 1, there exists no informative unbiased estimator for $q(\theta)$.

Acknowledgements

The author wishes to thank Professors Larry Brown and Richard Liu, for letting him read their work (Liu and Brown (1993)) prior to publication. Thanks to Professor Siva Sivaganesan and Ms. Clarissa Howison for their suggestions on the presentation of this paper.

References

[1] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics, Basic Ideas and Selected Topics.* Prentice Hall, Englewood Cliffs, New Jersey. MR443141

[2] Chen, J. (1995). A Hellinger Distance Inequality and its Applications to Variance-Mean Relationship in Estimation Problems. Unpublished.

[3] Donoho, D. L. and Liu, R. C. (1987). Geometrizing rates of convergence I. Technical Report, Department of Statistics, University of California, Berkeley.

[4] Doss, H. and Sethuraman, J. (1989). The price of bias reduction when there is no unbiased estimate. *Ann. Statist.* **17**, 440–442. MR981462

[5] Liu, R. C. and Brown, L. D. (1993). Nonexistence of informative unbiased estimators in singular problems. *Ann. Statist.* **Vol. 21**, 1–13. MR1212163

[6] Low, M. (1995). Bias-Variance Tradeoffs in Functional Estimation Problems. *Ann. Statist.* **Vol. 23**, 824–835. MR1345202

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 218–227
© Institute of Mathematical Statistics, 2004

Combining correlated unbiased estimators of the mean of a normal distribution

Timothy Keller[1] and Ingram Olkin[2]

National Agricultural Statistics Service and Stanford University

Abstract: There are many applications in which one seeks to combine multiple estimators of the same parameter. If the constituent estimators are unbiased, then the fixed linear combination which is minimum variance unbiased is well-known, and may be written in terms of the covariance matrix of the constituent estimators. In general, the covariance matrix is unknown, and one computes a composite estimate of the unknown parameter with the covariance matrix replaced by its maximum likelihood estimator. The efficiency of this composite estimator relative to the constituent estimators has been investigated in the special case for which the constituent estimators are uncorrelated. For the general case in which the estimators are normally distributed and correlated, we give an explicit expression relating the variance of the composite estimator computed using the covariance matrix, and the variance of the composite estimator computed using the maximum likelihood estimate of the covariance matrix. This result suggests that the latter composite estimator may be useful in applications in which only a moderate sample size is available. Details of one such application are presented: combining estimates of agricultural yield obtained from multiple surveys into a single yield prediction.

1. Introduction

The need to combine estimators from different sources arises in many fields of application. In agriculture estimates may come from different experimental stations; in the medical sciences there may be multi-sites or multiple studies; sample surveys may contain subsurveys at different locations; several laboratories might assay a sample of one. Often making a prediction requires the combination of estimators. The present analysis was motivated by a model to predict agricultural yield. However, the model is generic, and occurs in a variety of contexts. The specifics of the application are discussed in Section 5.

It is perhaps surprising that the earliest methods for combining estimators were nonparametric. Fisher (1932) and Tippett(1931) proposed methods for combining p-values obtained from independent studies. Fisher was motivated by agriculture and Tippett by industrial engineering. These methods have been used to combine the results of independent studies in meta-analysis.

The parametric problem was first posed by Cochran (1937), who was also motivated by an agricultural problem. For simplicity suppose that we have two estimators T_1 and T_2 of θ from a $\mathcal{N}(\theta, \sigma_1^2)$ and $\mathcal{N}(\theta, \sigma_2^2)$ population, respectively. The

[1]U.S. Department of Agriculture, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030-1504, USA. e-mail: keller-timothy@hotmail.com
[2]Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA. e-mail: iolkin@stat.stanford.edu
Keywords and phrases: meta-analysis, unbiased estimators, correlated estimators.
AMS 2000 subject classifications: 62H12, 62H10.

combined estimator

$$T = w_1 T_1 + w_2 T_2 \tag{1.1}$$

with

$$w_1 = \sigma_1^{-2}/\left(\sigma_1^{-2} + \sigma_2^{-2}\right), w_2 = \sigma_2^{-2}/\left(\sigma_1^{-2} + \sigma_2^{-2}\right) \tag{1.2}$$

is unbiased and has variance

$$Var(T) = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \leq \min\left(\sigma_1^2, \sigma_2^2\right). \tag{1.3}$$

Consequently, the combined estimator dominates either single estimator in terms of having a smaller variance.

In practice the variances are unknown, and estimates $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ independent of T_1, T_2, are substituted in w_1 and w_2, that is,

$$T^* = \hat{w}_1 T_1 + \hat{w}_2 T_2. \tag{1.4}$$

Of course, now the variance of T^* is no longer minimum variance, but it is unbiased.

Cochran's paper was the genesis for a sequence of papers to study the effect of using estimates of the variances. We briefly describe these in chronological order. Graybill and Deal (1959) started with the Cochran model and assumed that the estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are independent and that each arises from a sample of size larger than 9. Under this condition, they show that T^* is uniformly better than either T_1 or T_2, where better means smaller variance.

Zacks (1966) starts with the assumption that the ratio $\rho = \sigma_2^2/\sigma_1^2$ is unknown but is estimable, and creates an estimator

$$T^{(1)} = (\hat{\rho} T_1 + T_2)/(\hat{\rho} + 1), \tag{1.5}$$

where $\hat{\rho}$ is independent of T_1 and T_2. Then $T^{(1)}$ is unbiased. The efficiency of $T^{(1)}$ cannot be given in closed form, and Zacks (1966) provides graphs of the efficiency relative to the estimator $T^{(1)}$ with ρ replacing $\hat{\rho}$.

Seshadri (1974), motivated by balanced incomplete block (BIB) design considerations, assumes that there is an unbiased estimator $\hat{b}$ of the ratio $b = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$, independent of T_1 and T_2 . Then the estimator

$$T^{(2)} = (1 - \hat{b}) T_1 + \hat{b} T_2 \tag{1.6}$$

is unbiased, and $var\ T^{(2)} \leq \min\ (var\, T_1, var\, T_2)$ provided Var $\hat{b} \leq b^2$ and Var $(1 - \hat{b}) \leq (1 - b)^2$. The key point is that in certain BIB designs there is an intra-block and inter-block estimator, and also an estimator $\hat{b}$.

When the sample sizes of the two samples are equal to n, Cohen and Sackrowitz (1974) discuss estimators of the form

$$T^{(3)} = \hat{\alpha}_1 T_1 + \hat{\alpha}_2 T_2, \tag{1.7}$$

where α_i are functions of sample variances and are chosen with respect to a squared error loss function normalized by σ_1^2. They determine the sample size n for which $T^{(3)}$ is superior to either T_1 or T_2.

Because the estimators T_i of the mean and s_i^2 of the variances are location and scale estimators, Cohen (1974) considers a location-scale family as a more general construct than the normal family. Again, the combined estimator is

$$T^{(4)} = \hat{b}_1 T_1 + \hat{b}_2 T_1, \quad \hat{b}_1 + \hat{b}_2 = 1, \tag{1.8}$$

where now $\hat{b}_2 = c\hat{\sigma}_1^2/(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$, c is a suitably chosen constant, and $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are appropriately chosen estimators.

The extension from combining two estimators to combining k estimators from k normal populations $\mathcal{N}(\theta, \sigma_i^2)$, $i = 1, \ldots, k$, is discussed by Norwood and Hinkelmann (1977). Here

$$T^{(5)} = \hat{w}_1 T_1 + \cdots + \hat{w}_k T_k \tag{1.9}$$

with $\hat{w}_i = \hat{\sigma}_i^{-2} / \sum_1^k \hat{\sigma}_j^{-2}$. They show that $var\,(T^{(5)}) \leq \min\{var\ T_i\}$ if each sample size is greater than 9, or if some sample size is equal to 9, and the others are greater than 17.

For the case $k = 2$ Nair (1980) computes the variance of T^* as an infinite series, as a function of two parameters, σ_1^2 and $\alpha = n_1\sigma_1^2/n_2\sigma_2^2$. Of course, it is symmetric and can be restated as a function of σ_2^2 and $1/\alpha$.

Following the formulation of Cohen and Sackrowitz (1974), Kubokawa (1987) provides a family of minimax estimators under normalized quadratic loss functions. Green and Strawderman (1991) also consider quadratic loss and provide a James–Stein shrinkage estimator. The use of a quadratic loss function is extended to the multivariate case by Loh (1991), where now we have normal populations $\mathcal{N}(\theta, \mathbf{\Sigma_1})$ and $\mathcal{N}(\theta, \mathbf{\Sigma_2})$. As in the univariate case, there are estimators $\hat{\theta}_1, \hat{\theta}_2$ of the mean vectors and independent covariance matrix estimators S_1, S_2, each having a Wishart distribution. For the loss function

$$L(\hat{\theta}, \theta, \Sigma_1, \Sigma_2) = (\hat{\theta} - \theta)'(\Sigma_1^{-1} + \Sigma_2^{-1})(\hat{\theta} - \theta), \tag{1.10}$$

with Σ_1 and Σ_2 known the estimator

$$\hat{\theta} = \left(S_1^{-1} + S_2^{-1}\right)^{-1}\left(\Sigma_1^{-1}\hat{\theta}_1 + \Sigma_2^{-1}\hat{\theta}_2\right) \tag{1.11}$$

is shown to be best linear unbiased.

The model that we here consider is that there are k normal populations $\mathcal{N}(\theta, \sigma_i^2)$, $i = 1, \ldots, k$. This model was considered by Halperin (1961) who provided an extensive analysis in which the estimator of θ is a weighted combination of the individual means,which are permittted to be correlated. For this model Halperin (1961) obtains the same variance as given in (2.8) below. In the present analysis the estimator of θ is a weighted combination of any unbiased estimators, and thereby permits somewhat more flexibility. Our derivation makes use of invariance arguments. In a later paper, Krishnamoorthy and Rohatgi (1990) show that the simple arithmetic mean is dominated by a shrinkage estimator that takes advantage of the variances.

2. The correlated case

As our starting point suppose that the data available are k unbiased estimators $T_1, \ldots, T_k$ of θ. However, the vector $T = (T_1, \ldots, T_k)$ has covariance matrix Σ, for which there is a sample covariance matrix S having a Wishart distribution $\mathcal{W}(\Sigma; k, n)$. Furthermore, S and $(T_1, \ldots T_k)$ are independent.

When Σ is known, the linear estimator

$$\hat{\theta} = w_1 T_1 + \cdots + w_k T_k, \qquad w_1 + \cdots + w_k = 1, \tag{2.1}$$

with $w_i, i = 1, \ldots, k$, fixed is unbiased. Let $w = (w_1, \ldots, w_k)'$ and $e = (1, \ldots, 1)'$. For the choice

$$w' = \left(e'\Sigma^{-1}\right)/\left(e'\Sigma^{-1}e\right), \tag{2.2}$$

$\hat{\theta}$ is also minimum variance unbiased. Furthermore,

$$Var(\hat{\theta}) = \frac{e'\Sigma^{-1}[\mathcal{E}(T-\theta e)'(T-\theta e)]\Sigma^{-1}e}{(e'\Sigma^{-1}e)^2} = \frac{1}{e'\Sigma^{-1}e}. \tag{2.3}$$

That $Var(\hat{\theta})$ is minimum variance follows from the Cauchy–Schwartz inequality:

$$\big(w'\Sigma w\big)\big(e'\Sigma^{-1}e\big) \geq \big(w'e\big)^2 = 1 \tag{2.4}$$

with equality if and only if (2.2) holds. Also,

$$\big(e'\Sigma^{-1}e\big)^{-1} \leq \min\big\{\sigma_1^2, \ldots, \sigma_k^2\big\}, \tag{2.5}$$

which follows from (2.4) with $w \equiv e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$.

When Σ is unknown it is estimated by S, and we consider the candidate estimator

$$\tilde{\theta} = \big(e'S^{-1}T\big)/\big(e'S^{-1}e\big). \tag{2.6}$$

The estimator $\tilde{\theta}$ is unbiased and has variance

$$\begin{aligned} Var(\tilde{\theta}) &= \mathcal{E}_S\mathcal{E}_T\frac{e'S^{-1}[(T-\theta e)'(T-\theta e)]S^{-1}e}{(e'S^{-1}e)^2} \\ &= \mathcal{E}_S\frac{e'S^{-1}\Sigma S^{-1}e}{(e'S^{-1}e)^2}. \end{aligned} \tag{2.7}$$

In the next section we provide a proof of the basic result:

$$Var(\tilde{\theta}) = \left(\frac{n-1}{n-k}\right) Var(\hat{\theta}). \tag{2.8}$$

3. Proof of the main result

The Wishart density of S is

$$f(S) = C(k,n) \mid \Sigma \mid^{-n/2} \mid S \mid^{\frac{(n-k-1)}{2}} \quad \exp\left(-\frac{1}{2}\, tr\, \Sigma^{-1}S\right),\ S > 0, \tag{3.1}$$

where

$$C(k,n) = \left\{2^{\frac{nk}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma\left(\frac{n-i+1}{2}\right)\right\}^{-1},$$

and $\Sigma > 0$ (that is, Σ is positive definite).

Let $Y = \Sigma^{-\frac{1}{2}}S\Sigma^{-\frac{1}{2}}$, so that the density of Y is

$$f(Y) = C(k,n) \mid Y \mid^{\frac{(n-k-1)}{2}} \exp\left(-\frac{1}{2}\, tr\, Y\right), \quad Y > 0. \tag{3.2}$$

With $b = \Sigma^{-\frac{1}{2}}e$

$$Var(\tilde{\theta}) = \mathcal{E}\left[\frac{b'Y^{-2}b}{(b'Y^{-1}b)^2}\right]. \tag{3.3}$$

Because the density (3.2) is orthogonally invariant, that is, $\mathcal{L}(G'YG) = \mathcal{L}(Y)$ for any orthogonal matrix G, a judicious choice of G allows one to put (3.3) in a more convenient form. Let $e_1 = (1, 0, \ldots, 0)'$, and choose G so that the first row

of G is $b'/\sqrt{b'b}$ and the remaining $k-1$ rows of G complete an orthonormal basis for G. Then, by construction, $Gb = \sqrt{b'b}\, e_1$. Consequently, with $Z = G'YG$, (3.3) becomes

$$Var(\tilde{\theta}) = \mathcal{E}\left[\frac{e_1' Z^{-2} e_1}{(e_1' Z^{-1} e_1)^2}\right] \frac{1}{b'b}.$$

Note that $b'b = e'\Sigma^{-1}e$, and recall that $Var(\hat{\theta}) = e'\Sigma^{-1}e$, so that

$$Var(\tilde{\theta}) = \mathcal{E}\left[\frac{e_1' Z^{-2} e_1}{(e_1' Z^{-1} e_1)^2}\right] Var(\hat{\theta}) \tag{3.4}$$

Remark. For any vector a of unit length, and a positive definite matrix B, $a'B^2a \geq (a'Ba)^2$. Hence (3.4) demonstrates that $Var(\tilde{\theta}) \geq Var(\hat{\theta})$ under the hypothesis that S and $T = (T_1, \ldots, T_k)'$ are independent, but with no distributional assumptions on S or T.

Now the task of proving the theorem is reduced to computing the expectation on the right side of equation (3.4). Towards that end, partition the $k \times k$ matrix Z and its inverse as

$$Z = \begin{pmatrix} z_{11} & z_1' \\ z_1 & Z_{22} \end{pmatrix}, \quad Z^{-1} = \begin{pmatrix} \tilde{z}_{11} & \tilde{z}_1' \\ \tilde{z}_1 & \tilde{Z}_{22} \end{pmatrix},$$

where Z_{22} and $\tilde{Z}_{22}$ are both $(k-1) \times (k-1)$.

In what follows we make use of well-known relationships between the blocks of Z and Z^{-1}. (See, for instance, Anderson, 2003.) Employing these relationships, and that $(I - uu')^{-1} = I + \frac{uu'}{1-u'u}$ the expression inside the expectation brackets in (3.4) can be written as:

$$\frac{e_1' Z^{-2} e_1}{(e_1' Z^{-1} e_1)^2} = \frac{\tilde{z}_{11}^2 + \tilde{z}_1'\tilde{z}_1}{\tilde{z}_{11}^2} = 1 + z_{11}\, u' Z_{22}^{-1} u, \tag{3.5}$$

where $u = Z_{22}^{-1/2} z_1/\sqrt{z_{11}}$; then (3.4) becomes:

$$Var(\tilde{\theta}) = \left[1 + E(z_{11}\, u' Z_{22}^{-1} u)\right] Var(\hat{\theta}). \tag{3.6}$$

The density of Z has the form (3.2), which can be written as

$$\begin{aligned} &f(Z_{22}, z_{11}, u) \\ &\quad = C(k,n)|Z_{22}|^{\frac{(n-k)}{2}} \exp\left(-\frac{1}{2} tr Z_{22}\right) z_{11}^{\frac{n}{2}-1} \exp\left(-\frac{1}{2} z_{11}\right)(1-u'u)^{\frac{(n-k-1)}{2}}. \end{aligned} \tag{3.7}$$

Again, using orthogonal invariance, the expectation in (3.6) is

$$\mathcal{E}\left[z_{11} u' Z_{22}^{-1} u\right] = C(k,n) I_1 I_2 I_3, \tag{3.8}$$

where

$$\begin{aligned} I_1 &= \int_0^\infty z_{11}^{\frac{n}{2}+1} \exp\left(-\frac{1}{2} z_{11}\right) dz_{11} = \Gamma\left(\frac{n+2}{2}\right) 2^{\frac{n+2}{2}}, \\ I_2 &= \int_{u'u<_1} u'u(1-u'u)^{\frac{n-k-1}{2}} du = (k-1)\pi^{\frac{k-1}{2}} \Gamma\left(\frac{n-k+1}{2}\right) / 2\Gamma\left(\frac{n+2}{2}\right), \\ I_3 &= \int_{Z_{22}>0} \left(e_1' Z_{22}^{-1} e\right) \mid Z_{22} \mid^{\frac{n-k}{2}} \exp\left(-\frac{1}{2} tr Z_{22}\right) dZ_{22}. \end{aligned}$$

The integral I_2 can be evaluated using polar coordinates; it is also a Dirichlet Integral of Type-I. (See Sobel, Uppuluri and Frankowski, 1977). To simplify notation in I_3 let $Q = Z_{22}$, so that Q is a $(k-1)\times(k-1)$ matrix having a Wishart distributon $\mathcal{W}(I; k-1, n)$. Then $I_3 = \mathcal{E}(Q^{-1})_{11}/C(k-1, n)$. But this expectation is known (see e.g. Kshirsagar, 1978, p. 72) so that

$$I_3 = \left[(n-k)2^{\frac{n}{2}}\,\pi^{\frac{k-1}{2}}\,\Gamma\left(\frac{n-k+1}{2}\right)\right]^{-1}. \tag{3.9}$$

Combining these results we obtain

$$Var(\tilde{\theta}) = (1 + I_1 I_2 I_3)\, Var(\hat{\theta}) = \frac{n-1}{n-k}\, Var(\hat{\theta}). \tag{3.10}$$

4. Discussion of efficiency for $k = 2$ and $n = N - 1$

The result that $Var\ (\tilde{\theta}) = \frac{n-1}{n-k}\ Var\ (\hat{\theta})$ coincides with what intuition suggests: when $k = 1$, $Var\ (\tilde{\theta}) = Var\ (\hat{\theta})$; when $k > 1$, $Var\ (\tilde{\theta}) > Var\ (\hat{\theta})$, and for all k, $lim_N \to \infty\ Var\ (\tilde{\theta}) = Var\ (\hat{\theta})$. But the result gives more precise information that helps one to assess the efficiency of the Graybill-Deal estimator for a given sample size.

Consider the case $k = 2$, $N = n-1$. If, without loss of generality, we take $\sigma_{11} = \min\{\sigma_{11}, \sigma_{22}\}$, then Var $(\tilde{\theta}) \leq \min\,(\sigma_{11}, \sigma_{22})$ when

$$\frac{1}{N-3} \leq \frac{(\sigma_{11} - \sigma_{12})^2}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}. \tag{4.1}$$

In the special case for which cov $(T_1, T_2) = 0$, (4.1) is $1/(N-3) \leq \sigma_{11}/\sigma_{22} \leq 1$, which implies that $Var(\tilde{\theta}) < \min\,(\sigma_{11}, \sigma_{22})$ for all $N \geq 5$. Note that this does not contradict the previously quoted result of Graybill and Deal (1959); their hypothesis allows N_1 and N_2, the sample sizes for the respective constituent estimators, to be unequal; whereas the current theorem was derived under the assumption that $N_1 = N_2 = N$. When T_1 and T_2 are uncorrelated, there are corresponding sample sizes N_1 and N_2 used in estimating the variances. However, when the T's are correlated, the covariance matrix is estimated from a single sample of size N.

Writing $\sigma_{11} = \alpha^2\sigma_{22}, 0 \leq \alpha \leq 1$, and denoting the correlation between T_1 and T_2 by ρ, (4.1) can be written as

$$\frac{1}{N-3} \leq \frac{(\alpha - \rho)^2}{1 - \rho^2}. \tag{4.2}$$

Then it is apparent that if one varies the parameters α and ρ so that $\alpha - \rho \to 0$, the sample size N necessary for (4.2) to hold increases without bound. But this also is intuitive: $\alpha - \rho \to 0$ is equivalent to $\hat{\theta} \to T_1$. Given a rough initial estimate for the parameters α and ρ, one may use (4.2) to obtain some idea whether the Graybill-Deal estimator dominates the better of the two constituent estimators for a given sample size.

Taking the special case $\sigma_{11} = \sigma_{22}$, (4.2) becomes

$$\frac{1}{N-3} \leq \frac{1-\rho}{1+\rho}.$$

This form of equation (4.1) implies that the sample size for (4.1) to hold increases without bound as $\rho \to 1$. Once again, this is intuitive: to say ρ is close to 1 means the estimator T_2 provides essentially the same information about θ as the estimator T_1, and hence the composite estimator cannot be expected to provide much more information than that provided by T_1 alone.

5. An agricultural application: Forecasting yield

The National Agricultural Statistics Service (NASS), an agency of the USDA, makes monthly pre-harvest yield forecasts for the major US agricultural commodities at several geographic levels. In the final analysis, the official forecast of yield announced to the public is necessarily the result of a mixed process of both objective scientific technique and subjective expert judgment. Nevertheless, subjective expert judgement is most effective when it has an objective estimate of yield with which to commence its operation. Given an historical data series for the most important estimators of yield, and the corresponding published *final* yield, one can estimate the covariance structure and biases for those estimators. These are then the basis for computing a composite estimate of yield. The question of how best to use historical data to estimate the biases in the constituent estimators of yield is important in itself. In order to avoid a long digression, we pick up the discussion of the application at the point where a 'bias correction' has already been applied to the historical data; hence only the problem of estimating the covariance matrix remains.

Table 1 presents the predicted yield based on a biological yield model (T_1) and the predicted yield based on a survey of producer expectations (T_2). These data have been masked for security considerations. Make the following assumptions:

(1) The true yield β_i for year i is the yield published by NASS (Table 2) at the end of the growing season.

(2) T_1 and T_2 are independent.

(3) The covariance matrix is essentially constant over time.

Under these assumptions the maximum likelihood estimator for the covariance matrix based on the data in Table 1 is:

Table 1: Predicted yields (weight per area) of commodity Z for state X in month Y.

Year	Survey of biological yield	
1	88.0	87.5
2	82.5	80.0
3	83.0	86.5
4	73.5	79.0
5	79.0	84.5
6	82.0	83.5
7	83.0	79.8
8	80.8	84.0
9	81.0	83.0
10	79.0	79.0
11	64.0	76.0
12	80.5	83.8
13	83.0	87.0
14	81.5	78.5

$$S = \begin{pmatrix} 9.50 & 2.19 \\ 2.19 & 15.30 \end{pmatrix},$$

and the vector of weights for the linear combination of T_1 and T_2 which is the Graybill-Deal estimator of yield is $w' = (0.642, 0.358)$.

A word about the operational implementation of these ideas is in order. It is unreasonable to expect that the assumptions underlying the estimate of the covariance matrix hold for all time; hence, in practice, one envisions that yield data from a 'moving window' of N past years would be used to estimate the vector of coefficients, w, used to compute the composite estimate of yield for the *current* year. This concept has been tested by a cross-validation scheme in which each of $N+1$ years is sequentially treated as the 'current' year, and the remaining N years are treated as the 'past', where $N+1$ is the length of the relevant data series which is available; but, for the sake of a simple exposition, the calculations presented in Table 2 are based on all 14 years of data at once, the results of the cross-validation shceme being very similar.

Looking at Table 2, one notes that the root mean square error for the composite estimator was less than that of either constituent estimator of yield, and only slightly larger than the root mean square error for the yield forecast produced by the panel of commodity experts. Given that this panel was privy to a great many sources of information relevant to setting yield, in *addition* to the constituent estimators of yield, this is a remarkable result. One cannot hope to replace expert judgement with statistical methodology; nevertheless, these results demonstrate

Table 2:

Year	Composite Estimate ($\hat{\theta}$)	Panel of Experts	Final Published Yield (θ)
1	87.8	89.5	87.8
2	81.5	82.5	87.3
3	84.2	85.8	85.3
4	75.3	76.3	76.8
5	81.3	83.3	78.3
6	82.5	83.8	89.0
7	81.8	85.0	82.5
8	81.8	81.3	84.0
9	81.7	81.8	82.3
10	79.0	81.0	80.8
11	68.3	67.5	68.3
12	81.6	83.0	83.0
13	84.4	85.0	85.0
14	80.4	82.0	81.8

Root Mean Square Error:

Farmer Reported Yield	3.06
Biological Yield Model	3.92
Composite Estimator	2.68
Panel of Experts	2.58

that the techniques of composite estimation can provide a useful starting point for the overall process of setting a yield forecast.

Acknowledgement

The authors are grateful to Brian Cullis for many helpful comments and insights relating to the paper by Halperin (1961), and to the referee for helpful suggestions.

References

[1] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edition. John Wiley and Sons, New York. MR1990662

[2] Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments, Supplement to the *Journal of the Royal Statistical Society* **4** 102–118.

[3] Cohen, A. (1976). Combining estimates of location, *Journal of the American Statistical Association* **71** 172–175. MR426258

[4] Cohen, A. and Sackrowitz, H. B. (1974). On estimating the common mean of two normal distributions. *Annals of Statistics* **2** 1274–1282. MR365851

[5] Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.) Oliver and Boyd, London.

[6] Graybill, F. A. and Deal, R. B. (1959). Combining unbiased estimators, *Biometrics* **15** 543–550. MR107925

[7] Green, E. J. and Strawderman, W. E. (1991). A James–Stein type estimator for combining unbiased and possibly biased estimators, *Journal of the American Statistical Association* **86** 1001–1006. MR1146348

[8] Halperin, M. (1961). Almost linearly-optimum combination of unbiased estimates. *Journal of the American Statistical Association* **56** 36–43. MR141181

[9] Krishnamoorthy, K. and Rohatgi, V. K. (1990). Unbiased estimation of the common mean of a multivariate normal distribution. *Communications in Statistics – Theory and Methods* **19** (5) 1803–1810. MR1075503

[10] Kshirsagar, A. (1978). *Multivariate Analysis*, Marcel Dekker, Inc., New York. MR343478

[11] Kubokawa, T. (1987). Admissible minimax estimation of a common mean of two normal populations. *Annals of Statistics*, **15** 1245–1256. MR902256

[12] Loh, W.-L. (1991). Estimating the common mean of two multivariate normal distributions, *Annals of Statistics*, **19** 297–313. MR1091852

[13] Nair, K. A. (1980) Variance and distribution of the Graybill-Deal estimator of the common mean of two normal populations, *Annals of Statistics* **8** 212–216. MR557567

[14] Norwood, T. E. and Hinkelmann, K. Jr. (1977). Estimating the common mean of several normal populations, *Annals of Statistics* **5** 1047–1050. MR448679

[15] Raj, D. (1998). *Sampling Theory*, McGraw-Hill, New York. MR230440

[16] Seshadri, V. (1963). Constructing uniformly better estimators. *Journal of the American Statistical Association* **58** 172–175. MR145628

[17] Sobel, M., Uppuluri, R., and Frankowski, K. (1977). *Selected tables in mathematical statistics*, Vol. 4: Dirichlet Distribution – Type 1, American Mathematical Society, Providence, Rhode Island. MR423747

[18] Tippett, L. H. C. (1931). *The method of statistics*. Williams and Norgate, London.

[19] Zacks, S. (1966). Unbiased estimation of the common mean of two normal distributions based on small samples of equal size, *Journal of the American Statistical Association* **61** 467–476. MR207100

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 228–236
© Institute of Mathematical Statistics, 2004

An asymptotic minimax determination of the initial sample size in a two-stage sequential procedure

Michael Woodroofe[1]

University of Michigan

Abstract: When estimating the mean of a normal distribution with squared error loss and a cost for each observation, the optimal (fixed) sample size depends on the variance σ^2. A two-stage sequential procedure is to first conduct a pilot study from which σ^2 can be estimated, and then estimate the desired sample size. Here an asymptotic formula for the initial sample size in a two-stage sequential estimation procedure is derived–asymptotic as the cost of a single observation becomes small compared to the loss from estimation error. The experimenter must specify only the sample size, n_0 say, that would be used in a fixed sample size experiment; the initial sample size of the two-stage procedure is then the least integer greater than or equal to $\sqrt{n_0/2}$. The resulting procedure is shown to minimize the maximum Bayes regret, where the maximum is taken over prior distributions that are consistent with the initial guess n_0; and the minimax solution is shown to provide an asymptotic lower bound for optimal Bayesian choices for a wide class of prior distributions.

1. Introduction

It is indeed a pleasure to offer this tribute to Herman Rubin and to ponder his influence on my own work. I still remember the interest with which I read the papers on Bayes' risk efficiency [7] and [8] early in my career. From reading these papers, I gained an appreciation for the power of statistical decision theory and its interplay with asymptotic calculations that go beyond limiting distributions. These involved moderate deviations in the case of [7]. A central idea in [8] is the study of a risk functional, the integrated risk of a procedure with respect to a prior distribution that can vary over a large class. I have used this idea in a modified form in work on sequential point estimation and very weak expansions for sequential confidence intervals—[12, 13, 14], and the references given there. This idea is also present in Theorem 2 below. The connection between [12] and Bayes risk efficiency is notable here. The following is proved in [12], though not isolated: Suppose that it is required to estimate the mean of an exponential family with squared error loss and a cost for each observation and that the population mean is to be estimated by the sample mean. Then there is a stopping time which is Bayes risk non-deficient in the sense of [4]; that is, it minimizes a Bayesian regret asymptotically, simultaneously for all sufficiently smooth prior distributions.

The present effort combines tools from decision theory and asymptotic analysis to obtain a rule for prescribing the initial sample size in a two-stage sequential procedure for estimating the mean of a normal distribution. Unlike the fully sequential, or even three-stage, versions of the problem, Bayes risk non-deficiency is

[1]Statistics Department, University of Michigan, Ann Arbor, Michigan 48109, USA. e-mail: michaelw@stat.lsa.umich.edu

Keywords and phrases: Bayesian solutions, integrated risk and regret, inverted gamma priors, sequential point estimation, squared error loss.

AMS 2000 subject classifications: 62L12.

not possible with two-stage procedures, and the rule is obtained from minimaxity. The problem is stated in Section 2, and the minimax solution is defined. The rule requires the statistician to specify only the fixed sample size, n_0 say, that would have been used in a fixed sample size design, or to elicit such from a client. The minimax initial sample size is then the least integer that is greater than or equal to $\sqrt{n_0/2}$. The proof of asymptotic minimaxity is provided in Section 3. As explained in Section 4, the minimax solution is very conservative but, at least, provides an asymptotic lower bound for optimal Bayesian solutions for a wide class of prior distributions.

2. The problem

Let $X_1, X_2, \ldots \overset{ind}{\sim} \text{Normal}[\mu, \sigma^2]$, where $-\infty < \mu < \infty$ and $\sigma > 0$ are unknown, and consider the problem of estimating μ with loss of the form

$$L_a(n) = a^2\big(\bar{X}_n - \mu\big)^2 + n, \tag{1}$$

where $\bar{X}_n = (X_1 + \cdots + X_n)/n$. In (1), $a^2(\bar{X}_n - \mu)^2$ represents the loss due to estimation error, and n the cost of sampling. The units are so chosen that each observation costs one unit, and a is determined by the importance of estimation error relative to the cost of sampling. Also, the estimator has been specified as $\bar{X}_n$, leaving only the sample size n to be determined. If σ were known, then the expected loss plus sampling cost, $E_{\mu,\sigma^2}[L_a(n)] = a^2\sigma^2/n + n$, would be minimized when n is an integer adjacent to

$$N = a\sigma,$$

and in many ways the problem is one of estimating N. This will be done using the sample variances

$$S_n^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\big(X_i - \bar{X}_n\big)^2$$

for $n \geq 2$. Interest in two-stage sequential procedures for estimation originated with Stein's famous paper [9]. The problem has a long history, much of which is included in Chapter 6 of [5], but there seems to be no general agreement on the choice of the initial sample size m in two-stage procedures. Some additional references are provided in the last section.

A *two-stage procedure* consists of a pair $\delta = (m, \tilde{N})$ where $m \geq 2$ is an integer and $\tilde{N} = \tilde{N}(S_m^2)$ is an integer valued random variable for which $\tilde{N} \geq m$. The estimator of μ is then $\bar{X}_{\tilde{N}}$. For example, letting $\lceil x \rceil$ denote the least integer that is at least x,

$$\hat{N}_a = \max\big\{m, \lceil aS_m \rceil\big\} \tag{2}$$

satisfies the conditions for any $m \geq 2$. The choice of m has to be subjective at some level, because there is no data available when it is chosen. Here it is required only that the experimenter specify a prior guess, u say for σ, or even just the guess $n_0 = au$ for N. This seems a very modest requirement, since a fixed sample size experiment would have to include a prior guess for N. Given the prior guess, it is shown that

$$m_a = \max\left\{2, \left\lceil \sqrt{\frac{1}{2}n_0} \right\rceil\right\} \tag{3}$$

leads to a procedure that minimizes the maximum Bayes' regret in the class of prior distributions for which σ has prior mean u.

3. The theorem

The risk of a two stage procedure $\delta = (m, \tilde{N})$ is $R_a(\delta; \sigma^2) = E_{\mu,\sigma^2}[L_a(\tilde{N})]$. Using the Helmut transformation (for example, [11, p. 106]), it is easily seen that

$$R_a\big(\delta; \sigma^2\big) = E_{\sigma^2}\left[\frac{a^2\sigma^2}{\tilde{N}} + \tilde{N}\right], \tag{4}$$

which depends on σ^2, but not on μ. The difference

$$r_a\big(\delta, \sigma^2\big) = E_{\sigma^2}\left[\frac{a^2\sigma^2}{\tilde{N}} + \tilde{N}\right] - 2N,$$

is called the regret.

If ξ is a prior distribution over $[0, \infty)$, write P_ξ and E_ξ for probability and expectation in the Bayesian model, where $\sigma^2 \sim \xi$ and $S_2^2, S_3^2, \ldots$ are jointly distributed random variables; and write P_ξ^m and E_ξ^m for conditional probability and expectation given S_m^2. Then the integrated risk of a two-stage procedure δ with respect to a ξ is

$$\bar{R}_a(\delta, \xi) = \int_0^\infty R_a\big(\delta; \sigma^2\big)\xi\big\{d\sigma^2\big\} = E_\xi\left[\frac{a^2\sigma^2}{\tilde{N}} + \tilde{N}\right]$$

possibly infinite; and if $\int_0^\infty \sigma\xi\{d\sigma^2\} < \infty$, then the integrated regret of δ with respect to ξ is

$$\bar{r}(\delta, \xi) = \int_0^\infty r_a\big(\delta; \sigma^2\big)\xi\big\{d\sigma^2\big\} = E_\xi\left[\frac{a^2\sigma^2}{\tilde{N}} + \tilde{N} - 2N\right]$$

again possibly infinite. As noted above, the experimenter must specify $E_\xi(N)$, or equivalently, $E_\xi(\sigma)$. In fact, it is sufficient for the experimenter to specify an upper bound. For a fixed $u \in (0, \infty)$, let $\Xi = \Xi_u$ be the class of prior distributions for which

$$\int_0^\infty \sigma\xi\big\{d\sigma^2\big\} \le u; \tag{5}$$

and let $\Xi^o = \Xi_u^o$ be the class of ξ for which there is equality in (5). Also, let δ^a be the procedure $(m_a, \hat{N}_a)$ defined by (2) and (3) with $n_0 = au$.

Theorem 1. *For any $u \in (0, \infty)$.*

$$\inf_\delta \sup_{\xi\in\Xi} \bar{r}(\delta; \xi) \sim \sqrt{2n_0} \sim \sup_{\xi\in\Xi} \bar{r}\big(\delta^a; \xi\big)$$

as $a \to \infty$.

Proof. The proof will consist of showing first that

$$\limsup_{a\to\infty} \sup_{\xi\in\Xi} \frac{1}{\sqrt{a}}\bar{r}\big(\delta^a; \xi\big) \le \sqrt{2u} \tag{6}$$

and then that

$$\liminf_{a\to\infty} \sup_{\xi\in\Xi^o} \inf_\delta \frac{1}{\sqrt{a}}\bar{r}(\delta; \xi) \ge \sqrt{2u} \tag{7}$$

This is sufficient, since $\inf_\delta \sup_{\xi\in\Xi} \ge \sup_{\xi\in\Xi^o} \inf_\delta$. In the proofs of (6) and (7), there is no loss of generality in supposing that $u = 1$.

The Upper Bound. From (4) and (2),

$$R_a\big(\delta^a;\sigma^2\big) \le a\sigma^2 E_{\sigma^2}\left[\frac{1}{S_{m_a}}\right] + aE_{\sigma^2}(S_{m_a}) + m_a + 1. \tag{8}$$

Here

$$E_{\sigma^2}(S_m) = C(m)\sigma, \tag{9}$$

where

$$C(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m-1}{2}}\Gamma(\frac{m-1}{2})}$$

and Γ is the Gamma-function; and, similarly,

$$E\left(\frac{1}{S_m}\right) = \sqrt{\frac{m-1}{m-2}}\frac{1}{C(m-1)\sigma}. \tag{10}$$

A version of Stirling's Formula asserts that

$$\log\Gamma(z) = \left(z-\frac{1}{2}\right)\log(z) - z + \frac{1}{2}\log(2\pi) + \frac{1}{12z} + O\left(\frac{1}{z^3}\right)$$

as $z \to \infty$. See, for example, [1, p. 257]. It then follows from simple algebra that

$$C(m) = 1 - \frac{1}{4m} + O\left(\frac{1}{m^2}\right). \tag{11}$$

Let a be so large that $m_a \ge 3$. Then, combining (8) and (11),

$$\begin{aligned} R_a\big(\delta^a;\sigma^2\big) &\le a\sigma\left[\sqrt{\frac{m_a-1}{m_a-2}}\frac{1}{C(m_a-1)} + C(m_a)\right] + m_a + 1 \\ &= 2a\sigma + \frac{a\sigma}{2m_a} + m_a + 1 + a\sigma \times O\left(\frac{1}{m_a^2}\right), \end{aligned}$$

where $O(1/m)$ is a function only of m. So, for every $\xi \in \Xi = \Xi_1$,

$$\bar{r}_a(\delta^a;\xi) \le \frac{a}{2m_a} + m_a + 1 + a \times O\left(\frac{1}{m_a^2}\right) \le \sqrt{2a} + O(1),$$

establishing (6), since $n_0 = 2a$ when $u = 1$.

The Lower Bound. The lower bound (7) will be established by finding the Bayes procedure and a lower bound for the Bayes regret

$$\bar{r}_a(\xi) = \inf_\delta \bar{r}_a(\delta;\xi)$$

for a general prior distribution ξ and then finding priors $\xi_a \in \Xi^o$ for which $\liminf_{a\to\infty} \bar{r}_a(\xi_a)/\sqrt{a} \ge \sqrt{2}$.

Finding the Bayes procedure is not difficult. If the initial sample size is $m \ge 2$, then $\tilde{N}$ should be chosen to minimize the posterior expected loss $E_\xi^m[a^2\sigma^2/n + n]$ with respect to n. Clearly,

$$E_\xi^m\left[\frac{a^2\sigma^2}{n} + n\right] = \frac{a^2 V_m}{n} + n = 2a\sqrt{V_m} + \frac{1}{n}\big(n - a\sqrt{V_m}\big)^2 \tag{12}$$

where

$$V_m = E_\xi^m\big(\sigma^2\big)$$

So, (12) is minimized when n is the larger of m and an integer adjacent to $a\sqrt{V_m}$, leaving

$$\bar r_a(\xi) = \inf_{m\ge 2} E_\xi\left\{2a\sqrt{V_m} + \frac{1}{m}\big(m - a\sqrt{V_m}\,\big)_+^2 + \eta(a,m)\right\} - 2a,$$

where $(x)_+^2$ denotes the square of the positive part of x and $0 \le \eta(a,m) \le 1/m$. An alternative expression is

$$\bar r_a(\xi) = \inf_{m\ge 2} E_\xi\left\{2a\big[\sqrt{V_m} - U_m\big] + \frac{1}{m}\big(m - a\sqrt{V_m}\,\big)_+^2 + \eta(a,m)\right\}, \tag{13}$$

where

$$U_m = E_\xi^m(\sigma)$$

and $E_\xi(U_m) = E_\xi(\sigma) = 1$.

Suppose now that ξ is an inverted Gamma prior with density

$$\frac{1}{\Gamma(\frac{1}{2}\alpha)}\left(\frac{\beta}{2\sigma^2}\right)^{\frac{1}{2}\alpha}\exp\left[-\frac{\beta}{2\sigma^2}\right]\frac{1}{\sigma^2}, \tag{14}$$

where $\alpha > 1$ and $\beta > 0$. Equivalently $1/\sigma^2$ has a Gamma distribution with parameters $\alpha/2$ and $\beta/2$. Then

$$E(\sigma) = \frac{\Gamma(\frac{\alpha-1}{2})}{\Gamma(\frac{1}{2}\alpha)}\sqrt{\frac{\beta}{2}}. \tag{15}$$

Letting

$$W_m = (m-1)S_m^2$$

and applying (15) to the posterior distributions then leads to

$$U_m = \frac{\Gamma(\frac{\alpha+m-2}{2})}{\Gamma(\frac{\alpha+m-1}{2})}\sqrt{\frac{\beta + W_m}{2}}$$

and

$$V_m = E_\xi^m\big(\sigma^2\big) = \frac{\beta + W_m}{\alpha + m - 3} = B(\alpha+m-1)^2 \times U_m^2, \tag{16}$$

where

$$B(m) = \sqrt{\frac{m-1}{m-2}}C(m) = 1 + \frac{1}{4m} + O\left(\frac{1}{m^2}\right). \tag{17}$$

In order for the ξ of (14) to be in $\Xi^o = \Xi_1^o$, α and β must be so constrained that the right side of (15) equals one. Then $E_\xi(U_m) = 1$, $E_\xi(\sqrt{V_m}) = B(\alpha+m-1)$, and

$$\begin{aligned}\bar r_a(\xi) &= \inf_{m\ge 2} E_\xi\left\{2a\big[B(\alpha+m-1)-1\big]U_m + \frac{1}{m}\big(m - a\sqrt{V_m}\,\big)_+^2 + \eta(a,m)\right\}\\ &\ge \inf_{m\ge 2}\left\{2a\big[B(\alpha+m-1)-1\big] + (1-\epsilon)^2 m P_\xi\big[a\sqrt{V_m} \le \epsilon m\,\big]\right\}\end{aligned}$$

for any $\epsilon > 0$.

Observe that $B(\alpha+m-1)$ is positive and bounded away from 0 for $0<\alpha\le 1$ for each fixed $m\ge 2$. It follows that the term in braces on the right side of (13) is of order a for each fixed $m\ge 2$ when ξ is an inverted gamma prior with $0<\alpha\le 2$ and, therefore, that the minimizing $m=m_a$ approaches ∞ as $a\to\infty$. So, $\inf_{m\ge 2}$ in (13) can be replaced by $\inf_{m\ge m_0}$ for any m_0 for all sufficiently large a.

The marginal distribution of W_m is of the form

$$P_\xi[W_m\le w]=\int_0^w \frac{1}{\beta}g\Big(\frac{z}{\beta}\Big)\,dz,$$

where

$$g(z)=\frac{\Gamma(\frac{\alpha+m-1}{2})}{\Gamma(\frac{\alpha}{2})\Gamma(\frac{m-1}{2})}\frac{z^{\frac{m-3}{2}}}{(1+z)^{\frac{\alpha+m-1}{2}}}.$$

Clearly,

$$\begin{aligned}\int_c^\infty g(z)dz &\le \frac{\Gamma(\frac{\alpha+m-1}{2})}{\Gamma(\frac{\alpha}{2})\Gamma(\frac{m-1}{2})}\int_c^\infty\Big(\frac{1}{1+z}\Big)^{1+\frac{1}{2}\alpha}dz\\ &= \frac{2\Gamma(\frac{\alpha+m-1}{2})}{\alpha\Gamma(\frac{\alpha}{2})\Gamma(\frac{m-1}{2})}\Big(\frac{1}{1+c}\Big)^{\frac{1}{2}\alpha}\end{aligned}$$

for all $c>0$. So, there is a constant K for which

$$\int_c^\infty g(z)dz\le\frac{Km}{\sqrt{1+c}}$$

for all $1<\alpha\le 2$, $m\ge 2$ and $c>0$. Let ξ_a be an inverted gamma prior with $\beta_a=o(a^{-2})$ and α_a determined by the condition that $E_{\xi_a}(\sigma)=1$. Then $\alpha_a\to 1$ as $a\to\infty$. If $\epsilon>0$ is given, then

$$\begin{aligned}P_{\xi_a}\big[a\sqrt{V}_m\ge\epsilon m\big] &= P_{\xi_a}\Big[W_m\ge\epsilon^2\frac{m^2(m-2)}{a^2}-\beta_a\Big]\\ &\le \frac{2Ka}{\epsilon}\sqrt{\beta_a}\le\epsilon\end{aligned}$$

for all $m\ge 3$ and sufficiently large a. It follows that for any $m_0\ge 3$,

$$\bar r(\xi_a)\ge\inf_{m\ge 2}\Big\{2a\big[B(\alpha_a+m-1)-1\big]+(1-\epsilon)^3m\Big\}. \tag{18}$$

for all sufficiently large a. From (11) and (17) there is an m_0 for which $B(m)\ge l+(1-\epsilon)/4m$ for all $m\ge m_0-1$. Then

$$\bar r(\xi_a)\ge(1-\epsilon)\inf_{m\ge m_0}\Big[\frac{a}{2m}+(1-\epsilon)^2m\Big]\ge(1-\epsilon)^2\sqrt{2a}$$

for all sufficiently large a. Relation (7) follows since $\epsilon>0$ was arbitrary. □

4. The minimax solution as a lower bound

The minimax solution is very conservative in that it specifies a very small initial sample size. For example, if the prior guess for the best fixed sample size is 100, then the asymptotic minimax solution calls for an initial sample size of only 8; and if the prior guess is increased to 1000, then the initial sample size increases only

to 23. The asymptotic minimax solution approximates the Bayes procedure when the σ is small with high probability, but still has a fixed mean, as is clear from the nature of the inverted gamma prior that was used to obtain the lower bound. A statistician who can specify more about the prior distribution will take a larger initial sample size for large a and incur a smaller regret. For example, if $\sigma \geq \sigma_0 > 0$ with prior probability one, then the optimal initial sample size is at least $a\sigma_0$, and the Bayes regret is of order one as $a \to \infty$, assuming that σ has a finite prior mean. A more detailed study of the asymptotic properties of Bayes procedures suggests that optimal m is closely related to the behavior of the prior density near $\sigma^2 = 0$, a relationship that might be difficult to specify or elicit from a client. The inverted gamma priors (14) are an extreme case since the prior density approaches zero very rapidly as $\sigma^2 \to 0$ in this case. An advantage of the asymptotic minimax solution, of course, is that it does not require the statistician to elicit detailed prior information from a client.

The following result shows that the asymptotic minimax solution (3) provides an asymptotic lower bound for optimal Bayesian solutions for a very large class of prior distributions.

Theorem 2. *Suppose that $\xi\{0\} = 0$, that ξ has a continuously differentiable density on $(0, \infty)$, and that $\int_0^\infty \sigma^2 \xi\{d\sigma^2\} < \infty$. Let $m_a^* = m_a^*(\xi)$ be an optimal initial sample size for ξ. Then*

$$\lim_{a\to\infty} \frac{m_a^*}{\sqrt{a}} = \infty. \tag{19}$$

Proof. As above, there is no loss of generality in supposing that $\int_0^\infty \sigma\xi\{d\sigma^2\} = 1$. By (13),

$$\bar{r}_a(\xi) = \inf_{m\geq 2} \left[2ab_m + c_m(a) + \eta(a, m)\right],$$

where $b_m = E_\xi[\sqrt{V_m} - U_m]$, $c_m(a) = E_\xi[(m - a\sqrt{V_m})_+^2]/m$, and $0 \leq \eta(a, m) \leq 1/m$. Then

$$2a[b_{m_a^*} - b_{2m_a^*}] \leq c_{2m_a^*}(a) + \frac{1}{2m_a^*},$$

since $2ab_m + c_m(a) + \eta(a, m)$ is minimized when $m = m_a^*$ and $0 < \eta(a, m) \leq 1/m$. By Lemmas 1 and 2 below,

$$c_m(a) \leq mP_\xi\left[\sigma \leq \frac{m}{a}\right] \tag{20}$$

and

$$b_m - b_{2m} \geq \frac{\epsilon}{m} \tag{21}$$

for some $\epsilon > 0$ that does not depend on m. Combining the last three equations,

$$\frac{2a\epsilon}{m_a^*} \leq 2m_a^* P_\xi\left[\sigma \leq \frac{2m_a^*}{a}\right] + \frac{1}{2m_a^*}$$

and, therefore,

$$\frac{m_a^*}{\sqrt{a}} \geq \sqrt{\frac{\epsilon}{2P_\xi[\sigma \leq 2m_a^*/a]}},$$

for all sufficiently large a. Relation (19) follows directly, completing the proof, except for the proofs of the lemmas. □

Lemma 1. *Relation* (20) *holds.*

Proof. Using Jensen's Inequality twice, $(m-a\sqrt{V_m})_+^2 \le [E_\xi^m(m-a\sigma)]_+^2 \le E_\xi^m[(m-a\sigma)_+^2]$. So,

$$c_m(a) \le \frac{1}{m} E_\xi\big[(m-a\sigma)_+^2\big] \le m P_\xi\left[\sigma \le \frac{m}{a}\right],$$

as asserted. □

Lemma 2. *There is an $\epsilon > 0$ for which relation* (21) *holds.*

Proof. Since $E_\xi(U_m) = E_\xi(\sigma)$ for all m, $b_m - b_{2m} = E_\xi[\sqrt{V_m} - \sqrt{V_{2m}}]$. Next, since $V_m - V_{2m} = 2\sqrt{V_m}(\sqrt{V_m} - \sqrt{V_{2m}}) - (\sqrt{V_{2m}} - \sqrt{V_m})^2$ and $E_\xi^m(V_{2m} - V_m) = 0$,

$$b_m - b_{2m} = E_\xi\left[\frac{(\sqrt{V_{2m}} - \sqrt{V_m})^2}{2\sqrt{V_m}}\right].$$

From Laplace's method, for example, [6],

$$V_m = S_m^2 + O\left(\frac{1}{m}\right)$$

*w.p.*1 (P_{σ^2}) for each $\sigma^2 > 0$ and, therefore, *w.p.*1 (P_ξ). That $\sqrt{m}(\sqrt{V_{2m}} - \sqrt{V_m})$ has a non-degenerated limiting distribution follows directly, and then

$$\liminf_{m\to\infty} m E_\xi\left[\frac{(\sqrt{V_{2m}} - \sqrt{V_m})^2}{2\sqrt{V_m}}\right] > 0$$

by Fatou's Lemma. Relation (21) follows. □

5. Remarks and acknowledgments

The smoothness condition on the prior in Theorem 2 can probably be relaxed. In the proof, it was used to derive the relation $V_m - S_m^2 = O(1/m)$ *w.p.*1, and this is a smaller order of magnitude that is needed.

If ξ is an inverted gamma prior with a fixed $\alpha > 1$ and $\beta > 0$, then

$$r_a(\xi) + \frac{a}{m_a^*(\xi)} = O\big[\sqrt{\log(a)}\,\big].$$

This may be established by combining techniques from the proofs of Theorems 1 and 2.

Bayesian solutions to two-stage sequential estimation problems have been considered by several authors—notably [2, 3], and [10].

The normality assumption has been used heavily, to suggest the estimators for μ and σ^2 and also for special properties of these estimators in (4), (9) and (10). It is expected that similar results may be obtained for multiparameter exponential families and other loss structures, and such extensions are currently under investigation in the doctoral work of Joon Lee. Extensions to a non-parametric context are more speculative.

It is a pleasure to acknowledge helpful conversations with Bob Keener, Joon Sang Lee, and Anand Vidyashankar and helpful comments from Anirban Dasgupta. This research was supported by the National Science Foundation and the U.S. Army Research Office.

References

[1] Abromowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions.* National Bureau of Standards.

[2] Cohen, A. and Sackrowitz, H. (1985). Bayes double sampling procedures, *Ann. Statist.* **12**, 1035–1049. MR751290

[3] Cohen, A. and Sackrowitz, H. (1994). Double sampling estimation when cost depends on the parameter, *Statistical Decision Theory and Related Topics* **V**, 253–266. MR1286306

[4] Hodges, J. L. Jr. and Lehmann, E. L. (1970). Deficiency, *Ann. Math. Statist.* **41**, 783–801. MR272092

[5] Malay, G., Mukhopadhyay, N. and Sen, P. K. (1997). *Sequential Estimation.* Wiley. MR1434065

[6] Johnson, R. (1970). Asymptotic expansions associated with posterior distributions, *Ann. Math. Statist.* **41**, 851–864. MR263198

[7] Rubin, H. and Sethuraman, J. (1965). Probabilities of moderate deviations, *Sankhya, Ser. A* **27**, 325–346. MR203783

[8] Rubin, H. and Sethuraman, J. (1965). Bayes risk efficiency, *Sankhya, Ser. A* 347–356. MR207112

[9] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance, *Ann. Math. Statist.* **16**, 243–258. MR13885

[10] Treder, R. P. and Sedransk, J. (1996). Bayesian double sampling plans with normal distributions, *J. Amer. Statist. Assn* **91**, 782–790. MR1395745

[11] Woodroofe, M. (1982). *Non-linear Renewal Theory in Sequential Analysis.* S.I.A.M. MR660065

[12] Woodroofe, M. (1985). Asymptotic local minimaxity in sequential point estimation, *Ann. Statist.* **13**, 676–688. Correction: **17**, p. 452. MR790565

[13] Woodroofe, M. (1986). Very weak expansions for sequential confidence intervals, *Ann. Math. Statist.* **14**, 1049–1067. MR856805

[14] Woodroofe, M. and Coad, D. S. (1997). Corrected confidence intervals for sequentially designed experiments, *Statistica Sinica* **7**, 53–74. MR1441144

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 237–249
© Institute of Mathematical Statistics, 2004

Estimating gradient trees

Ming-Yen Cheng[1], Peter Hall[2] and John A. Hartigan[3]

National Taiwan University, Australian National University and Yale University

Abstract: With applications to cluster analysis in mind, we suggest new approaches to constructing tree diagrams that describe associations among points in a scatterplot. Our most basic tree diagram results in two data points being associated with one another if and only if their respective curves of steepest ascent up the density or intensity surface lead toward the same mode. The representation, in the sample space, of the set of steepest ascent curves corresponding to the data, is called the gradient tree. It has a regular, octopus-like structure, and is consistently estimated by its analogue computed from a nonparametric estimator which gives consistent estimation of both the density surface and its derivatives. We also suggest 'forests', in which data are linked by line segments which represent good approximations to portions of the population gradient tree. A forest is closely related to a minimum spanning tree, or MST, defined as the graph of minimum total length connecting all sample points. However, forests use a larger bandwidth for constructing the density-surface estimate than is implicit in the MST, with the result that they are substantially more orderly and are more readily interpreted. The effective bandwidth for the MST is so small that even the corresponding density-surface estimate, let alone its derivatives, is inconsistent. As a result, relationships that are suggested by the MST can change considerably if relatively small quantities of data are added or removed. Our trees and forests do not suffer from this problem. They are related to the concept of gradient traces, introduced by Wegman, Carr and Luo (1993) and Wegman and Carr (1993) for purposes quite different from our own.

1. Introduction

Gradient trees capture topological features of multivariate probability densities, such as modes and ridges. In this paper we suggest methods for estimating gradient trees based on a sample of n observations from the density. Each estimator is in the form of a tree with $n-1$ linear links, connecting the observations. The methods will be evaluated in terms of their accuracy in estimating the population gradient tree, and their performance for real data sets. We also propose a new technique for describing, and presenting information about, neighbour relationships for spatial data.

To define a gradient tree, note that the gradient curves of a multivariate density f are the curves of steepest ascent up the surface $\mathcal{S}$ defined by $y = f(x)$. The representations of gradient curves, in the sample space, will be called *density ascent lines*, or DALs. The tree-like structure that they form is the gradient tree. This theoretical quantity may be estimated by replacing f by a nonparametric density estimator, $\hat{f}$ say, and then following the prescription for computing DALs and the gradient tree.

[1]Department of Mathematics, National Taiwan University, Taipei 106, Taiwan. e-mail: cheng@math.ntu.edu.tw

[2]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. e-mail: halpstat@maths.anu.edu.au

[3]Department of Statistics, Yale University, Box 2179, Yale Station, New Haven, CT 06520, USA. e-mail: john.hartigan@yale.edu

Keywords and phrases: density ascent line, density estimation, forest, gradient trace, minimum spanning tree, nearest neighbour methods, ridge estimation, tree diagram.

AMS 2000 subject classifications: primary 62H30; secondary 62H20.

A gradient tree may be viewed as a modification the concept of a 'gradient trace', introduced by Wegman, Carr and Luo (1993) and Wegman and Carr (1993). The goal of these authors was to use gradient traces to compute 'k-skeletons', which are k-dimensional analogues of the mode and represent nonlinear regression-like summary statistics. Our purpose is quite different. We view gradient trees as a tool for cluster analysis, and argue that in this context the concept has advantages over more familiar methodology such as minimum spanning trees, or MSTs, introduced by Florek *et al.* (1951); see also Friedman and Rafsky (1981, 1983).

An MST is the graph of minimum total length connecting all sample points. It is an estimator of the gradient tree that arises when we take $\hat{f}$ to be the most basic of nearest neighbour density estimators, in which the estimate at each point is inversely proportional to a monotone function of the distance to the closest sample point. However, this is a poor estimator of the population density, let alone its gradient, and so it is not surprising that the MST is a poor estimator of the corresponding population gradient tree. We suggest gradient tree estimators that are asymptotically consistent for the corresponding population gradient tree, and which also improve on the MST for small sample sizes.

We also suggest algorithms for drawing 'forests', using either the full dataset or subsets that have been identified by the gradient tree. Like the MST, a forest provides information about relationships among neighbouring data, but like our gradient tree it has the advantage that it is based on relatively accurate, and statistically consistent, information about gradients. In contrast with the MST, a forest is based on directed line segments, with the direction corresponding to movement up an estimate $\widehat{\mathcal{S}}$ of the surface $\mathcal{S}$. Our approach to constructing a forest allows the experimenter to choose, when describing relationships between points, how much emphasis will be given to a relatively conventional Euclidean measure of closeness of the points, and how much will be given to a measure of closeness related to movement up $\widehat{\mathcal{S}}$.

Although we work mainly in the bivariate case, our methods are certainly not limited to two dimensions. One way of treating high-dimensional data is of course to form bivariate scatterplots by projection, and apply our methods to the individual plots. Tools for manipulating two-dimensional projections of three- or higher-dimensional data include Asimov's (1985) grand tour, Tierney's (1990) Lisp-Stat, or Swayne, Cook and Buja's (1991) XGobi; see also Cook, Buja, Cabrera and Hurley's (1995) grand-tour projection-pursuit.

Moreover, density ascent lines and gradient trees have analogues when the sample space is of arbitrarily high dimension, rather than simply bivariate. (Analogues of forests may be constructed too, but the formula for a certain penalty term that is needed to define a forest is more complex in higher dimensions.) Hence, rather than compute these quantities for bivariate scatterplots, their multivariate forms (represented as lines in space, rather than lines in the plane) could be calculated and then viewed through their bivariate projections, or through rotations of trivariate projections.

Density-based approaches to assessing relationship have also been considered by Hartigan (1975), who took clusters to be maximal connected sets (that enjoyed at least a certain level of likelihood) of points of density exceeding a certain level. See also the discussion of tree diagrams by Hartigan (1982). Alternative approaches include methods based on measures of distance that satisfy the triangle inequality (e.g. Jardine and Sibson, 1971; Hubert, 1974) and techniques founded on parametric mixtures (e.g. Everitt, 1974; Kuiper and Fisher, 1975). Wishart (1969) was an early user of near neighbour methods to construct clusters.

Pruzansky, Tversky and Carroll (1982) compared spatial and tree representations of data.

2. Gradient trees and ridges

We begin by defining a 'true' density ascent line, when the density f of the bivariate distribution of a general data point X is assumed known. Then we discuss computation of this line, and calculation of its sample version.

Let $\mathcal{S}$ be the surface defined by the equation $y = f(x)$, and assume that both the first derivatives of f are continuous everywhere. Suppose too that the set of positive density is connected, and contains at most a finite number of stationary points. A *density ascent line* (DAL) for f, starting at a point x in the plane Π that denotes the sample space, is defined to be the projection, into Π, of the trajectory formed by climbing $\mathcal{S}$ in the direction of steepest ascent. Henceforth we shall call the 'projection' of a three-dimensional structure into Π, the 'representation' of that structure in Π, and reserve the term 'projection' for other purposes.

If the trajectory on $\mathcal{S}$ is represented as the locus of points $(x^{(1)}(s), x^{(2)}(s), y(s))$, where $s \in (0, s_0)$ is a convenient parameter such as distance along the trajectory from one of its ends, then the corresponding DAL will be the curve formed by the locus of points $(x^{(1)}(s), x^{(2)}(s))$, for $s \in (0, s_0)$, in Π. If f_1, f_2 denote the derivatives of f in the two coordinate directions then the curve of steepest ascent is in the direction (f_1, f_2), and is well defined except at stationary points of the density. The gradient tree is the collection of closures of DALs.

Next we give more detail about a DAL, and then an explicit method for computing one. Let $D(f) = (f_1^2 + f_2^2)^{1/2}$ denote the length of $\nabla f = (f_1, f_2)$, and put $\omega_j = f_j / D(f)$ and $\omega = (\omega_1, \omega_2)$. Then, for $x \in \mathcal{S}$, $\omega(x)$ is the unit vector in Π representing the direction of steepest ascent up $\mathcal{S}$, at the point $(x, f(x)) \in \mathcal{S}$. The DAL that passes through $x \in \Pi$ may be thought of as having been obtained, starting at a point on the line, by stepping along the line in the direction indicated by ω. Formally, the DAL that passes through $x \in \Pi$ may be represented by the infinitesimal transformation, $x \mapsto x + \omega(x)\, ds$, where ds is an element of displacement along the DAL, denoting the length of one of the aforementioned steps.

This suggests the following algorithm for computation. Given $x_0 \in \Pi$, and a small positive number δ, consider the sequence of points $\mathcal{P} \equiv \{x_j : -\infty < j < \infty\}$ defined by $x_j = x_{j-1} + \omega(x_{j-1})\, \delta$ and $x_{-j} = x_{1-j} + \omega(x_{1-j})\, \delta$, for $j \geq 1$. Thus, the DAL that passes through x_0 represents the limit, as $\delta \to 0$, of the sequence $\mathcal{P}$. The algorithm is convenient for numerical calculation, provided we stop before reaching places where $D(f)$ vanishes.

In empirical work, where we compute estimators of DALs, we of course replace f, f_1, f_2 in the algorithm by their estimators $\hat{f}, \hat{f}_1, \hat{f}_2$. We used the algorithm described above, with a suitably small value of δ, to calculate the empirical DALs shown in Section 4. Alternatively, one could recognise that DALs are integral lines of the gradient field of a smooth density function, implying that in principle they could be computed using an ordinary differential equation solver.

There is no commonly accepted definition of a *ridge* (or antiridge) of a surface such as $\mathcal{S}$, and in fact four different approaches, framed in terms of indices of 'ridgeness', were suggested by Hall, Qian and Titterington (1992). The following is related to the second definition there, and is chosen partly for ease of computation in the present context; its representation in Π is easily calculated from the functional $D(f)$. Moreover, the representation is itself a DAL, and it admits an elementary (and computable) generalisation to high-dimensional settings.

Given a point P on $\mathcal{S}$, let $\Pi' = \Pi'(P)$ denote the plane that contains P and is parallel to Π, and let $\mathcal{C}$ be the curve formed by the intersection of Π' with $\mathcal{S}$. If the steepest ascent curve up $\mathcal{S}$, starting from P, is perpendicular to $\mathcal{C}$ at P, then we say that P is a point on a ridge (or an antiridge) of $\mathcal{S}$. The ridge or antiridge itself is a locus of such points, and is the curve of steepest ascent on $\mathcal{S}$ that passes through P. (Therefore, its representation in Π is a DAL.) The point P is on a ridge, rather than an antiridge, if the curvature at P of the curve formed by the intersection of $\mathcal{S}$ with a plane perpendicular to Π, and containing P, is negative; and on an antiridge if the curvature is positive.

A ridge can bifurcate at a point which represents a location on $\mathcal{S}$ where three or more ridges join. The trajectories of steepest ascent that climb up the surface between two ridges meeting at a bifurcation point B, necessarily join one another at B. From there they have a common path, along an ascending ridge that leads away from B; and they continue together until they terminate at a local maximum, perhaps passing through other bifurcation points on the way.

The representation, in the plane Π, of a ridge and a bifurcation point will be called a ridge line (RL) and a branchpoint, respectively. The DALs corresponding to the representations (in Π) of ridges have different paths until they meet their first branchpoint, after which they are the same until they terminate at a mode. An RL is essentially what Wegman and Carr (1993) call 1-skeleton, the main difference being in the definition of a ridge.

Therefore, the DALs that comprise a gradient tree do have a tree-like structure, in the following way. Individual points in the sample space, representing leaves of the tree, are at first linked to branchpoints through distinct DAL paths. Beyond the first branchpoint the consolidated bundle of DAL paths, representing a branch of the tree, may be joined at subsequent branchpoints by other branches, until they finally reach a mode.

In theory, more complex structures are also possible, for example when two branches lead away from a branchpoint and come together again at a mode or at another branchpoint. However, it is rare in practice for such features to occur in DALs computed from data via nonparametric density estimators, and so we shall not consider them further here.

Two points $x_1, x_2 \in \Pi$ that are linked to the same mode by a DAL, may be said to lie in the same cluster. Thus, DALs divide the plane into clusters. Ridge lines divide the sample space in a DEFANGED different manner, in a sense orthogonal to the division into clusters. They give neither a subclassification nor a higher-level classification, but provide information of a different type, as follows.

If the ridge that produced an RL were almost horizontal, and lay between two local maxima of $\mathcal{S}$, occurring at points $x_{\max,1}$ and $x_{\max,2}$, say, in Π, then the points along that RL would have no clear allocation to the clusters corresponding to $x_{\max,1}$ and $x_{\max,2}$. Therefore, the RL would represent a watershed in the division of the sample space into clusters. On the other hand, a point that lay on either side of, and sufficiently close to, the RL would be more definitively allocated to just one of the clusters represented by $x_{\max,1}$ and $x_{\max,2}$.

More generally, we might fairly say that points that lie on one side or other of an RL are less ambiguously associated with their corresponding mode, at least if they are sufficiently close to the RL, than are points that lie directly on the RL. Indeed, if two points $x_1, x_2 \in \Pi$ lie on opposite sides of, and sufficiently close to, an RL, then all points x_3 that lie between x_1 and x_3 can be said to be more ambiguously associated with their corresponding modes than either x_1 or x_2.

In addition to their role in defining such a gradation of the sample space, the fact that RLs of density or intensity estimators represent the 'backbone' and 'ribs' of the structure of those quantities means that they provide valuable quantitative information about structure. Indeed, they are sometimes used to approximate the locations of physical structures associated with scatterplots, for example positions of the subterranean fault lines that give rise to earthquake epicentres (see Jones and Stewart, 1997).

Relative to ridge lines, antiridge lines have more connection with clustering in the usual sense, since they represent boundaries between regions where points are assigned to different clusters. However, they are typically computed from relatively little data, and so their locations may not be known as precisely as those of ridge lines.

Next we describe a method for locating, and computing, an RL, given the density f. A locus of points on $\mathcal{S}$, all of which have the same height above Π, is called a level set of $\mathcal{S}$. Its representation in Π is a contour of $\mathcal{S}$. An RL may be reached from another point in Π by moving around a contour. The orientation of the contour passing through x is the direction of the unit vector $\omega_{\rm perp}(x)$, say, defined as being orthogonal to $\omega(x)$ and determined up to a change of sign. Therefore, the contour is defined by the infinitesimal transformation $x \mapsto x \pm \omega_{\rm perp}(x)\, ds$, where ds is an infinitesimal unit of length around the contour. The point at which this contour cuts an RL is a local minimum of $D(f)$; a local maximum corresponds to cutting the representation in Π of an antiridge.

Hence, to find a point x on an RL we move around the contour, computing $D(f)$ as we go, until we find a local minimum of $D(f)$. Then, moving along the RL is equivalent to moving up the DAL starting from x, or down the DAL leading to x; we have already described how this may be done. It is helpful to note that turning points of $D(f)$ are solutions of the equation

$$f_{12}\left(f_1^2 - f_2^2\right) = f_1 f_2 \left(f_{11} - f_{22}\right),$$

where $f_{ij}(x) = \partial^2 f(x_1, x_2)/\partial x_i\, \partial x_j$. Of course, descending the DAL that defines a ridge is equivalent to traversing the line defined by $x \mapsto x - \omega(x)\, ds$, where now ds is an infinitesimal unit of length along the DAL.

More generally, if the sample space Π is p-dimensional, where $p \geq 2$; and if we define $D = (\sum_i f_i^2)^{1/2}$, where f_i equals the derivative of f in the direction of the ith coordinate direction, for $1 \leq i \leq p$; then a ridge line or antiridge line is a locus in Π of turning points of $D(f)$. It may be calculated by generalising the method suggested above. DEFANGED A practicable, computational algorithm for an RL may be obtained as before, replacing the infinitesimal ds by a small positive number δ. The empirical version, in which density f is replaced by the density estimator $\hat{f}$, also follows as before; we used this method to compute the RLs shown in Section 4. Tests for significance of empirical modes may be based on work of Silverman (1981), Hartigan and Hartigan (1985), Müller and Sawitzki (1991) or Cheng and Hall (1999), for example.

3. Forests based on distance and density

While the minimum spanning tree is not consistent for the population gradient tree, it provides some information about relationships among neighbouring data values. In this section we suggest a regularisation of the minimum spanning tree in which links between observations are penalised if they are not sufficiently close to estimated density ascent lines. It may be applied to a subset $\mathcal{Y} = \{Y_1, \ldots, Y_N\}$

of the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, for example to those data that are linked to the same mode in the gradient tree, as well as to the full sample.

Let $\|Y_i - Y_j\|$ denote Euclidean distance in the sample space Π, and let $d(Y_i, Y_j)$ be some other measure of distance between Y_i and Y_j. It is not necessary that $d(\cdot, \cdot)$ be a metric; appropriate definitions of d are powers of Euclidean distance in Π, i.e. $d(Y_i, Y_j) \equiv \|Y_i - Y_j\|^s$, and powers of Euclidean distance on $\widehat{\mathcal{S}}$, i.e.

$$d(Y_i, Y_j) \equiv \left[\|Y_i - Y_j\|^2 + \{\hat{f}(Y_i) - \hat{f}(Y_j)\}^2\right]^{s/2},$$

where $s > 0$. In our numerical work in Section 4 we shall use the first of these definitions, with $s = 2$.

Now add a penalty to $d(Y_i, Y_j)$, proportional to the squared length of the projection of $Y_i - Y_j$ orthogonal to $\widehat{\omega}(Y_i)$. (Here, $\widehat{\omega}(x)$ denotes the empirical form of $\omega(x)$, computed with $\hat{f}$ replacing f.) Equivalently, the penalty is proportional to the area of the triangle that has one side equal to the length of the line joining Y_i and Y_j, and another equal to the length of the representation in Π of a straight-line approximation, of the same length as the previous side, to the gradient curve. The area in question equals half the value of $\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2$, if the vertex of the triangle is at Y_i. We apply these penalties in proportion to a tuning parameter $t \geq 0$, obtaining symmetrically and asymmetrically penalised versions, respectively, of $d(Y_i, Y_j)$:

$$D(Y_i, Y_j) = d(Y_i, Y_j) + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2\right] \quad \text{or} \quad (3.1)$$

$$D(Y_i, Y_j) = d(Y_i, Y_j) + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2\right] + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_j)\}^2\right]. \quad (3.2)$$

Using a large value of t amounts to placing more emphasis on point pairs whose interconnecting line segment lies close to a gradient curve.

We are now in a position to construct the forest corresponding to the dataset $\mathcal{Y}$ and the penalised distance measure D. Given Y_i, we draw a directed line segment from Y_i to Y_j if and only if Y_j minimises $D(Y_i, Y_j)$ over all points Y_j for which $\hat{f}(Y_j) > \hat{f}(Y_i)$. The forest is the set of these directed segments. If $\mathcal{Y}$ is a cluster, and if we adjoin to $\mathcal{Y}$ the unique mode associated with that structure, then with probability 1 there is exactly one point Y_i (the mode) in $\mathcal{Y}$ for which the directed line segment does not exist. As we climb higher up the surface the directed line segments tend to coalesce, producing a tree structure sprouting from the mode (although it was constructed from the opposite direction).

If we define $D(\cdot, \cdot)$ as at (3.1) then taking $t = 0$ produces a forest that is similar in both definition and appearance to the minimum spanning tree, although based on directed line segments. Choosing a relatively large value of t imposes greater penalty for not walking as nearly as possible along the DAL that starts at Y_i, when passing from Y_i to Y_j. The extent to which line segments cross over in the forest may be reduced by increasing t, thereby forcing the direction of movement on $\widehat{\mathcal{S}}$ to give more emphasis to the uphill component of motion. The advantage of (3.2) over (3.1) is that in the former the tree treats the notions of 'uphill' and 'downhill' symmetrically, but in practice, forests defined by (3.1) and (3.2) are virtually identical.

4. Numerical examples

Rees (1993) determined the 'proper motions' of 515 stars in the region of the globular cluster M5. Using the proper motions and radial velocity dispersions he estimated the probability that each star belonged to the cluster. The analysis below is

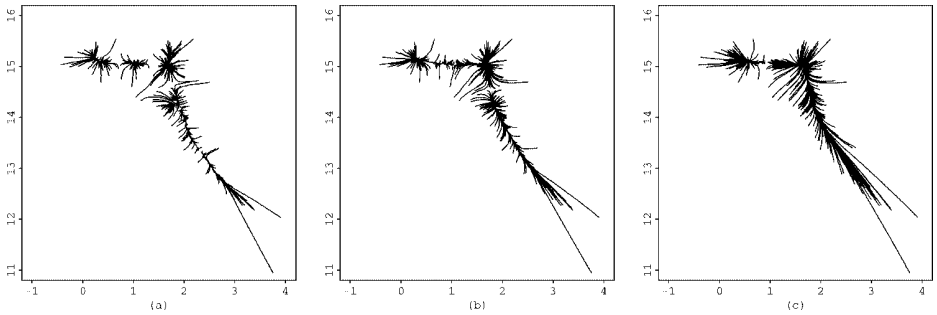


Figure 1: *Steepest Ascent Trees.* Panels (a), (b) and (c) depict DALs for the smoothed nearest neighbour estimator corresponding to $k = 25, 50, 100$, respectively.

based on the Herzprung-Russell diagram, a plot of magnitude versus temperature, for the 463 stars that were determined by Rees to have probability of at least 0.99 of belonging to the cluster.

We employed two different versions of $\hat{f}$. Both were nearest neighbour methods, which we chose for reasons that were both pragmatic (the adaptivity of NN methods means that they have less tendency than other density estimation techniques to suffer from spurious islands of mass) and didactic (NN methods are commonly used in classification problems). The first version of $\hat{f}$ was a standard k'th nearest neighbour estimator, with $\hat{f}(x)$ equal to $k/(n\pi r^2)$ where $r = r(x)$ was the smallest number such that the circle centred on x and with radius r contained just k points. The second density estimator was a smoothed version of the first, equal to $2k/(n\pi r^2)$ where r was the solution of

$$\sum_{i=1}^{n} \left\{ 1 - \left(\frac{\|X_i - x\|}{r} \right)^2 \right\}^+ = k.$$

See Section 5 for discussion of this technique. Since our graphs remain unchanged if we multiply $\hat{f}$ by a constant factor then it is not necessary to normalise, and so the factor $k/n\pi$ may be dropped.

Figure 1 depicts the gradient tree, or collection of DALs, for $k = 25, 50, 100$. In constructing figures 1 and 2 we used only the second, smoothed nearest neighbour estimator $\hat{f}$. Note that as k increases the number of empirical modes decreases; the number is $7, 4, 2$ for $k = 25, 50, 100$ respectively. The gradient trees indicate which points are most closely associated with the respective modes. The orientations and spacings of the tentacles of these 'octopus diagrams' provide information about the steepness of $\hat{f}$ in different places.

Figure 2 shows the RLs for the same values of k. Ridge lines are depicted by solid lines, and antiridge lines by dashed lines. The main RL, in the lower right of the figure, is clearly depicted; it is in a sense the backbone of the surface defined by the density estimator. Other RLs represent relatively minor 'creases' in the surface, and play more the role of 'ribs'.

The gradient trees provide only minimal information about interpoint relationships. Detail of that type is more readily supplied by forests, depicted in figures 3 and 4 for the two respective density estimators. We used the distance function defined at (3.1), with $d(Y_i, Y_j) = \|Y_i - Y_j\|^2$. The six panels in each figure represent different pairs of values of the smoothing parameter $k = 25, 50, 100$ and gradient

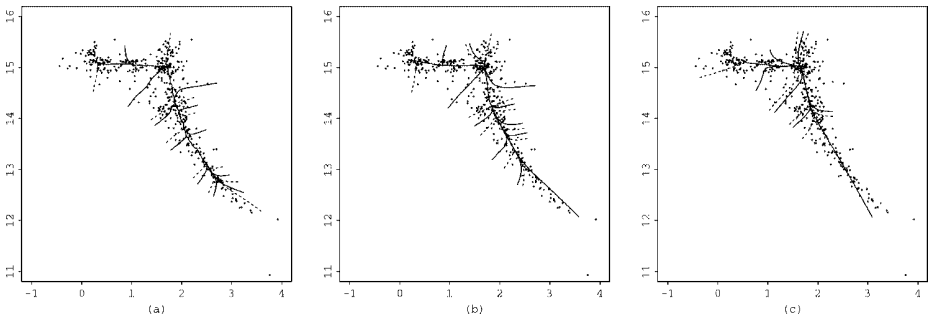


Figure 2: *Ridge Projections.* Panels (a), (b) and (c) show the ridge lines (solid) and antiridge lines (dashed) corresponding to the respective DALs in figure 1. To illustrate relationships to the data, a scatterplot of the data is included in each panel.

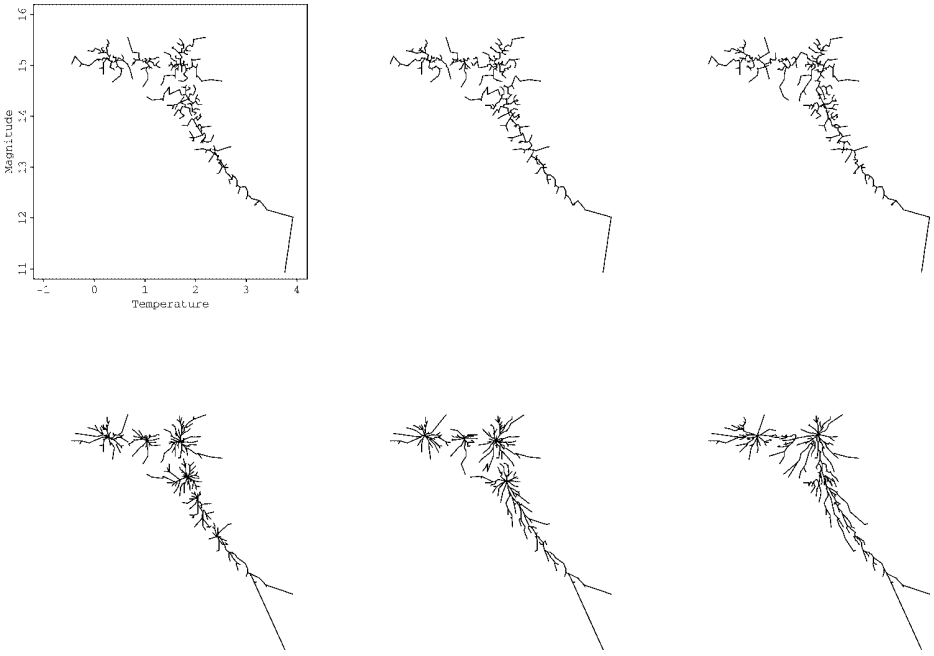


Figure 3: *Forests.* Forests drawn using the unsmoothed nearest neighbour estimator, with $t = 0$ (top row) and $t = 10$ (bottom row), and $k = 25$, 50, 100 (columns 1–3).

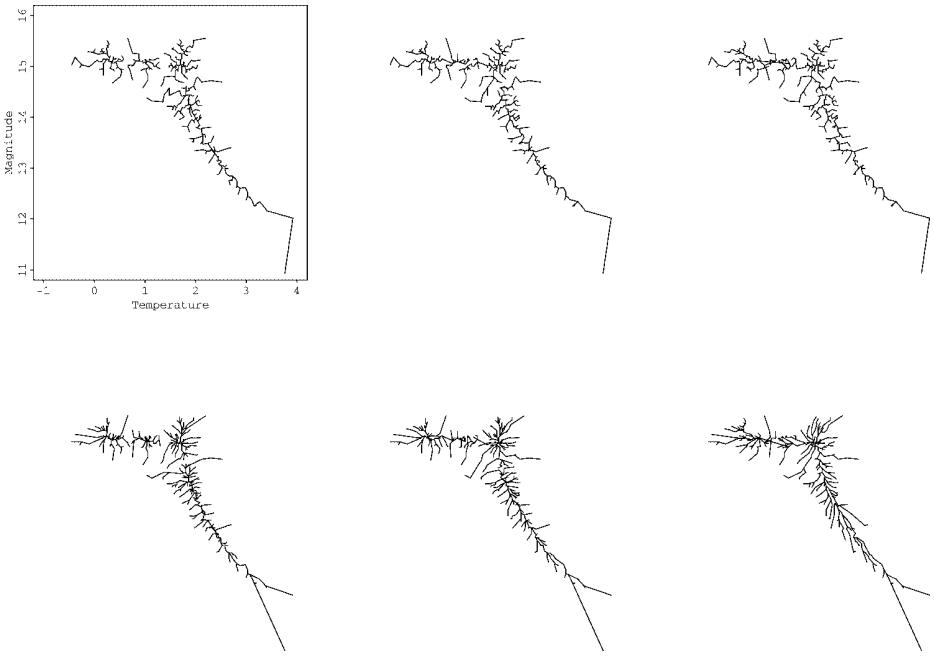


Figure 4: *Forests.* Forests drawn using the smoothed nearest neighbour estimator, with panels ordered as in figure 3.

weight $t = 0, 10$. Taking $t = 0$ produces directed line segments based almost entirely on distances between points, except that the direction of the segment is always that of increasing estimated density. The resulting forest is comparable to the minimum spanning tree, and its links have almost random orientation. On the other hand, using $t = 10$ gives heavy weight to segments that lie close to the representation in Π of the estimated gradient curve, and (for both density estimators) produces a more orderly presentation of the links.

Overall, the data show strong evidence of a northwest to southeast ridge, and at least three modes. Smoothing the density estimator produces some regularisation of forests, but choice of k has much greater effect on our graphs than estimator type.

In order to further illustrate performance of the gradient tree approach, these methods, along with two conventional graphical tools (contour plots and perspective mesh plots), were applied to two simulated data sets. In these examples, which are discussed below, smoothed nearest neighbour estimators were employed whenever estimation of the density and its gradients were required.

In the first example, 500 random variates were generated from the bimodal Normal mixture,

$$0.7\,N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) + 0.3\,N\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}0.26 & -0.13\\-0.13 & 0.65\end{pmatrix}\right). \qquad (4.1)$$

The smoothing parameter was $k = 45$, and gradient weight was $t = 10$. The data, contour plots, and perspective mesh plots based on the density estimator, are shown in panels (a) and (b) of figure 5, which provide evidence of bimodality. However, the density ascent lines, ridge lines and forests, depicted in panels (c), (d) and (e)

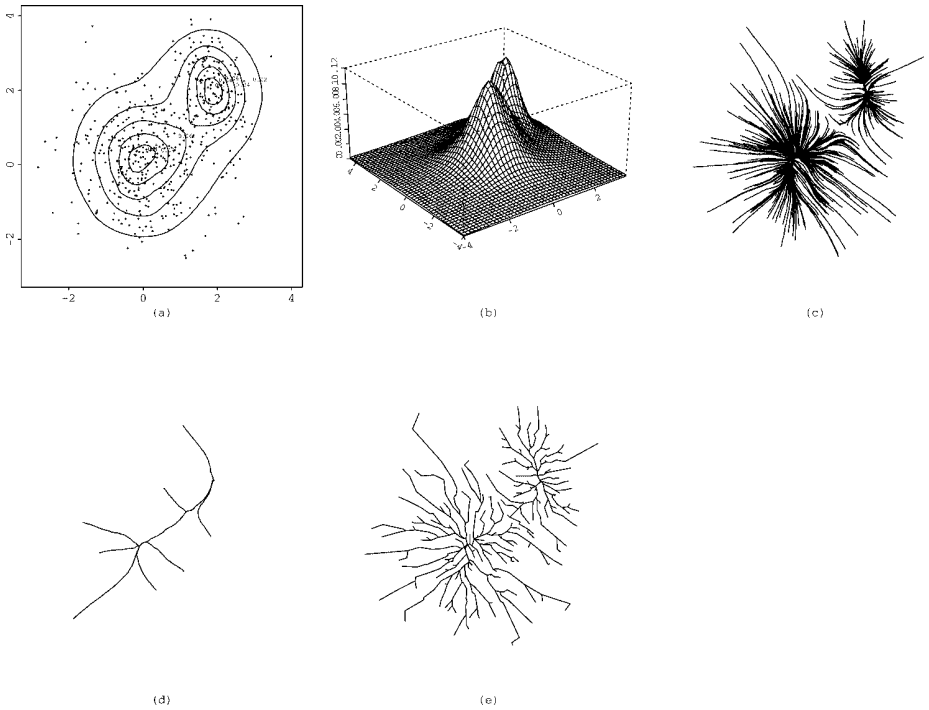


Figure 5: *Bimodal data example.* A scatterplot of 500 random numbers simulated from model (4.1) is shown in panel (a). Panels (a), (b), (c), (d) and (e) depict respectively contour plots, a perspective mesh plot, density ascent lines, ridge lines, and forests based on the smoothed nearest neighbour estimator with $k = 45$ and $t = 10$.

of figure 5, show more clearly than panels (a) and (b) structure of the surface, and in particular the locations of the two modes and the steepest ascent directions up the surface.

Each of the graphical methods illustrated in panels (c) and (d) divides the 500 data points into two subgroups, in which each point is connected to the centre of the subgroup to which it belongs. The directions of the density ascent curves, and hence information about the way in which the surface increases as one moves in different directions, are conveyed much better by these two graphics than by those in panels (a) and (b). Most importantly, panels (c) and (d) allow the reader to extract point-to-point relationships from the data to a significant extent; such information cannot be so readily obtained from the contour plot (panel (a)) or the perspective mesh plot (panel (b)).

The second example is of data simulated from a model, described below, which has more complex structure than that described at (4.1). Let U, V, W, Z be independent random variables, with U and V having the $N(0, 0.06^2)$ distribution, W being uniformly distributed on the interval $(-1, 1)$, and Z having density $g(z) = 0.2z+0.5$ for $|z| \le 1$. Put

$$X = \mathrm{sgn}(W)\,(0.6 - Z)\,I(-1 \le Z \le 0.6) + U, \qquad Y = Z + V, \tag{4.2}$$

where $I(\cdot)$ denotes the indicator function. The surface defined by the joint density of (X, Y) has two ridges, represented by the lines $x = \pm(0.6 - y)$ for $-1 \le y \le 0.6$,

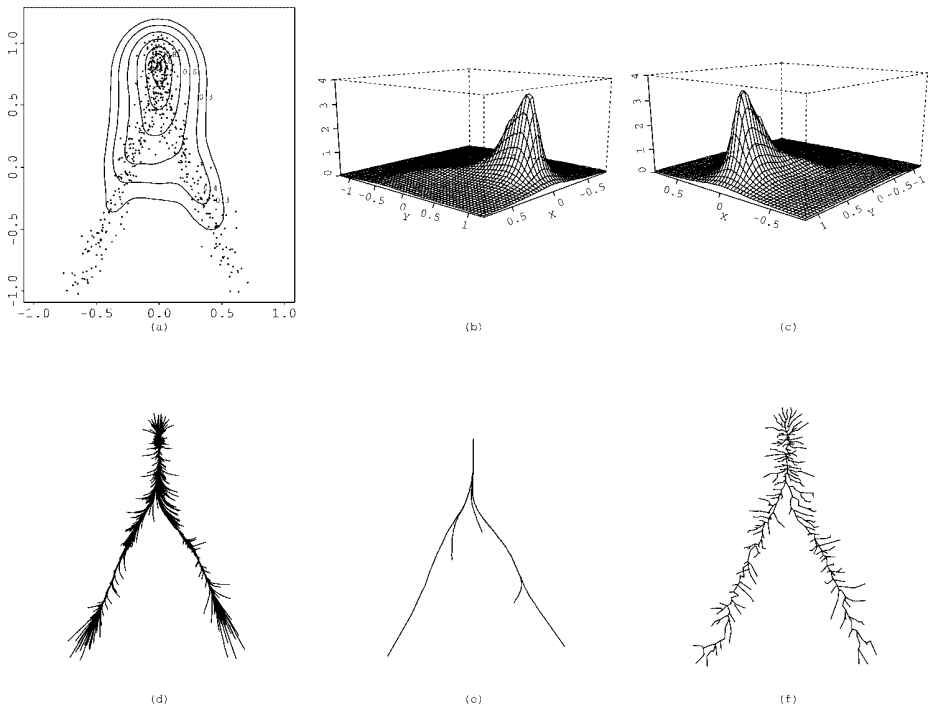


Figure 6: *Ridge data example.* A scatterplot of 500 observations simulated from model (4.2), and the corresponding contour plots, are shown in panel (a). Perspective mesh plots from different angles, showing the two ridge branches, are given in panels (b) and (c). Panels (d), (e) and (f) depict respectively density ascent lines, ridge lines, and forests. Graphics here used the smoothed nearest neighbour estimator with $k = 55$ and $t = 10$.

which merge at $(0, 0.6)$ and then continue together along the line $x = 0$ until the point $(0, 1)$ is reached. The height of the surface increases steadily as one travels along any of these ridges in a direction that has a northbound component.

We generated 500 observations from model at (4.2). The smoothing parameter was taken to be $k = 55$, and the gradient weight was $t = 10$. Panel (a) of figure 6 incorporates a scatterplot of the dataset. The contour plots and perspective mesh plots, given in panels (a)–(c) of figure 6, provide only a vague impression of the biridge nature of the data. In contrast, the density ascent lines, ridge and antiridge lines and forests, shown in panels (d)–(f) of figure 6, provide substantially less ambiguous information about the ridges and, more generally, about the nature of the scatterplot.

The tree and forest structures in different datasets, for example those in our last two examples, are readily compared. In particular the very different characters of the 'octopus plots' (tree structures made up of density ascent lines) in panel (c) of figure 5, and panel (d) of figure 6, are immediately apparent. The first shows two approximately symmetric clusters about single centres, with little evidence of ridges, while the second demonstrates marked asymmetry and 'ridginess'. Likewise, the forests in panel (d) of figure 5, and panel (f) of figure 6, show very different hierarchical structures. The first demonstrates a relatively low level of relationship among different points in the cluster, with many of the branches of the forest joining

the cluster relatively close to the respective mode, and so being related to other branches (and hence other points in the cluster) largely through that mode. On the other hand, panel (f) of figure 6 shows a strong degree of hierarchy, with each branch of the forest joining its respective 'ridge branch' after travelling only a short distance, and being linked to other branches though the ridge.

5. Density estimators and theory

The two-dimensional nearest-neighbour density estimators used in Section 4 may be described as follows. Given a kernel K, put $\hat{f}(x) = \hat{f}(x|R) = R/nh_x^2$ where h_x is given by

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_x}\right) = R.$$

If K is the uniform kernel, equal to $1/\pi$ within a region $\mathcal{R}$ and 0 elsewhere, then this prescription requires h_x to be such that R data values are contained within the region $x \oplus h_x\mathcal{R}$, which of course is the standard near-neighbour construction. A disadvantage of the uniform kernel, however, is that the resulting estimator is very rough. The second approach discussed in Section 4 uses a bivariate form of the Epanechnikov kernel. Alternatively we could use bivariate biweight or triweight kernels.

We employed the same value of R for all x, so that the bandwidth h_x was relatively small in regions of high data density. Assuming that $R = R(n) \to \infty$ and $R/n \to 0$ as $n \to \infty$ it may be shown that $h_x \sim \{R/n\kappa_1 f(x)\}^{1/2}$ as $n \to \infty$, where $\kappa_j = \int K(v)^j\, dv$. In particular, the effective bandwidth is of size $(R/n)^{1/2}$. Assuming that K is symmetric and f has two bounded derivatives, the bias and variance of $\hat{f}$ are of sizes R/n and $(n/R^3)^{1/2}$, respectively. Therefore, optimal mean-square performance of the estimator $\hat{f}$ is obtained with R of size $n^{5/7}$, in which case mean squared error equals $O(n^{-4/7})$, just as it would be for a traditional second-order kernel estimator. Variance is asymptotic to $(nf^5\kappa_1^3/R^3)^{1/2}\kappa_2$.

Note particularly that, using bandwidths of these sizes, our gradient estimators are consistent for the true gradients. That is not true for the implicit gradient estimators employed in a minimum spanning tree, which are in effect based on a bandwidth that is of size $n^{-1/2}$. This means that the error-about-the-mean term in the estimator of f, let alone for estimators of the derivatives of f, does not converge to zero, which accounts for the haphazard, complex structure of minimum spanning tree diagrams.

References

[1] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Statist. Comput.* **6**, 128–143. MR773286

[2] Cheng, M.-Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *J. Roy. Statist. Soc. Ser. B* **60**, 579–590 MR1625938

[3] Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand tour and projection pursuit. *J. Computat. Graph. Statist.* **4**, 155–172.

[4] Everitt, B. S. (1974). *Cluster Analysis*. Halstead, London. MR455213

[5] Florek, K., Lukaszwicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble finit. *Colloq. Math.* **2**, 282–285. MR48832

[6] Friedman, J. H. and Rafsky, L. C. (1981). Graphics for the multivariate two-sample problem. (With discussion.) *J. Amer. Statist. Assoc.* **76**, 277–295. MR624331

[7] Friedman, J. H. and Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377–391. MR696054

[8] Hall, P., Qian, W. and Titterington, D. M. (1992). Ridge finding from noisy data. *J. Computat. Graph. Statist.* **1**, 197–211. MR1270818

[9] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York. MR405726

[10] Hartigan, J. A. (1982). Classification. In: *Encyclopedia of Statistical Sciences* **2**, Eds. S. Kotz and N. L. Johnson, pp. 1–10. Wiley, New York. MR670950

[11] Hartigan, J. A. and Hartigan, P. M. (1985). The DIP test of unimodality. *Ann. Statist.* **13**, 70–84. MR773153

[12] Hubert, L. J. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *J. Amer. Statist. Assoc.* **69**, 698–704. MR373170

[13] Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. Wiley, New York. MR441395

[14] Jones, R. H. and Stewart, R. C. (1997). A method for determining significant structures in a cloud of earthquakes. *J. Geophysical Res.* **102**, 8245–8254.

[15] Kuiper, F. K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**, 777–784.

[16] Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746. MR1147099

[17] Pruzansky, S., Tversky, A. and Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika* **47**, 3–24.

[18] Rees Jr., R. F. (1993). New proper motions in the globular cluster M5. *Astron. J.* **106**, 1524–1532.

[19] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97–99. MR610384

[20] Swayne, D. F., Cook, D. and Buja, A. (1991). XGobi: Interactive dynamic graphics in the X window system with a link to S. In *ASA Proceedings of the Section on Statistical Graphics*, pp. 1–8.

[21] Tierney, L. (1990). *LISP-STAT, An Object-Oriented Environment for Statistics and Dynamic Graphics*. Wiley, New York.

[22] Wegman, E. J. and CARR, D. B. (1993). Statistical graphics and visualization. In: *Handbook of Statistics* **9***: Computational Statistics*, Ed. C. R. Rao, pp. 857–958. North Holland, Amsterdam.

[23] Wegman, E. J., Carr, D. B. and LUO, Q. (1993). Visualizing multivariate data. In: *Multivariate Analysis: Future Directions*, Ed. C. R. Rao, pp. 423–466. North Holland, Amsterdam. MR1246351

[24] Wishart, D. (1969). A generalization of nearest neighbour which reduces chaining effects. In: *Numerical Taxonomy*, Ed. A. J. Cole, pp. 282–311. Academic, London.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 250–254
© Institute of Mathematical Statistics, 2004

Conservative bounds on extreme P-values for testing the equality of two probabilities based on very large sample sizes

Herman Chernoff[1]

Harvard University

Abstract: With very large sample sizes the conventional calculations for tests of the equality of two probabilities can lead to very small P-values. In those cases, the large deviation effects make inappropriate the asymptotic normality approximations on which those calculations are based. While reasonable interpretations of the data would tend to reject the hypothesis in those cases, it is desireable to have conservative estimates which don't underestimate the P-value. The calculation of such estimates is presented here.

1. Introduction

There are several excellent alternatives for testing the hypothesis that $p_1 = p_2$ where p_1 and p_2 are probabilities governing two binomial samples. These include the Yates continuity correction and the Fisher Exact test and several others based on the asymptotic normality of the observed proportions. All these test procedures have the desireable property that the calculated P-value does not depend on the unknown common probability under the hypothesis. There is a slight problem with the Fisher exact test, i.e., it is not strictly appropriate for the problem because the calculated probability is conditional on the values of the margins, which are not fixed in advance. The problem is considered slight because the information in the margins is quite small Chernoff (2004).

In a legal case the problem arose where there were 7 successes out of 16 trials for one sample and 24 successes out of 246 in the second sample. It is clear that the hypothesis is not plausible in the light of these data. Since the various alternative tests provide substantially different calculated P-values, all very small, it was considered wise to present a very conservative P-value. While one sample size was substantial, the other was quite modest. Neither was so large that modern computers would be frustrated by calculating the exact P-value rather than relying on asymptotic theory. One consequence of such an approach is that the P-value is no longer independent of the unknown value of the nuisance parameter, the common value of the probabiities under the hypothesis. This problem is dealt with in several publications (Berger and Boos (1994), Chernoff (2003)). A crucial aspect of the difficulty in using asymptotic theory is that in extreme cases where the P-values are very very small, we are in the tails of the distribution and asymptotic normality no longer fits in these large deviation cases.

A new problem recently came to my attention, where both sample sizes are enormous, i.e. $n_1 = 19,479$ and $n_2 = 285,422$, Here agains there are several cases

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA. e-mail: chernoff@stat.harvard.edu

Keywords and phrases: P-values, Chernoff bound, Yates continuity correction, Poisson.
AMS 2000 subject classifications: primary 62F03; secondary 62B10.

where we have a large deviation problem, and asymptotic normality is not appropriate, and probably not conservative. How should we deal with this problem in this example where ordinary high speed computers may find it difficult to provide exact calculations such as were feasible in the previous case? The Chernoff bound, originally derived by H. Rubin, provides a method of deriving an upper bound on the desired probability which is convenient to calculate.

2. The Poisson approximation

While the normal approximation is unreliable, the Poisson approximation may be better. In any case, it is to be used here merely to provide an initial approximation for the quantities required for the binomial calculation. We outline the analysis which provides a solution assuming the Poisson approximation fits.

The main tool to deliver a conservative bound on the P-value is the Chernoff bound, first derived by Herman Rubin, using a Chebyshev type of inequality, that states that if $d \geq E(X)$,

$$P(X \geq d) \leq E\big(e^{t(X-d)}\big)$$

for all t. The right hand side attains its minimum for $t \geq 0$.

Let X_1 and X_2 be the number of successes in n_1 and n_2 independent trials with common probability p, and let

$$D = \frac{X_1}{n_1} - \frac{X_2}{n_2},$$

Using the Poisson approximation to the binomial distribution, we shall derive the curve in the (p,d) space for which the bound on $\log(P(D \geq d))$,

$$q = \log\big(\inf_t E\big(e^{t(D-d)}\big)\big)$$

attains a given value, for $d > 0$. Under the assumption that the number of successes in each trial has a Poisson distribution, we have

$$Q(t,d) = \log\big(E\big(e^{t(D-d)}\big)\big) = -dt + n_1 p\big(e^{t/n_1} - 1\big) + n_2 p\big(e^{-t/n_2} - 1\big).$$

Differentiating with respect to t, the value of t which minimizes Q satisfies

$$e^{t/n_1} - e^{-t/n_2} = d/p = a$$

while

$$Q(t,d) = pr(t,a)$$

where

$$r(t,a) = -at + n_1\big(e^{t/n_1} - 1\big) + n_2\big(e^{t/n_2} - 1\big).$$

For each value of t, there is a corresponding value of a for which t is optimal and a corresponding value of r. Let $p = q/r$ and $d = ap$. As t varies these values of p and d trace out the (p,d) curve corresponding to the given value of $q \geq \log(P)$.

3. The binomial case

We use the Poisson calculation to get a first approximation in the derivation of the $(p.d)$ curves for the binomial case. In the previous section we obtained values of p and d for each value of t. Here we will keep both p and q fixed, and starting with the value of t, we find

$$Q(t,d) = \log E\big(e^{t(D-d)}\big) = -td + n_1 \log\big(1 - p + pe^{t/n_1}\big) + n_2 \log\big(1 - p + pe^{-t/n_2}\big)$$

and the value of d for which Q is minimized by the given value of t is given by

$$d(t) = (1-p)\left(\frac{1}{1-p+pe^{-t/n_2}} - \frac{1}{1-p+pe^{t/n_1}}\right).$$

We note that

$$d'(t) = p(1-p)\left(\frac{e^{t/n_1}}{n_1(1-p+pe^{t/n_1})^2} + \frac{e^{-t/n_2}}{n_2(1-p+pe^{-t/n2})^2}\right).$$

Insofar as $Q(t, d(t))$ varies from the specified value of q, we apply the Newton iteration to modify t. This leads t to the new value $t + (q - Q(t, d(t))/Q'(t)$ where

$$Q'(t) = \partial Q/\partial t + d'(t)\partial Q/\partial d = -td'(t).$$

Thus t goes into $t - (q - Q)/td'(t)$.

If the new value of t and $d(t)$ do not provide $Q(t, d(t))$ close enough to the desired value q, one may iterate again. Finally we have for each initial value of t and the given value of q a new point (p, d) for the curve of specified $q \geq \log(P(D \geq d))$.

While the curves we have obtained of (p, d) values for a given value of q are useful, they don't resolve the inverse problem in which we may be interested. That is, how do we calculate a bound on the P-value for a given p and d? A series of curves provided above would be useful to get rough approximations for a set of cases with given n_1 and n_2, but do not provide a reasonable precise algorithm should that be desired. To obtain the bound on the P-value, we start with the estimate of p given by $p = (X_1 + X_2)/(n_1 + n_2)$. Assusming that value is fixed, we approximate t, assuming t is small compared to n_1 and n_2, by

$$t = \frac{dn_1n_2(1-p)}{(n_1+n_2)p}$$

This value of t together with the observed value of D yields $Q(t, D)$ and $d(t)$. Insofar as $d(t)$ differs from D, we modify t by the Newton method to $t + (D - d(t))/d'(t)$. With this new value of t, we recalculate Q and $d(t)$ and interate until $d(t)$ is approximately D. Then the bound on the P-value is given by e^Q assuming our estimate of p is accurate. Since the range of possible values of p is quite limited under the hypothesis, we can see how much the P-value changes by considering potential alternative values of p.

4. Summary

For the case of very large sample sizes, with data quite inconsistent with the hypothesis that two binomial distributions have the same value of p, we anticipate very small P-values. The usual calculations are unreliable because large deviation effects make the asymptotic normality on which these calculations depend unreliable. While it is clear in such cases that the hypothesis is false, it is often desireable to have a conservative bound on the P-value. The Chernoff bound provides such a result. We provide the basis for three algorithms. One provides the (p, d) values for which given bounds on the value of $\log(P)$ are attained assuming that a Poisson approximation to the binomial distribution is acceptable. This algorithm is used as a starting point in calculating the curve of (p, d) values for the binomial distribution. Finally we show how to calculate the conservative bound for the P-value in the binomial case.

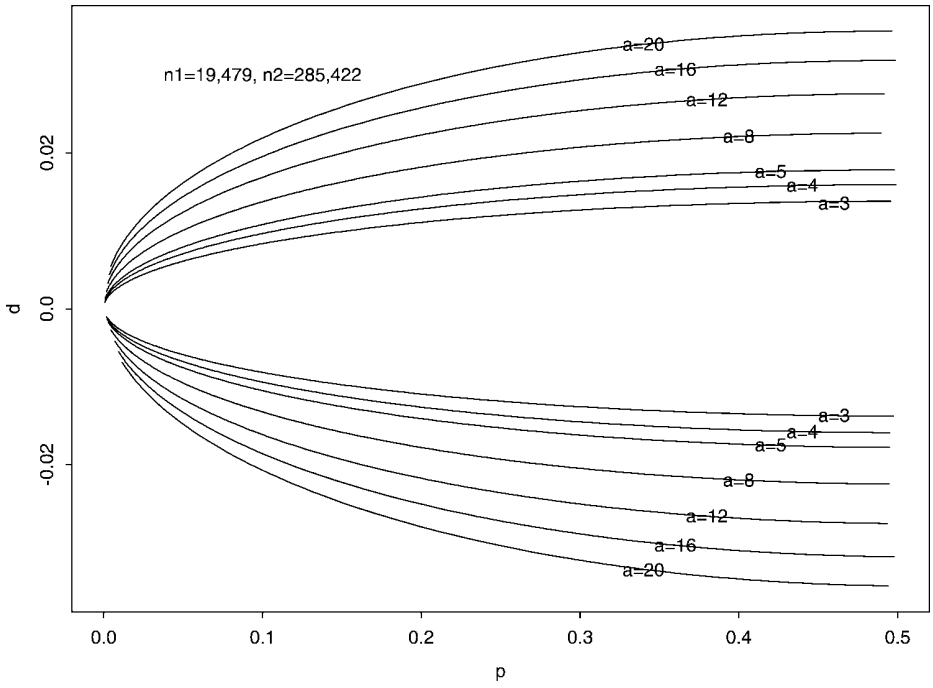


Figure 1:

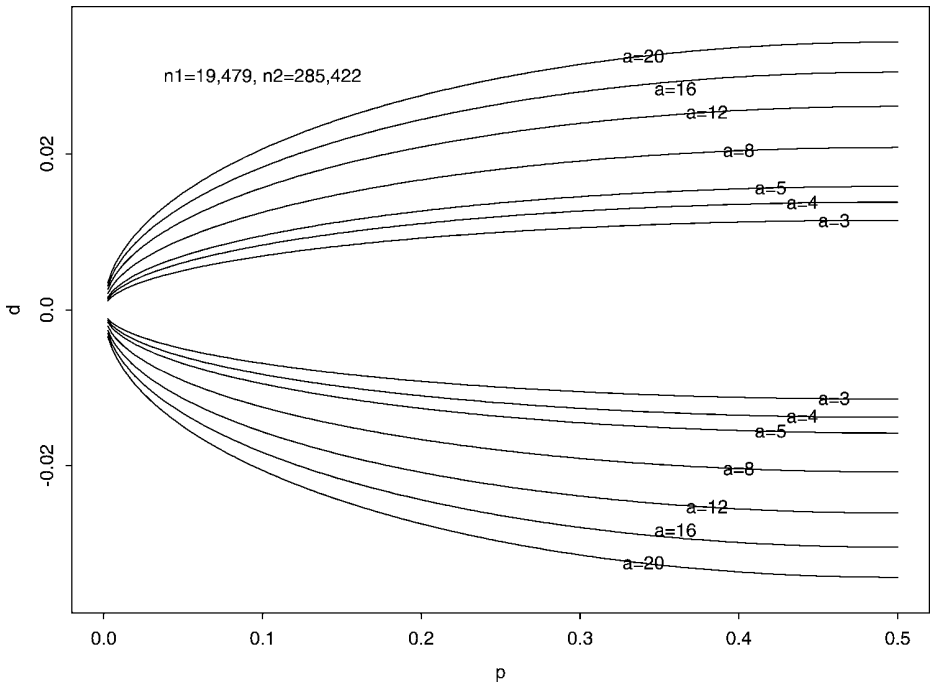


Figure 2:

We have in Figure 1, the (p, d) values for the case $P = 10^{-a}$ where a takes on the values 3,4,5,8,12,16, and 20, $n_1 = 19,479$, $n_2 = 285,422$, and we use the binomial distribution. In Figure 2 we use the calculation for the Yates continuity correction where p represents the estimate of the common probability.

In both of these cases we have calculated one sided P-values. The calculation for negative values of D can be obtained by interchanging n_1 and n_2 after replacing D by its absolute value.

References

[1] Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, **89**, 1012–1016. MR1294746

[2] Chernoff, H. (2003). Another View of the Classical Problem of Comparing Two Probabilities, *J. Iranian Statist. Soc.*, **1**, 35–54. MR1981752

[3] Chernoff, H. (2004). Information for testing the equality of two probabilities, from the margins of the 2×2 table. *J. Statist. Plann. Inference*, **121**, 209–214. MR2038817

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 255–286
© Institute of Mathematical Statistics, 2004

Detecting a target in very noisy data from multiple looks

Jiashun Jin[1]

Purdue University

Abstract: Consider an imaging situation with extremely high noise levels, hidden in the noise there may or may not be a signal; the signal—when present—is so faint that it cannot be reliably detected from a single frame of imagery. Suppose now multiple frames of imagery are available. Within each frame, there is only one pixel possibly containing a signal while all other pixels contain purely Gaussian noise; in addition, the position of the signal moves around randomly from frame to frame. Our goal is to study how to reliably detect the existence of the signal by combining all different frames together, or by "multiple looks".

In other words, we are considering the following testing problem: test whether all normal means are zeros versus the alternative that one normal mean per frame is non-zero. We identified an interesting range of cases in which either the number of frames or the contrast size of the signal is not large enough, so that the alternative hypothesis exhibits little noticeable effect on the bulk of the tests or for the few most highly significant tests. With careful calibration, we carried out detailed study of the log-likelihood ratio for a precisely-specified alternative. We found that there is a threshold effect for the above detection problem: for a given amplitude of the signal, there is a critical value for the number of frames—the detection boundary—above which it is possible to detect the presence of the signals, and below which it is impossible to reliably detect it. The detection boundary is explicitly specified and graphed.

In addition, we show that above the detection boundary, the likelihood ratio test would succeed by simply accepting the alternative when the log-likelihood ratio exceeds 0. We also show that the newly proposed Higher Criticism statistic in [11] is successful throughout the same region of number (of frames) vs. amplitude where the likelihood ratio test would succeed. Since Higher Criticism does not require a specification of the alternative, this implies that Higher Criticism is in a sense optimally adaptive for the above detection problem. The phenomenon found for the Gaussian setting above also exists for several non-Gaussian settings.

1. Introduction

Consider a situation in which many extremely noisy images are available. In each image frame, there is only one pixel containing a signal with all other pixels containing purely Gaussian noise. For any single frame, the signal is so faint that it is impossible to detect, and in addition, the position of the signal moves around randomly from frame to frame. The goal is to study how to detect a signal hidden in the extremely noisy background by combining all different frames together; i.e. by "multiple looks". This is a mathematical caricature of situations faced in two applied problems.

1. *Speckle Astronomy.* In earth-based telescope imaging of astronomical objects, atmospheric turbulence poses a fundamental obstacle. The image of the object

[1]Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA. e-mail: jinj@stat.purdue.edu

Keywords and phrases: detection boundary, Gaussian mixtures, higher criticism, infinitely divisible law, likelihood ratio, multiple looks.

AMS 2000 subject classifications: primary 62G10, 62F05; secondary 62E20.

is constantly moving around in the field of view; with a regular exposure time, an image of what should be a sharp point becomes highly blurred. A possible approach is to take many pictures with very short exposure time for each picture; the exposure time is so short that during exposure the position of the object hardly changes. However, this causes a new problem: the exposure time being so short that few photons accumulate, therefore we are unable to clearly see the object in any single frame. Technology nowadays enables us to easily collect hundreds or thousands of frames of pictures; from one frame to another, the position of the galaxy/star (if it exists) randomly moves around within the frame. The goal is to find out roughly at what amplitude it becomes possible to tell, from m realizations, that there is something present above usual background, see [2]. In this example, we are trying to detect, but not to estimate.

2. *Single Particle Electron Microscopy (SPEM).* In traditional crystallography, the image taken is actually the superposition of the scattering intensity across a huge number (10^{23}) of fundamental cells of the crystal, the superposed image lacks phase, and can only resolve the modulus of the Fourier Transform (FT) of the image. However we need to see images with phase correctly resolved. A possible solution to this is the *single particle EM*, see [25]. This method enables us to see correctly phased image from a single surface patch of frozen non-crystallized specimen; however this caused a new problem: the image is extremely noisy, there is little chance to see the molecule from any single image. On the other hand, technology nowadays can easily take large numbers (10^{10}) of different frames of image; however from one frame to the another, the position of the molecule randomly moves around the whole frame. However, by combining these huge numbers of frames of images, we hope we can reliably estimate the shape of the molecule. The question here is: what are the fundamental limits of resolution? If we can't "see" the molecule in any one image, and the molecule is moving around, can we still recover the image? In this example, the question is to estimate; however the first step for estimation is to make sure the things you want to estimate are actually there, and so problem of detection is an essential first step.

1.1. The multiple-looks model

Motivated by the examples in the previous section, suppose that we have independent observations $X_j^{(k)}$, $1 \le j \le n$, $1 \le k \le m$ (we reserve i for $\sqrt{-1}$), here j is the index for different pixels in each frame, and k is the index for different frames. As we have m frames and n pixels per frame, we have in total N observations, where

$$N \equiv m \cdot n. \tag{1.1}$$

For simplicity, assume that the signal, if it exists, is contained in one pixel for each frame. We want to tell which of the following two cases is true: whether each frame contains purely Gaussian noise, or that exactly one pixel per frame contains a signal (of fixed amplitude) but all other pixels are purely Gaussian noise and that the position of the signal randomly changes from frame to frame.

Formally, the observations obey:

$$X_j^{(k)} = \mu\delta_{j_0(k)}(j) + z_j^{(k)}, \qquad 1 \le j \le n, \qquad 1 \le k \le m, \tag{1.2}$$

where

$$z_j^{(k)} \overset{i.i.d}{\sim} N(0,1),$$

μ is the amplitude of the signal, and $j_0(k)$ is the position of the signal. Here for any fixed k, $j_0(k)$ is random variable taking values in $\{1,2,\ldots,n\}$ with equal probability, independent with each other as well as $z_j^{(k)}$, and where $\delta_{j_0(k)}(\cdot)$ is the Dirac sequence:

$$\delta_{j_0(k)}(j) = \begin{cases} 1, & j = j_0(k), \\ 0, & j \neq j_0(k). \end{cases} \tag{1.3}$$

The problem is to find out: given μ and n, what's the minimum value of $m = m^*$ such that we are able reliably to distinguish (1.2) from the pure noise model $X_j^{(k)} = z_j^{(k)}$.

Translating our problem into precise terms, we are trying to hypothesis test the following:

$$H_0 : X_j^{(k)} = z_j^{(k)}, \qquad 1 \le j \le n,\ 1 \le k \le m, \tag{1.4}$$

$$H_1^{(n,m)} : X_j^{(k)} = \mu\delta_{j_0(k)}(j) + z_j^{(k)}, \qquad 1 \le j \le n,\ 1 \le k \le m, \tag{1.5}$$

we call this testing model as *multiple-looks* model. Here, H_0 denotes the global intersection null hypothesis, and $H_1^{(n,m)}$ denotes a specific element in its complement. Under $H_1^{(n,m)}$, for each fixed k, there is only one observation $X_{j_0(k)}^{(k)}$ among $\{X_j^{(k)}\}_{j=1}^n$ containing a signal with amplitude μ, and the index $j_0(k)$ is sampled from the set $\{1,2,\ldots,n\}$ with equal probability, independently with k as well as $z_j^{(k)}$; in total, we have N observations which are normally distributed with zero mean, except m of them have a common nonzero mean μ.

Suppose we let $m = n^r$ for some exponent $0 < r < 1$ (or equivalently $m = N^{r/(1+r)}$). For r in this range, the number of nonzero means is too small to be noticeable in any sum which is in expectation of order N, so it cannot noticeably affect the behavior of bulk of the distribution. Let

$$\mu = \mu_n = \sqrt{2s\log n}, \qquad 0 < s < 1; \tag{1.6}$$

for s in this range, $\mu_n < \sqrt{2\log n}$, the nonzero means are, in expectation, smaller than the largest $X_j^{(k)}$ from the true null component hypotheses, so the nonzero means cannot have a visible effect on the upper extremes. For the calibrations we choose in this way, there is only a tiny fraction of observations with elevated mean, and the elevated mean is only of moderate significances.

1.2. Log-likelihood ratio and limit law

Obviously, with μ, n, and m fixed and known, the optimal procedure is the Neyman-Pearson likelihood ratio test (LRT), [28]. The log-likelihood ratio statistic for problems (1.4)–(1.5) is:

$$LR_{n,m} = \sum_{k=1}^{m} LR_n^{(k)},$$

where for any $1 \le k \le m$,

$$LR_n^{(k)} = LR_n^{(k)}\big(\mu, n; X_1^{(k)}, \ldots, X_n^{(k)}\big) \equiv \log\left(\frac{1}{n}\sum_{j=1}^{n} e^{\mu\cdot X_j^{(k)} - \mu^2/2}\right).$$

Fixed $0 < s < 1$ and n large, when $r \approx 0$ is relatively small, as the overall evidence for the existence of the signal is very weak, the null hypothesis and the alternative hypothesis merge together, and it is not possible to separate them; but when r gets larger, say $r \approx 1$, the evidence for the existence of the signal will get strong enough so that the null and the alternative separate from each other completely. Between the stage of "not separable" and "completely separable", there is a critical stage of "partly separable"; a careful study of this critical stage is the key for studying the problem of hypothesis testing (1.4)–(1.5).

In terms of log-likelihood ratio (LR), this particular critical stage of "partly separable" can be interpreted as, for any fixed s and $\mu_n = \sqrt{2s\log n}$, there is a critical number $m^* = m^*(n,s)$ such that as $n \to \infty$, LR_{n,m^*} converges weakly to non-degenerate distributions ν^0 and ν^1 under the null and the alternative respectively; since typically ν^0 and ν^1 overlap, the null and the alternative are partly separable.

This turns out to be true. In fact, we have the following theorem:

Theorem 1.1. *For parameter $0 < s < 1$, let $\mu_n = \mu_{n,s} = \sqrt{2s\log n}$, and*

$$m^* = m^*(n,s) \equiv \begin{cases} n^{1-2s}, & 0 < s \le 1/3, \\ \sqrt{2\pi}\cdot \mu_{n,s}\cdot n^{-(1-s)^2/(4s)}, & 1/3 < s < 1, \end{cases}$$

then as $n \to \infty$:

1. *When $0 < s < \frac{1}{3}$,*

$$\text{under } H_0 : LR_{n,m^*} \Longrightarrow N(-1/2, 1),$$
$$\text{under } H_1^{(n,m^*)} : LR_{n,m^*} \Longrightarrow N(1/2, 1).$$

2. *When $s = \frac{1}{3}$,*

$$\text{under } H_0 : LR_{n,m^*} \Longrightarrow N(-1/4, 1/2),$$
$$\text{under } H_1^{(n,m^*)} : LR_{n,m^*} \Longrightarrow N(1/4, 1/2).$$

3. *When $\frac{1}{3} < s < 1$,*

$$\text{under } H_0 : LR_{n,m^*} \Longrightarrow \nu_s^0, \qquad \text{under } H_1^{(n,m^*)} : LR_{n,m^*} \Longrightarrow \nu_s^1,$$

where ν_s^0 and ν_s^1 are distributions with characteristic functions $e^{\psi_s^0}$ and $e^{\psi_s^1}$ respectively, and

$$\psi_s^0(t) = \int_{-\infty}^{\infty} \big[e^{it\log(1+e^z)} - 1 - ite^z\big] e^{-\frac{1+s}{2s}z}\, dz, \tag{1.7}$$

$$\psi_s^1(t) = \psi_s^0(t) + \int_{-\infty}^{\infty} [e^{it\log(1+e^z)} - 1] e^{-\frac{1-s}{2s}z}\, dz. \tag{1.8}$$

In fact, the difference between LR_{n,m^*} under H_1^{n,m^*} and LR_{n,m^*} under H_0 weakly converges to 1, 1/2, and ν_s^* according to $s < 1/3$, $s = 1/3$ and $s > 1/3$, here ν_s^* is the distribution with characteristic function $e^{[\psi_s^1 - \psi_s^0]}$.

It was shown in [26, Chapter 2] that the laws ν_s^0 and ν_s^1 in Theorem 1.1 are in fact infinitely divisible. In Section 6.3, we discuss several other issues about ν_s^0

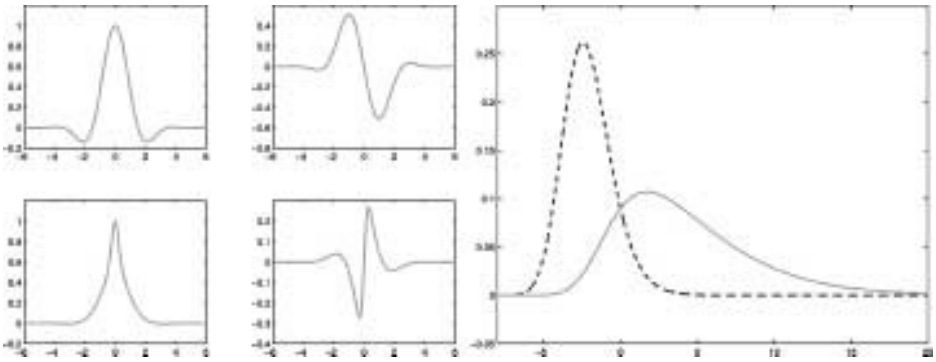


Figure 1: Left panel: Characteristic functions for $\nu^0_{.5}$ (top) and $\nu^1_{.5}$ (bottom). Left column: real parts, right column: imaginary parts. Right panel: Density functions for $\nu^0_{.5}$ (left) and $\nu^1_{.5}$ (right). The mean values of them are approximately -2.09 and 4.19, and variance of them are approximately 2.57 and 20 respectively.

and ν^1_s, where we view ν^0_s as a special example of $\nu^0_{s,\gamma}$, and ν^1_s as a special example of $\nu^1_{s,\gamma}$, with $\gamma = 2$. In short, both ν^0_s and ν^1_s have a bounded continuous density function, and a finite first moment as well a finite second moment. The mean value of ν^0_s is negative, and the mean value of ν^1_s is positive; in comparison, ν^0_s has a smaller variance than ν^1_s. In Figure 1, we plot the characteristic functions and density functions for ν^0_s and ν^1_s respectively with $s = 1/2$.

In [8], adapting to our notations, Burnashev and Begmatov studied the limiting behavior of $LR_{n,m}$ with $m = 1$, see more discussion in Section 7.3, as well as Section 4. In addition, the LRT and its optimality has been widely studied, see [6, 14], etc., and have also been discussed for various settings of detection of signals in a Gaussian noise setting, see [3, 4, 13], and also [29] for example.

1.3. Detection boundary

Theorem 1.1 implies that there is a *threshold effect* for the detection problem of (1.4)–(1.5). Dropping some lower order terms when necessary, (namely $\sqrt{2\pi} \cdot \mu_{n,s}$ in the case $1/3 < s < 1$), m^* would be reduced into a clean form: $m^* = n^{\rho^*(s)}$, where

$$\rho^*(s) = \begin{cases} 1 - 2s, & 0 < s \le 1/3, \\ \dfrac{(1-s)^2}{4s}, & 1/3 < s < 1. \end{cases} \tag{1.9}$$

Consider the curve $r = \rho^*(s)$ in the s–r plane. The curve separates the square $\{(s,r) : 0 < s < 1,\ 0 < r < 1\}$ into two regions: the region above the curve or the detectable region, and the region below the curve or the undetectable region; we call $r = \rho^*(s)$ the *detection boundary*. See the left panel of Figure 4 for illustrations, also see the left panel of Figure 5, where the curve corresponds to $\gamma = 2$ is $r = \rho^*(s)$. Theorem 1.1 implies that, roughly say, LR_{n,m^*} weakly converges to different non-degenerate distributions when (s,r) falls exactly on the detection boundary. We now study what will happen when (s,r) moves away from the detection boundary. On one hand, when (s,r) moves towards the interior of the detectable region, in comparison, we will have a lot more available observations while at the same time the amplitude is the same; so intuitively, $LR_{n,m}$ will put almost all mass at $-\infty$

under the null, and at ∞ under the alternative; this implies that the null and alternative separate from each other completely. On the other hand, when (s, r) moves towards the interior of the undetectable region, conversely, we have much fewer observations than we need, so the null and the alternative would both concentrate their mass around 0; more subtle analysis in Section 4 gives a much stronger claim: by appropriate normalization, $LR_{n,m}$ weakly converges to the *same* non-degenerated distribution, under H_0 as well as under $H_1^{(n,m)}$, and this non-degenerate distribution has a bounded continuous density function; thus the null and the alternative do completely merge together and are not separable. Precisely, we have the following Theorem. Recall that the Kolmogorov-Smirnov distance $\|\cdot\|_{KS}$ between any two cdf's G and G' is defined as:

$$\big\|G - G'\big\|_{KS} = \sup_t \big|G(t) - G'(t)\big|;$$

back to our notation $m = n^r$, here m depends only on n and r, which is not the critical $m^* = m^*(n, s)$ as in Theorem 1.1.

Theorem 1.2. *Let $\mu_n = \mu_{n,s} = \sqrt{2s\log n}$ and $m = n^r$.*

1. *When $r > \rho^*(s)$, consider the likelihood ratio test (LRT) that rejects H_0 when $LR_{n,m} > 0$, the sum of Type I and Type II errors tends to 0:*

$$P_{H_0}\{\textit{Reject } H_0\} + P_{H_1^{(n,m)}}\{\textit{Accept } H_0\} \to 0, \qquad n \to \infty.$$

2. *When $r < \rho^*(s)$,*

$$\lim_{n\to\infty} \big\|F_0^{(n,m)} - F_1^{(n,m)}\big\|_{KS} = 0,$$

where $F_0^{(n,m)}$ and $F_1^{(n,m)}$ are the cdf's of $LR_{n,m}$ under H_0 and $H_1^{(n,m)}$ respectively. As a result, for any test procedure, the sum of Type I and Type II errors tends to 1:

$$P_{H_0}\{\textit{Reject } H_0\} + P_{H_1^{(n,m)}}\{\textit{Accept } H_0\} \to 1, \qquad n \to \infty.$$

1.4. Higher criticism and optimal adaptivity

If we think of the s - r plane, $0 < s < 1$, $0 < r < 1$, we are saying that throughout the region $r > \rho^*(s)$, the alternative can be detected reliably using the likelihood ratio test (LRT). Unfortunately, as discussed in [11], the usual (Neyman-Pearson) likelihood ratio requires a precise specification of s and r, and misspecification of (s, r) may lead to failure of the LRT. Naturally, in any practical situation we would like to have a procedure which does well throughout this whole region without knowledge of (s, r). Hartigan [18] and Bickel and Chernoff [7] have shown that the usual generalized likelihood ratio test $\max_{\epsilon,\mu}\{[dP_1^{(n)}(\epsilon,\mu)/dP_0^{(n)}](X)\}$ has nonstandard behavior in this setting; in fact the maximized ratio tends to ∞ under H_0. It is not clear that this test can be relied on to detect subtle departures from H_0. Ingster [21] has proposed an alternative method of adaptive detection which maximizes the likelihood ratio over a finite but growing list of simple alternative hypotheses. By careful asymptotic analysis, he has in principle completely solved the problem of adaptive detection in the Gaussian mixture model (2.2)–(2.3) which we will introduce in Section 2; however, this is a relatively complex and delicate procedure which is tightly tied to the narrowly-specified Gaussian mixture model (2.2)–(2.3).

It would be nice to have an easily-implemented and intuitive method of detection which is able to work effectively throughout the whole region $0 < s < 1, r > \rho^*(s)$, which is not tied to the narrow model (2.2)–(2.3), and which is in some sense easily adapted to other (nonGaussian) mixture models. Motivated by these, we have developed a new statistic *Higher Criticism* in [11], where we have shown that the Higher Criticism statistic is optimally adaptive for detecting sparse Gaussian heterogeneous mixtures, as well as many other non-Gaussian settings.

To apply the Higher Criticism in our situation, let us convert the observations into the p-values. Let $p_j^{(k)} = P\{N(0,1) > X_j^{(k)}\}$ be the p-value for observation $X_j^{(k)}$, and let the $p_{(\ell)}$ denote the p-values *sorted in increasing order*, (recall $N = n \cdot m$):

$$p_{(1)} < p_{(2)} < \cdots < p_{(N)},$$

so that under the intersection null hypothesis the $p_{(\ell)}$ behave like order statistics from a uniform distribution. With this notation, the Higher Criticism is:

$$HC_N^* = \max_{1 \le \ell \le \alpha_0 \cdot N} \sqrt{N}[\ell/N - p_{(\ell)}]/\sqrt{p_{(\ell)}(1 - p_{(\ell)})},$$

where $0 < \alpha_0 < 1$ is any constant. Under the null hypothesis H_0, HC_N^* is related to the normalized uniform empirical process. Intuitively, under H_0, the p-values $p_j^{(k)}$ can be viewed as independent samples from $U(0,1)$. Adapting to the notations of [11], let $F_N(t) = \frac{1}{N}\sum_{\ell=1}^{N} 1_{\{p_{(\ell)} \le t\}}$, then the uniform empirical process is denoted by:

$$U_N(t) = \sqrt{N}[F_N(t) - t], \qquad 0 < t < 1,$$

and the normalized uniform empirical process by

$$W_N(t) = U_N(t)/\sqrt{t(1-t)}.$$

Under H_0, for each fixed t, $W_N(t)$ is asymptotically $N(0,1)$, and

$$HC_N^* = \max_{0<t<\alpha_0} W_N(t).$$

See [11] for more discussion. The following theorem is proved in [11]:

Theorem 1.3. *Under the null hypothesis H_0, as $N \to \infty$,*

$$\frac{HC_N^*}{\sqrt{2\log\log N}} \to_p 1.$$

It then follows if we threshold HC_N^* at $\sqrt{4\log\log N}$, the Type I error would equal to 0 asymptotically; moreover, thresholding at $\sqrt{4\log\log N}$ also gives a Type II error which equals to 0 asymptotically:

Theorem 1.4. *Consider the Higher Criticism test that rejects H_0 when*

$$HC_N^* > \sqrt{4\log\log N}. \tag{1.10}$$

For every alternative $H_1^{(n,m)}$ defined in (1.4)–(1.5) *above where r exceeds the detection boundary $\rho^*(s)$—so that the likelihood ratio test rejects H_0 at 0 would have negligible sum of Type I and Type II errors – the test based on Higher Criticism in* (1.10) *also has negligible sum of Type I and Type II errors:*

$$\left[P_{H_0}\{\textit{Reject } H_0\} + P_{H_1^{(n,m)}}\{\textit{Accept } H_0\}\right] \to 0, \qquad n \to \infty.$$

Roughly speaking, everywhere in the s–r plane where the likelihood ratio test would completely separate the two hypotheses asymptotically—the Higher Criticism will also completely separate the two hypotheses asymptotically; since it doesn't require any specification of parameters s and r, the Higher Criticism statistic is in some sense *optimally adaptive*. Of course, in the cases where the s–r relation falls *below* the detection boundary, all methods fail.

It is interesting to notice here the phenomena that the detection boundary $r = \rho^*(s)$ is partly linear ($s < 1/3$) and partly curved ($s > 1/3$); the curve only has up to the first order continuous derivatives at $s = 1/3$. As discussed in [11] or [26, Chapters 2–5], this phenomena implies that the detection problem of (1.4)–(1.5) is essentially different for the cases $0 < s \leq 1/3$ and $1/3 < s < 1$. Intuitively, when (s, r) is close to the curved part, statistics based on those a few largest observations would be able to effectively detect, while when (s, r) is close to the linear part, statistics based on a few largest observations (such as Max, Bonferroni, FDR) will fail, and only the newly proposed statistic Higher Criticism, or the Berk–Jones statistic which is asymptotically equivalent to the Higher Criticism in some sense [5, 11], is able to efficiently detect. As the study is similar to that in [11], we skip further discussion. However, in Section 2.2, we will explain this phenomenon from the angle of analysis.

1.5. Summary

We have considered a setting in which we have multiple frames of extremely noisy images, in each frame, hidden in the noise there may or may not be some signals, and the signal—when present—is too faint to be reliably detected from a single frame, and the position of the signal moves randomly across the whole frame. For fixed contrast size of the signal and the number of pixels in each frame, there is a critical number of frames—the detection boundary—above which combining all frames together gives a full power detection for the existence of the signal, and below which it is impossible to detect.

Above the detection boundary, the Neyman-Pearson LRT gives a full power detection. However, to implement LRT requires a specification of the parameters, and misspecification of the parameters may lead to the failure of the LRT. Motivated by this, we proposed a non-parametric statistic Higher Criticism in [11], which doesn't require such a specification of parameters; the Higher Criticism statistic gives asymptotically equal detection power to that of LRT. The Higher Criticism statistic only depends on p-values and can be used in many other settings.

Moreover, the detection boundary is partly linear and partly curved; compare the case when parameters are near the curved part and the case that the parameters are near the linear part, the detection problem is essentially different. Asymptotically, for the first case, statistics based on the largest a few observations are able to efficient to detect; however, for the second case, such statistics will totally fail, but the Higher Criticism statistic is still able to efficiently detect.

Below the detection boundary, asymptotically, all tests will completely fail for detection, even when all parameters are known.

The approach developed here seems applicable to a wide range of settings of non-Gaussian noise. In Section 6, we extend the Gaussian noise setting to the Generalized Gaussian noise setting.

1.6. Organization

The remaing part of the paper is organized as follows.

Sections 2–3 are for the proof of Theorem 1.1. In Section 2, we introduce a Gaussian mixture model, which we expect to be an "approximation" of the multiple-looks model, or Model 1.4–1.5; in comparison, this Gaussian mixture model is easier to study, and thus provides a bridge for studying the multiple-looks model. We then validate this expectation in Section 3 by showing that, with carefully chosen parameters, the difference between the log-likelihood ratios of these two models are indeed negligible; Theorem 1.1 is the direct result of those studies in Sections 2–3.

Second, we prove Theorem 1.2 in Section 4, and Theorem 1.4 in Section 5.

Next, in Section 6, we extend the study in Section 2 on the Gaussian mixture to non-Gaussian settings.

Finally, in Section 7, we briefly discuss several issues related to this paper. Section 8 is a technical Appendix.

2. Gaussian mixture model, and its connection to multiple looks model

Model (1.4)–(1.5) can be approximately translated into a Gaussian mixture model by "random shuffling". In fact, recall that the observations $\{X_j^{(k)}\}$ are collected frame by frame; suppose we arrange the $X_j^{(k)}$'s in a row according to the natural ordering:

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}, \ldots, X_1^{(m)}, X_2^{(m)}, \ldots, X_n^{(m)},$$

we then randomly shuffle them and rearrange back into frames, according to the ordering after the shuffling; we denote the resulting observations by $\{\tilde{X}_j^{(k)} : 1 \le j \le n, 1 \le k \le m\}$.

Of course under H_0, the above random shuffling won't have any effect and the joint distribution of $X_j^{(k)}$ is the same as that of $\{\tilde{X}_j^{(k)}\}$. However, if $H_1^{(n,m)}$ is true, then $\tilde{X}_j^{(k)}$ will have a slightly different distribution than that of $X_j^{(k)}$, which, approximately, can be viewed as sampled from a Gaussian mixture:

$$\bar{X}_j^{(k)} \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon N(\mu, 1), \qquad 1 \le j \le n, \qquad 1 \le k \le m, \tag{2.1}$$

with

$$\epsilon = \epsilon_n = n^{-1}, \qquad \mu = \mu_n = \mu_{n,s} = \sqrt{2s\log n}.$$

The difference between $\{X_j^{(k)}\}$ and $\{\bar{X}_j^{(k)}\}$ is that under $H_1^{n,m}$, $\{X_j^{(k)}\}$ has exactly a fraction $1/n$ of nonzero means in each frame while the $\{\bar{X}_j^{(k)}\}$ has such a fraction only in expectation. Moreover, the problem of hypothesis testing the multiple looks model (1.4)–(1.5) is approximately equivalent to hypothesis testing:

$$H_0 : \bar{X}_j^{(k)} \overset{i.i.d}{\sim} N(0,1), \qquad 1 \le j \le n,\ 1 \le k \le m, \tag{2.2}$$

$$\mathcal{H}_1^{(n,m)} : \bar{X}_j^{(k)} \overset{i.i.d}{\sim} (1-1/n)N(0,1) + (1/n)N(\mu_n, 1), \qquad 1 \le j \le n,\ 1 \le k \le m. \tag{2.3}$$

In this paper, we refer this model as the Gaussian mixture model, in contrast to the multiple-looks model. Since the random shuffling has no effect on the null hypothesis, we still use H_0 to denote the null hypothesis; however, we use $\mathcal{H}_1^{n,m}$ to denote the new alternative hypothesis. Moreover, we denote the likelihood ratio

statistic of Model (2.2)–(2.3) by $\mathcal{LR}_{n,m}$, in contrast to $LR_{n,m}$ of Model (1.4)–(1.5). Notice here:

$$\mathcal{LR}_{n,m} = \mathcal{LR}_{n,m}\big(\mu_n, n; \bar{X}_1^{(1)}, \ldots, \bar{X}_n^{(1)}, \ldots, \bar{X}_1^{(m)}, \ldots, \bar{X}_n^{(m)}\big) = \sum_{k=1}^{m}\sum_{j=1}^{n} \mathcal{LR}_j^{(k)},$$

where

$$\mathcal{LR}_j^{(k)} = \mathcal{LR}\big(\mu_n, n; \bar{X}_j^{(k)}\big) \equiv \log\left(1 - \frac{1}{n} + \frac{1}{n}e^{\mu_n \bar{X}_j^{(k)} - \mu_n^2/2}\right).$$

There are two important reasons for introducing the Gaussian mixture model above. First, as the multiple-looks model can be converted into the Gaussian mixture model by random shuffling, we expect that these two models are closely related. In fact, compare the two log-likelihood ratios: $LR_{n,m}$ and $\mathcal{LR}_{n,m}$: on one hand, as we will see in Section 3, with particularly chosen parameters (s, r), the difference between $LR_{n,m}$ and $\mathcal{LR}_{n,m}$ is in fact negligible; on the other hand, clearly, $\mathcal{LR}_{n,m}$ has a much simpler form than that of $LR_{n,m}$, and thus it is much easier to analyze $\mathcal{LR}_{n,m}$ than $LR_{n,m}$. In short, the study of the Gaussian mixture model will provide an important bridge for studying the multiple-looks model.

The second important reason is that, the Gaussian mixture model itself is of importance and has many interesting applications. In [11], we mentioned three application areas where situations as in Model (2.2)–(2.3) might arise: *early detection of bio-weapons use*, *detection of covert communications*, and *meta-analysis with heterogeneity*. There are many other potential applications in signal processing e.g., [22, 23, 24].

The main result on the problem of hypothesis testing the Gaussian mixture model, or Model (2.2)–(2.3) is the following.

Theorem 2.1. *For parameter $0 < s < 1$, let $\mu_n = \mu_{n,s} = \sqrt{2s\log n}$ and*

$$m^* = m^*(n,s) = \begin{cases} n^{1-2s}, & 0 < s \le 1/3, \\ \sqrt{2\pi}\cdot \mu_{n,s}\cdot n^{-(1-s)^2/(4s)}, & 1/3 < s < 1, \end{cases}$$

then as $n \to \infty$,

1. *When $0 < s < 1/3$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(-1/2, 1), \qquad \text{under } H_0,$$
$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(1/2, 1), \qquad \text{under } \mathcal{H}_1^{n,m^*}.$$

2. *When $s = 1/3$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(-1/4, 1/2), \qquad \text{under } H_0,$$
$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(1/4, 1/2), \qquad \text{under } \mathcal{H}_1^{n,m^*}.$$

3. *When $1/3 < s < 1$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} \nu_s^0, \quad \text{under } H_0, \qquad \mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} \nu_s^1, \quad \text{under } \mathcal{H}_1^{n,m^*},$$

where ν_s^0 and ν_s^1 are the same as in Theorem 1.1.

Similarly, there is a *threshold effect* for the hypothesis testing of the Gaussian mixture model, and so the detection boundary. In the s-r plane, the detection boundary of the Gaussian mixture model is:

$$r = \rho^*(s),$$

which is exactly the same as that of the multiple-looks model; see more discussion on the Gaussian mixture model in [11].

Ingster [20] studied a similar problem and noticed similar threshold phenomena, see more discussions in Section 7.3. There are many other studies on the detection of Gaussian mixtures using LRT, see [9, 16], and [17] for example.

2.1. Proof of Theorem 2.1

For the proof of Theorem 2.1, the approach below is developed independently and is different from that in [20]; the approach below is also generalized to the settings of non-Gaussian mixture which we will discuss in Section 6.

Denote the density function of $N(0,1)$ by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{2.4}$$

To prove Theorem 2.1, we start with the following key lemma:

Lemma 2.1. *With $\mu_n = \mu_{n,s}$ as defined in Theorem* 2.1*,*

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big[e^{it\log(1+e^z)} - 1 - ite^z\big] e^{-\frac{1+s}{2s}z} \phi\Big(\frac{z}{\mu_n}\Big)\, dz \\
&\quad = \begin{cases}
-\dfrac{it+t^2+o(1)}{2} \cdot \mu_n \cdot n^{\frac{(1-3s)^2}{4s}}, & 0 < s < 1/3, \\
-\dfrac{it+t^2+o(1)}{4} \cdot \mu_n, & s = 1/3, \\
\dfrac{1}{\sqrt{2\pi}} \psi_s^0(t) + o(1), & 1/3 < s < 1,
\end{cases}
\end{aligned} \tag{2.5}$$

and

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big[e^{it\log(1+e^z)} - 1\big] e^{-\frac{1-s}{2s}z} \phi\Big(\frac{z}{\mu_n}\Big)\, dz \\
&\quad = \begin{cases}
\big(it + o(1)\big) \cdot \mu_n \cdot n^{\frac{(1-3s)^2}{4s}}, & 0 < s < 1/3, \\
\dfrac{it+o(1)}{2} \cdot \mu_n, & s = 1/3, \\
\dfrac{1}{\sqrt{2\pi}} \big[\psi_s^1(t) - \psi_s^0(t)\big] + o(1), & 1/3 < s < 1,
\end{cases}
\end{aligned} \tag{2.6}$$

where $\psi_s^0(t)$ and $\psi_s^1(t)$ are defined in Theorem 1.1*.*

Let $N^* = N^*(n,s) = n \cdot m^*(n,s)$, to prove Theorem 2.1, it is sufficient to show that:

$$\text{under } H_0: \qquad Ee^{it\mathcal{L}\mathcal{R}_j^{(k)}} = \begin{cases}
1 - \big(it + t^2 + o(1)\big)/\big(2N^*\big), & 0 < s < \dfrac{1}{3}, \\
1 - \big(it + t^2 + o(1)\big)/\big(4N^*\big), & s = \dfrac{1}{3}, \\
1 + \big(\psi_s^0(t) + o(1)\big)/N^*, & \dfrac{1}{3} < s < 1,
\end{cases} \tag{2.7}$$

and

$$\text{under } \mathcal{H}_1^{(n,m^*)}: \qquad Ee^{it\mathcal{LR}_j^{(k)}} = \begin{cases} 1+\big(it-t^2+o(1)\big)/\big(2N^*\big), & 0<s<\dfrac{1}{3}, \\ 1+\big(it-t^2+o(1)\big)/\big(4N^*\big), & s=\dfrac{1}{3}, \\ 1+\big(\psi_s^1(t)+o(1)\big)/N^*, & \dfrac{1}{3}<s<1; \end{cases} \tag{2.8}$$

in fact, by $Ee^{it\mathcal{LR}_{n,m^*}} = (Ee^{it\mathcal{LR}_j^{(k)}})^{N^*}$, a direct result of (2.7)–(2.8) is that as $n\to\infty$, we have the following point-wise convergences:

$$\text{under } H_0: \qquad Ee^{it\mathcal{LR}_{n,m^*}} \to \begin{cases} e^{-(it+t^2)/2}, & 0<s<1/3, \\ e^{-(it+t^2)/4}, & s=1/3, \\ e^{\psi_s^0}, & 1/3<s<1, \end{cases}$$

and

$$\text{under } \mathcal{H}_1^{(n,m^*)}: \qquad Ee^{it\mathcal{LR}_{n,m^*}} \to \begin{cases} e^{(it-t^2)/2}, & 0<s<1/3, \\ e^{(it-t^2)/4}, & s=1/3, \\ e^{\psi_s^1}, & 1/3<s<1, \end{cases}$$

and Theorem 2.1 follows.

We now show (2.7). Under H_0, notice that:

$$Ee^{it\mathcal{LR}_j^{(k)}} = \int_{-\infty}^{\infty} e^{it\log(1-1/n+(1/n)e^{\mu_n z-\mu_n^2/2})}\phi(z)\,dz \tag{2.9}$$

$$= e^{it\log(1-1/n)}\cdot\int_{-\infty}^{\infty} e^{it\log(1+\frac{1}{n-1}e^{\mu_n z-\mu_n^2/2})}\phi(z)\,dz; \tag{2.10}$$

rewrite:

$$\int_{-\infty}^{\infty} e^{it\log(1+\frac{1}{n-1}e^{\mu_n z-\mu_n^2/2})}\phi(z)\,dz \tag{2.11}$$

$$= 1+\frac{it}{n}+\int_{-\infty}^{\infty}\big[e^{it\log(1+(1/n)e^{\mu_n z-\mu_n^2/2})}-1-it\cdot(1/n)e^{\mu_n z-\mu_n^2/2}\big]\phi(z)\,dz + O\big(1/n^2\big); \tag{2.12}$$

the key of the analysis is using the substitution $e^{z'} = (1/n)e^{\mu_n z-\mu_n^2/2}$:

$$\int_{-\infty}^{\infty}\big[e^{it\log(1+(1/n)e^{\mu_n z-\mu_n^2/2})}-1-it\cdot(1/n)e^{\mu_n z-\mu_n^2/2}\big]\phi(z)\,dz \tag{2.13}$$

$$= \frac{1}{\mu_n}e^{-\frac{(1+s)^2}{8s^2}\mu_n^2}\int_{-\infty}^{\infty} e^{-\frac{1+s}{2s}z}\big[e^{it\log(1+e^z)}-1-ite^z\big]\phi\left(\frac{z}{\mu_n}\right)dz; \tag{2.14}$$

combining (2.9)–(2.14) with Lemma 2.1 gives (2.7).

The proof of (2.8) is similar. Under $\mathcal{H}_1^{n,m^*}$,

$$Ee^{it\mathcal{LR}_j^{(k)}} = (1-1/n)\cdot\int_{-\infty}^{\infty} e^{it\log(1-1/n+(1/n)e^{\mu_n z-\mu_n^2/2})}\phi(z)\,dz \tag{2.15}$$

$$+(1/n)\cdot\int_{-\infty}^{\infty} e^{it\log(1-1/n+(1/n)e^{\mu_n z-\mu_n^2/2})}\phi(z-\mu_n)\,dz, \tag{2.16}$$

the first term can be analyzed similarly as in the case under H_0, as for the second term, similarly we have:

$$\int_{-\infty}^{\infty} e^{it\log(1-1/n+(1/n)e^{\mu_n z-\mu_n^2})}\phi(z-\mu_n)\,dz \tag{2.17}$$

$$= 1+\int_{-\infty}^{\infty}\big[e^{it\log(1+(1/n)e^{\mu_n z+\mu_n^2/2})}-1\big]\phi(z)\,dz+O(1/n) \tag{2.18}$$

$$= 1+\frac{1}{\mu_n}e^{-\frac{(1-s)^2}{8s^2}\mu_n^2}\int_{-\infty}^{\infty}\big[e^{it\log(1+e^z)}-1\big]e^{-\frac{1-s}{2s}z}\phi\Big(\frac{z}{\mu_n}\Big)\,dz+O(1/n), \tag{2.19}$$

combining (2.15)–(2.19) with (2.9) and Lemma 2.1 gives (2.10).

This concludes the proof of Theorem 2.1. □

2.2. Proof of Lemma 2.1

As we mentioned before, an interesting phenomenon for the detection of the multiple-looks model is that, the detection boundary is partly linear and partly curved; the whole curve only has up to the first order continuous derivatives. As the intuition for why this phenomenon happens had been developed in [11], here we try to understand the phenomenon from the angle of analysis.

In fact, take (2.5) for example, as $\mu_n\to\infty$, the integration

$$\int_{-\infty}^{\infty}\big[e^{it\log(1+e^z)}-1-ite^z\big]e^{-\frac{1+s}{2s}z}\phi\Big(\frac{z}{\mu_n}\Big)\,dz \tag{2.20}$$

behaves totally different for the cases $0<s<1/3$ and $1/3<s<1$. The reason is that, by dropping the term $\phi(z/\mu_n)$, the integrand in (2.20) is absolute integrable if and only if $(1+s)/(2s)<2$, or equivalently $1/3<s<1$; to see this, notice that the only possible place could make the integration to diverge is $z=-\infty$, observe that when $z<0$ and $|z|$ very large:

$$e^{it\log(1+e^z)}-1-ite^z\sim e^{2z}, \tag{2.21}$$

it immediately follows that the integration diverges if and only if $(1+s)/2s<2$, or $1/3<s<1$.

As a result, when $1/3<s<1$, (2.5) follows easily by Dominated Convergence Theorem. In fact, recall the definition of ψ_s^0 and by noticing the point-wise convergence of $\phi(z/\mu_n)$ to $1/\sqrt{2\pi}$, we have:

$$\int_{-\infty}^{\infty}e^{-\frac{1+s}{2s}z}\big[e^{it\log(1+e^z)}-1-ite^z\big]\phi\Big(\frac{z}{\mu_n}\Big)=\frac{1}{\sqrt{2\pi}}\psi_s^0(t)+o(1).$$

However, when $0<s\le 1/3$, the integration goes to to ∞ as $\mu_n\to\infty$, so we need to analyze differently. In fact, using (2.21), we have:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\big[e^{it\log(1+e^z)}-1-ite^z\big]e^{-\frac{1+s}{2s}z}\phi\Big(\frac{z}{\mu_n}\Big)\,dz\\
&\quad=\int_{-\infty}^{0}\big[e^{it\log(1+e^z)}-1-ite^z\big]e^{-\frac{1+s}{2s}z}\phi\Big(\frac{z}{\mu_n}\Big)\,dz+O(1)\\
&\quad=-\frac{1}{2}\big(it+t^2\big)\bigg[\int_{-\infty}^{0}e^{2z}\cdot e^{-\frac{1+s}{2s}z}\cdot\phi\Big(\frac{z}{\mu_n}\Big)\,dz\bigg]\big(1+o(1)\big)+O(1)\\
&\quad=-\frac{1}{2}\big(it+t^2\big)\mu_n e^{\frac{(1-3s)^2}{8s^2}\mu_n^2}\big(1+o(1)\big).
\end{aligned}$$

The remaining part of the proof is similar, so we skip it. See [26, Chapter 2] for a more detailed proof. □

3. Proof of Theorem 1.1

As we mentioned in Section 2, the multiple-looks model (1.4)–(1.5) can be converted into the Gaussian mixture model (2.2)–(2.3) by random shuffling, we thus expect the difference between the log-likelihood ratios LR_{n,m^*} and $\mathcal{LR}_{n,m^*}$ to be negligible, or

$$LR_{n,m^*} = \mathcal{LR}_{n,m^*} + o_p(1). \tag{3.1}$$

As a result, the limiting behavior of LR_{n,m^*} would be asymptotically the same as that of $\mathcal{LR}_{n,m^*}$ in Theorem 2.1.

Motivated by these, our approach for proving Theorem 1.1 is to, first validate (3.1), and then, combine (3.1) with Theorem 2.1.

We now show the cases under H_0 and under H_1^{n,m^*} separately.

First, under H_0. For $z_j^{(k)} \overset{iid}{\sim} N(0,1)$, $1 \le j \le n, 1 \le k \le m$, let:

$$v^{(k)} = v^{(k)}\big(\mu_n, n; z_1^{(k)}, z_2^{(k)}, \ldots, z_n^{(k)}\big) \triangleq \frac{1}{n}\left[\sum_{j=1}^{n} e^{\mu_n \cdot z_j^{(k)} - \mu_n^2/2}\right], \tag{3.2}$$

$$u^{(k)} = u^{(k)}\big(\mu_n, n; z_1^{(k)}, z_2^{(k)}, \ldots, z_n^{(k)}\big) \triangleq \left(\prod_{j=1}^{n}\left[1 - \frac{1}{n} + \frac{1}{n}e^{\mu_n \cdot z_j^{(k)} - \mu_n^2/2}\right]\right) - v^{(k)}, \tag{3.3}$$

then under H_0, by symmetry:

$$LR_{n,m^*} = \sum_{k=1}^{m^*} \log\big(v^{(k)}\big), \qquad \mathcal{LR}_{n,m^*} = \sum_{k=1}^{m^*} \log\big(u^{(k)} + v^{(k)}\big);$$

intuitively, since for a sequence of small numbers a_j, $\prod_{j=1}^{n}(1+a_j) \approx 1 + \sum_1^n a_j$, so:

$$u^{(k)} + v^{(k)} \approx 1 + \sum_{j=1}^{n}\left[-\frac{1}{n} + \frac{1}{n}e^{\mu_n z_j^{(k)} - \mu_n^2/2}\right] = v^{(k)};$$

we thus expect that the difference between LR_{n,m^*} and $\mathcal{LR}_{n,m^*}$ is indeed negligible. Let

$$w^{(k)} \triangleq \frac{u^{(k)}}{v^{(k)}}, \tag{3.4}$$

then:

$$\mathcal{LR}_{n,m^*} - LR_{n,m^*} = \sum_{k=1}^{m^*} \log\big(1 + w^{(k)}\big),$$

the following Lemma validates the heurism, or (3.1), under the null hypothesis H_0:

Lemma 3.1. *If $z_j^{(k)} \overset{i.i.d}{\sim} N(0,1)$, $1 \le j \le n$, $1 \le k \le m$, then for $\mu_n = \sqrt{2s\log(n)}$ and*

$$m^* = \begin{cases} n^{(1-2s)}, & 0 < s \le \dfrac{1}{3}, \\ \sqrt{2\pi} \cdot \mu_n \cdot n^{\frac{(1-s)^2}{4s}}, & \dfrac{1}{3} < s < 1, \end{cases}$$

we have:

$$\sum_{k=1}^{m^*} \log\big(1 + w^{(k)}\big) \to_p 0.$$

Combining Lemma 3.1 with Theorem 2.1 gives Theorem 1.1 under H_0.

Now under H_1^{n,m^*}, $X_j^{(k)} = \mu\delta_{j_0(k)}(j) + z_j^{(k)}$, where $j_0(k)$ uniformly distributed over $\{1, 2, \ldots, n\}$; so by symmetry:

$$LR_{n,m^*} =_D \sum_{k=1}^{m^*}\left[\log\left(\frac{1}{n}\left[e^{\mu_n z_1^{(j)}+\mu_n^2/2} + \sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}\right]\right)\right],$$

and we can rewrite:

$$\begin{aligned} LR_{n,m^*} &= \left[\sum_{k=1}^{m^*}\log\left(\frac{1}{n}\sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}\right)\right] \\ &\quad + \left[\sum_{k=1}^{m^*}\log\left(1 + \frac{1}{[\sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}]/n}\cdot\frac{1}{n}e^{\mu_n z_1^{(k)}+\mu_n^2/2}\right)\right]. \quad (3.5)\end{aligned}$$

By the study for the case under H_0, the first term on the right hand side above weakly converges to:

$$\sum_{k=1}^{m^*}\log\left(\frac{1}{n}\sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}\right) \stackrel{w}{\Longrightarrow} \begin{cases} N(-1/2, 1), & 0 < s < 1/3, \\ N(-1/4, 1/2), & s = 1/3, \\ \nu_s^0, & 1/3 < s < 1, \end{cases} \quad (3.6)$$

with ν_s^0 defined in Theorem 1.1, so all we need to study is the second term. The following Lemma is proved in [26, Chapter 4].

Lemma 3.2. *Fixed $0 < a < \frac{1}{2}$, with $\mu_n = \mu_{n,s} = \sqrt{2s\log n}$, then for $z_j^{(k)} \stackrel{iid}{\sim} N(0,1)$, $1 \le j \le n$,*

$$P\{v^{(k)} \le a\} \le 2e^{-[\frac{(2a-1)^2}{8}\mu_n\cdot n^{(1-s)}(1+o(1))]}, \qquad n \to \infty, \qquad \text{for any } k \ge 1.$$

With some elementary analysis, Lemma 3.2 implies:

$$\frac{1}{v^{(k)}} \to 1, \qquad \text{in probability and in } L^p, \qquad \forall p > 0. \quad (3.7)$$

Now back to the second term on the right hand side of (3.5), or:

$$\left[\sum_{k=1}^{m^*}\log\left(1 + \frac{1}{[\sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}]/n}\cdot\frac{1}{n}e^{\mu_n z_1^{(k)}+\mu_n^2/2}\right)\right];$$

inspired by(3.7), we expect that there will be only a negligible change if we replace the messy term $[(1/n)\sum_{j=2}^{n} e^{\mu_n z_j^{(k)}-\mu_n^2/2}]$ by 1 for all k; this turns out to be true, and we have the following lemma:

Lemma 3.3. *For $\mu_n = \mu_{n,s}$ and $m^* = m^*(n,s)$ defined in Theorem* 1.1*, if $z_j^{(k)} \stackrel{i.i.d}{\sim} N(0,1)$, $1 \le j \le n$, $1 \le k \le m^*$, then:*

$$\sum_{k=1}^{m^*}\left[\log\left(1 + \frac{1}{n}e^{\mu_n\cdot z_1^{(k)}+\mu_n^2/2}\right) - \log\left(1 + \frac{1}{\frac{1}{n}\sum_{j=2}^{n} e^{\mu_n\cdot z_j^{(k)}-\mu_n^2/2}}\cdot\frac{1}{n}e^{\mu_n\cdot z_1^{(k)}+\mu_n^2/2}\right)\right] \to_p 0.$$

Applying Lemma 3.3 directly to (3.5) gives:

$$LR_{n,m^*} =_D \left[\sum_{k=1}^{m^*}\log\left(\frac{1}{n}\sum_{j=2}^{n}e^{\mu_n z_j^{(k)}-\mu_n^2/2}\right)\right]+\left[\sum_{k=1}^{m^*}\log\left(1+\frac{1}{n}e^{\mu_n z_1^{(k)}+\mu_n^2/2}\right)\right]+o_p(1). \tag{3.8}$$

But for the second term in (3.8), observe that for any t, by substitution $e^{z'} = e^{\mu_n z_1^{(1)}+\mu_n^2/2}$,

$$E\big[e^{it\log(1+\frac{1}{n}e^{\mu_n z_1^{(1)}+\mu_n^2/2})}\big] = 1+\frac{1}{\mu_n}e^{-\frac{(1-s)^2}{8s^2}\mu_n^2}\cdot\int\big[e^{it\log(1+e^z)}-1\big]e^{-\frac{1-s}{2s}z}\phi(z/\mu_n)\,dz,$$

by independency:

$$E\big[e^{it\cdot\sum_{k=1}^{m^*}\log(1+\frac{1}{n}e^{\mu_n z_1^{(k)}+\mu_n^2/2})}\big] = \big(E\big[e^{it\log(1+\frac{1}{n}e^{\mu_n z_1^{(1)}+\mu_n^2/2})}\big]\big)^{m^*},$$

we then derive:

$$\sum_{k=1}^{m^*}\log\left(1+\frac{1}{n}e^{\mu_n\cdot z_1^{(k)}+\mu_n^2/2}\right)\overset{w}{\Longrightarrow}\begin{cases}1, & 0<s<1/3,\\ 1/2, & s=1/3,\\ \nu_s^*, & 1/3<s<1,\end{cases} \tag{3.9}$$

where ν_s^* is the distribution with characteristic function $e^{[\psi_s^1(t)-\psi_s^0(t)]}$; inserting (3.6) and (3.9) into (3.8) gives the proof of Theorem 1.1 under H_1^{n,m^*}. □

3.1. Proof of Lemma 3.1

A detailed proof of Lemma 3.1 is available in [26, Chapter 4]. In this section, we will only illustrate the main ideas for the proof, while skipping the technical details. Direct calculations show that:

$$1+w^{(k)} \geq (1-1/n)^n\cdot\frac{\prod_{j=1}^{n}[1+(1/n)e^{\mu_n\cdot z_j^{(k)}-\mu_n^2/2}]}{\frac{1}{n}\sum_{j=1}^{n}[e^{\mu_n\cdot z_j^{(k)}-\mu_n^2/2}]}\geq(1-1/n)^n,$$

so when $n\geq 2$, there is a constant $C>0$, such that:

$$\big|\log(1+w^{(k)})-w^{(k)}\big|\leq C\cdot\big(w^{(k)}\big)^2,$$

and to show Lemma 3.1, it is sufficient to show:

$$\sum_{k=1}^{m^*}w^{(k)}\to_p 0,\qquad \sum_{k=1}^{m^*}\big[w^{(k)}\big]^2\to_p 0. \tag{3.10}$$

Split:

$$w^{(k)} = u^{(k)}+u^{(k)}\cdot\left(\frac{1}{v^{(k)}}-1\right)\cdot 1_{\{v^{(k)}\geq 1/3\}}+u^{(k)}\cdot\left(\frac{1}{v^{(k)}}-1\right)\cdot 1_{\{v^{(k)}<1/3\}},$$

and

$$\big[w^{(k)}\big]^2 = \big[w^{(k)}\big]^2\cdot 1_{\{v^{(k)}<1/3\}}+\big[w^{(k)}\big]^2\cdot 1_{\{v^{(k)}\geq 1/3\}};$$

using Lemma 3.2, the remaining part of the proof is careful analysis, see [26, Chapter 4] for details. □

3.2. Proof of Lemma 3.3

It is sufficient to show:

$$\sum_{k=1}^{m^*}\left[\log\left(1+\frac{1}{n}e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}\right)-\log\left(1+\frac{1}{v^{(k)}}\cdot\frac{1}{n}e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}\right)\right]\to_p 0,$$

where $z^{(k)}\overset{iid}{\sim}N(0,1)$ and are independent of $\{v^{(k)}\}_{k=1}^{m^*}$. But since for any $x,y\geq 0$, $\log(1+x)-\log(1+y)=(x-y)/(1+x)+r(x,y)$, where the reminder term $|r(x,y)|\leq C(x-y)^2$ for some constant C, so all we need to show is as $n\to\infty$:

$$\sum_{k=1}^{m^*}\left[\frac{(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}{1+(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}\left(\frac{1}{v^{(k)}}-1\right)\right]\to_p 0, \tag{3.11}$$

and

$$\sum_{k=1}^{m^*}\left[\left(\frac{1}{v^{(k)}}-1\right)\cdot(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}\right]^2\to_p 0; \tag{3.12}$$

or equivalently, for any fixed t:

$$\begin{aligned} Ee^{it[\frac{(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}{1+(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}(\frac{1}{v^{(k)}}-1)]} &= 1+o\left(\frac{1}{m^*}\right),\\ Ee^{it[\frac{1}{n}e^{\mu_n\cdot z^{(k)}+\frac{\mu_n^2}{2}}\cdot(\frac{1}{v^{(k)}}-1)]^2} &= 1+o\left(\frac{1}{m^*}\right). \end{aligned} \tag{3.13}$$

Similar to the proof of Theorem 2.1, using substitution $e^{z'}=\frac{1}{n}e^{\mu_n\cdot z+\mu_n^2/2}$, we then rewrite:

$$\begin{aligned} &E\left(e^{it[\frac{(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}{1+\frac{1}{n}e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}}(\frac{1}{v^{(k)}}-1)]}-1\right)\\ &\quad=\frac{1}{\mu_n}\mu_n^{-\frac{(1-s)^2}{8s^2}\mu_n^2}\int_{-\infty}^{\infty}E\left[e^{it(v^{(k)}-1)\frac{e^z}{1+e^z}}-1\right]e^{-\frac{1-s}{2s}z}\cdot\phi\left(\frac{z}{\mu_n}\right)dz, \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} &E\left(e^{it[(\frac{1}{v^{(k)}}-1)\cdot(1/n)e^{\mu_n\cdot z^{(k)}+\mu_n^2/2}]^2}-1\right)\\ &\quad=\frac{1}{\mu_n}\mu_n^{-\frac{(1-s)^2}{8s^2}\mu_n^2}\int_{-\infty}^{\infty}E\left[e^{it[(v^{(k)}-1)e^z]^2}-1\right]e^{-\frac{1-s}{2s}z}\cdot\phi\left(\frac{z}{\mu_n}\right)dz, \end{aligned} \tag{3.15}$$

where on the right hand side, the expectation inside the integral sign is with respect to the law of $v^{(k)}$. Again by Lemma 3.2, the remaining part of the proof is careful analysis. See [26, Chapter 4] for the technical details. This concludes the proof of Lemma 3.3. □

4. Proof of Theorem 1.2

We prove Theorem 1.2 for the cases $r>\rho^*(s)$ and $0<r<\rho^*(s)$ separately.

For the case $r > \rho^*(s)$, by the definition of m^* and m, for (s, r) in this range, $m/m^* \to \infty$ as $n \to \infty$. First we consider the case under H_0, let:

$$\begin{aligned} a_n &= \begin{cases} \sqrt{m/m^*}, & 0 < s < 1/3, \\ \sqrt{m/(2m^*)}, & s = 1/3, \\ \sqrt{m/m^*} \cdot \sqrt{-(\psi_s^0)''(0)}, & 1/3 < s < 1, \end{cases} \\ b_n &= -\begin{cases} m/(2m^*), & 0 < s < 1/3, \\ m/(4m^*), & s = 1/3, \\ (m/m^*)(-\psi_s^0)'(0), & 1/3 < s < 1; \end{cases} \end{aligned} \tag{4.1}$$

roughly say, b_n is the mean value of $LR_{n,m}$, and a_n is the standard deviation of $LR_{n,m}$. By Theorem 1.1 and elementary analysis, it follows that $[LR_{n,m} - b_n]/a_n \stackrel{w}{\Longrightarrow} N(0,1)$, and thus $LR_{n,m}/\sqrt{m/m^*} \to_p -\infty$ under H_0. Similar argument shows $LR_{n,m}/\sqrt{m/m^*} \to_p \infty$ under $H_1^{(n,m)}$, this concludes the proof of Theorem 1.2 in this case.

We now consider the case $r > \rho^*(s)$. First, we briefly explain why the proof is non-trivial. Recall that, $LR_{n,m}$ converges to 0 in probability, under the null as well as under the alternative—which is a direct result of the studies of Sections 2–3; however, this claim alone is not sufficient for proving Theorem 1.2 in this case: the Kolmogorov–Smirnov distance between two random sequences could tend to 1 even when both of them tend to 0 in probability, the culprit is the discontinuity of the cdf function of ν_0 (here ν_0 denote the point mass with all mass at 0).

However, recall that given a cdf F which is a continuous function, then for any sequence of cdf's such that $F_n \stackrel{w}{\Longrightarrow} F$, we have:

$$\lim_{n\to\infty} \|F_n - F\|_{KS} = 0, \tag{4.2}$$

see, for example, [12]. Motivated by this, we need a stronger claim of the limiting behavior of $LR_{n,m}$. Namely, for any fixed (s, r) in this range, we hope to find a sequence of numbers $\{\ell_n = \ell_{n,s,r}\}_{n=1}^{\infty}$ such that:

$$\ell_n \cdot LR_{n,m} \stackrel{w}{\Longrightarrow} F, \tag{4.3}$$

both under the H_0 and $H_1^{(n,m)}$, where F is some continuous cdf function.

This turns out to be true. Consider the following sub-regions of the undetectable region $\{(s, r) : 0 < s < 1, 0 < r < \rho^*(s)\}$:

Ω_a. $0 < s \le 1/4$ and $0 < r < \rho^*(s)$, or $1/4 < s < 1/3$ and $4s - 1 < r < \rho^*(s)$.

Ω_b. $1/4 < s < 1/3$ and $r = 4s - 1$.

Ω_c. $1/3 < s < 1$ and $0 < r < \rho^*(s)$, or $1/4 < s \le 1/3$ and $r < 4s - 1$,

the following theorem is proved in the Appendix:

Theorem 4.1. *For $\mu_n = \mu_{n,s} = \sqrt{2s \log n}$, and*

$$m = \begin{cases} n^r, & (s, r) \in \Omega_a \cup \Omega_b, \\ \sqrt{2\pi} \cdot \mu_n \cdot n^r, & (s, r) \in \Omega_c, \end{cases}$$

let $\ell_n = \ell_{n,\tau} = n^{\tau/2}$, *where*

$$\tau = \tau(s,r) = \begin{cases} 1-2s-r, & (s,r) \in \Omega_a \cup \Omega_b, \\ 2(1+s-2\sqrt{s(1+r)}), & (s,r) \in \Omega_c, \end{cases}$$

then under H_0 *as well under* $H_1^{(n,m)}$,

$$\ell_n \cdot LR_{n,m} \stackrel{w}{\Longrightarrow} \begin{cases} N(0,1), & (s,r) \in \Omega_a, \\ N(0,1/2), & (s,r) \in \Omega_b, \\ \frac{1}{\sqrt{2\pi}}\tilde{\nu}^0_{s,\tau}, & (s,r) \in \Omega_c, \end{cases}$$

where $\tilde{\nu}^0_{s,\tau}$ *is the distribution with characteristic function* $e^{\tilde{\psi}^0_{s,\tau}}$, *and* $\tilde{\psi}^0_{s,\tau}(t) = \int_{-\infty}^{\infty}(e^{ite^z} - 1 - ite^z)e^{-\frac{1+s-\tau/2}{2s}z}dz$.

Adapting to our notations, Burnashev and Begmatov [8] has studied the limiting behavior of $LR_{n,m}$, with $m = 1$.

We remark here that in Theorem 4.1, the log term in the calibration of m is chosen for convenience. A similar result will be true if we take $m = n^r$ without any log term, and at the same time adding some log term to ℓ_n.

We now finish the proof of Theorem 1.2 in this case. To do so, we first check that $\tilde{\nu}^0_{s,\tau}$ indeed has a bounded continuous density function. In fact, by substitution $x = te^z$, we can rewrite:

$$\tilde{\psi}^0_{s,\tau}(t) = -|t|^{(1+s-\tau/2)/(2s)} \cdot e^{\pm i\pi\cdot\xi/2}, \tag{4.4}$$

where in $\pm$ the upper sign prevails for $t > 0$, and ξ is a complex number determined by:

$$e^{i\pi\cdot\xi/2} = -\int \left[e^{ix} - 1 - ix\right] \cdot |x|^{-(1+3s-\tau/2)/(2s)}\, dx;$$

with τ defined above and $(s,r) \in \Omega_c$, by elementary analysis, $1 < (1+s-\tau/2)/(2s) < 2$, and that $\tilde{\nu}^0_{s,\tau}$ has a bounded density function.

Now let $F_{s,r}$ be the cdf of $N(0,1)$, $N(0,1/2)$, and $\tilde{\nu}^0_{s,\tau}$ according to $(s,r) \in \Omega_a$, Ω_b, and Ω_c, notice that $F_{s,r}$ is a continuous function; now for any fixed (r,s) in the undetectable region, combining (4.3) with Theorem 4.1 gives:

$$\lim_{n\to\infty} \left\|F_0^{(n,m)} - F_1^{(n,m)}\right\|_{KS} \le \lim_{n\to\infty} \left[\left\|F_0^{(n,m)} - F_{s,r}\right\|_{KS} + \left\|F_1^{(n,m)} - F_{s,r}\right\|_{KS}\right] = 0; \tag{4.5}$$

it then follows that, for any sequence of thresholds $\{t_n\}_{n=1}^{\infty}$, the thresholding procedure that reject H_0 when $LR_{n,m} \ge t_n$ has an asymptotically equal to 1 of sum of Type I and Type II errors, uniformly for all sequences $\{t_n\}_{n=1}^{\infty}$:

$$\lim_{n\to\infty} \left[P_{H_0}\{LR_{n,m} \ge t_n\} + P_{H_1^{n,m}}\{LR_{n,m} < t_n\}\right] = 1.$$

Last, since for fixed r,s, and n, among all tests, the Neyman–Pearson likelihood ratio test with a specific threshold has the smallest sum of Type I and Type II errors, see, for example, [28], it then follows that the sum of Type I and Type II errors for any test tends 1. This concludes the proof of Theorem 1.2 in this case. □

Remark. We now give a short remark about the distribution of $\tilde{\nu}^0_{s,\tau}$. First, it was pointed out in [15] that, for a characteristic function e^{ψ} with ψ in the form as that in (4.4), its corresponding distribution has a finite pth moment if and only if $p < (1+s-\tau)/(2s)$; thus $\tilde{\nu}^0_{s,\tau}$ has a finite first moment, but not a finite second or higher moment. Second, it would be interesting to study whether (or when) $\tilde{\nu}^0_{s,\tau}$ is a stable law; $\tilde{\nu}^0_{s,\tau}$ is a stable law if and only if that in (4.4), $|\xi| \leq 2-(1+s-\tau)/(2s)$, see, for example, [15]; we skip further discussion.

5. Proof of Theorem 1.4

To prove Theorem 1.2, we note that it is sufficient to show

$$\lim_{n\to\infty} P_{H_1^{(n,m)}}\big\{HC_N^* \leq \sqrt{4\log\log N}\,\big\} = 0. \tag{5.1}$$

The key for proving (5.1) is to argue that the distribution of HC_N^* under $H_1^{(n,m)}$ will keep the unchanged if we replace the original sampling procedure by the following simple procedure: draw independently a total of N samples, with the first m from $N(\mu_n, 1)$ and the remaining $N-m$ from $N(0,1)$; we refer the latter as the *simplified* sampling.

In fact, if we use $\mathcal{HC}_N^*$ to denote the Higher Criticism statistic based such samples obtained by simplified sampling. Compare $\mathcal{HC}_N^*$ with HC_N^*, for any set of integers $1 \leq j_1, j_2, \ldots, j_m \leq n$, let $E_{\{j_1,j_2,\ldots,j_m\}}$ be the event:

$$E_{\{j_1,j_2,\ldots,j_m\}} = \big\{j_0(1) = j_1, j_0(2) = j_2, \ldots, j_0(m) = j_m\big\};$$

by symmetry, conditional on $E_{\{j_1,j_2,\ldots,j_m\}}$, HC_N* equals to $\mathcal{HC}_N^*$ in distribution:

$$\big[HC_N^* | E_{\{j_1,j_2,\ldots,j_m\}}\big] =_D \mathcal{HC}_N^*,$$

we thus conclude:

$$HC_N^* =_D \mathcal{HC}_N^*.$$

By the above analysis, it is clear that to show (5.1), it is sufficient to show:

$$\lim_{n\to\infty} P\big\{\mathcal{HC}_N^* \leq \sqrt{4\log\log N}\,\big\} = 0; \tag{5.2}$$

where the probability is evaluated for samples obtained by the simplified sampling. The proof of (5.2) is similar to the proof of Theorem 1.2 in [11], and we skip the technical detail. □

6. Extension

In this section, we extend our studies to certain non-Gaussian settings, or the Generalized-Gaussian settings. The Generalized Gaussian (Subbotin) distribution $\mathrm{GN}_\gamma(\mu)$ has density function $\phi_\gamma(x-\mu)$ where

$$\phi_\gamma(x) = \frac{1}{C_\gamma}\exp\left(-\frac{|x|^\gamma}{\gamma}\right), \qquad C_\gamma = 2\Gamma\left(\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}-1}. \tag{6.1}$$

This class of distributions was introduced by M. T. Subbotin 1923 ([31]) and has been discussed in [27, p. 195]. The Gaussian is one member of this family: namely, the one with $\gamma = 2$. The case $\gamma = 1$ corresponds to the Double Exponential (Laplace) distribution, which is a well-understood and widely-used distribution. The case $\gamma < 1$ is of interest in image analysis of natural scenes, where it has

been found that wavelet coefficients at a single scale can be modelled as following a Subbotin distribution with $\gamma \approx 0.7$. This suggests that various problems of image detection, such as in watermarking and steganography, could reasonably use the model above. A direct extension of the Gaussian mixture model (2.2)–(2.3) is the following:

$$H_0 : \bar{X}_j^{(k)} \overset{i.i.d}{\sim} GN_\gamma(0), \qquad 1 \le j \le n,\ 1 \le k \le m \tag{6.2}$$

$$\mathcal{H}_1^{(n,m)} : \bar{X}_j^{(k)} \overset{i.i.d}{\sim} (1-1/n)GN_\gamma(0) + (1/n)GN_\gamma(\mu), \quad 1 \le j \le n,\ 1 \le k \le m, \tag{6.3}$$

where we choose the calibrations in a similar way to that in the Gaussian setting:

$$\mu = \mu_{n,\gamma,s} = \big(\gamma s \log(n)\big)^{1/\gamma}, \qquad m = n^r, \qquad 0 < s < 1,\ 0 < r < 1. \tag{6.4}$$

Similar to the Gaussian case, for r and s in this range, this is again a very subtle problem. Recall that we mentioned in Section 1, the Gaussian Mixture model provides an important bridge for studying the (Gaussian) multiple-looks model, and which is also easier to study. For this reason, in this section, we will focus on the extension of Gaussian mixture model only. It would be interesting to work on a direct extension of Model (1.4)–(1.5), or non-Gaussian multiple-looks model; heuristically, based on Theorems 6.1 and 6.2 below, parallel results for Theorems 1.2 and 1.4 should still hold if we replace the Gaussian noise setting by the Generalized-Gaussian noise setting.

In this section, we will drop the subscript γ whenever there is no confusion.

6.1. Log-likelihood ratio and limit law

In this section, parallely to the Gaussian case, we discuss the limit law of the log-likelihood ratio statistic. Let $g(z|\mu) = g(z|\mu,\gamma) \equiv e^{(|z|^\gamma - |z-\mu|^\gamma)/\gamma}$, then the log-likelihood ratio of testing Model (6.2)–(6.3) is $\mathcal{LR}_{n,m} = \mathcal{LR}_{n,m,s,\gamma} = \sum_{k=1}^m \sum_{j=1}^n \mathcal{LR}_j^{(k)}$, where

$$\mathcal{LR}_j^{(k)} = \mathcal{LR}_{j,s,\gamma}^{(k)} = \log\big(1 - 1/n + (1/n)g\big(\bar{X}_j^{(k)}|\mu,\gamma\big)\big); \tag{6.5}$$

We now discuss the cases $\gamma > 1$ and $0 < \gamma \le 1$ separately.

First for the case $\gamma > 1$. This case includes the Gaussian ($\gamma = 2$) as a special case. Adapting to the notations in [26, Chapter 3], let

$$\begin{aligned} s_0(\gamma) &= \big(2^{\frac{1}{\gamma-1}} - 1\big)^\gamma / \big(2^{\frac{\gamma}{\gamma-1}} - 1\big), \\ a_1(\gamma) &= \big[1 - (1/2)^{1/(\gamma-1)}\big]^{1-\gamma}, \\ b_1(\gamma) &= \big[1 - 2^{\frac{1}{\gamma-1}}\big] / \big[(1 - 2^{\frac{1}{1-\gamma}}\big]^{\frac{1}{\gamma-2}}, \end{aligned}$$

and $x_s = x_s(\gamma)$ be the unique solution of the equation

$$x^\gamma - (x-1)^\gamma = \frac{1}{s}, \qquad x > 1;$$

notice here $\gamma = 2$ corresponds to the Gaussian case: $a_1(2) = 1$, $b_1(2) = -1$, $s_0(2) = 1/3$, and $x_s(2) = (1+s)/(2s)$, which are the same as we derived before. The main result for the case $\gamma > 1$ is the following theorem:

Theorem 6.1. *For parameter $0 < s < 1$, let $\mu_n = \mu_{n,s,\gamma} \equiv (\gamma \cdot s \cdot \log n)^{1/\gamma}$,*

$$m^* = m^*(n,s,\gamma) \equiv \begin{cases} (1/C_\gamma) \cdot \big[2\pi/\big((1-\gamma)b_1(\gamma)\big)\big]^{1/2} \cdot \mu_n^{1-\gamma/2} \cdot n^{1-a_1(\gamma)\cdot s}, & 0 \le s \le s_0(\gamma), \\ C_\gamma \cdot \mu_n^{\gamma-1} \cdot n^{s\cdot(x_s(\gamma))^\gamma}, & s_0(\gamma) < s < 1, \end{cases}$$

and $\mathcal{LR}_{n,m^*} \equiv \mathcal{LR}_{n,m^*,s,\gamma}$, then as $n \to \infty$:

1. *When $0 < s < s_0(\gamma)$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N\Big(-\frac{1}{2}, 1\Big), \qquad \text{under } H_0,$$
$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N\Big(\frac{1}{2}, 1\Big), \qquad \text{under } \mathcal{H}_1^{n,m^*}.$$

2. *When $s = s_0(\gamma)$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(-1/4, 1/2), \qquad \text{under } H_0,$$
$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N(1/4, 1/2), \qquad \text{under } \mathcal{H}_1^{n,m^*}.$$

3. *When $s_0(\gamma) < s < 1$,*

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} \nu^0_{s,\gamma}, \qquad \text{under } H_0, \qquad \mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} \nu^1_{s,\gamma}, \qquad \text{under } \mathcal{H}_1^{(n,m^*)}.$$

where $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$ are distributions with characteristic functions $e^{\psi^0_{s,\gamma}}$ and $e^{\psi^1_{s,\gamma}}$ respectively, and with $w_{s,\gamma} = x_s(\gamma)/[\frac{1}{s\cdot(x_s(\gamma)-1)^{\gamma-1}} - 1]$,

$$\psi^0_{s,\gamma}(t) = \int_{-\infty}^{\infty} \big[e^{it\log(1+e^z)} - 1 - ite^z\big] e^{-[1+w_{s,\gamma}]\cdot z}\, dz, \tag{6.6}$$

$$\psi^1_{s,\gamma}(t) = \psi^0_{s,\gamma}(t) + \int_{-\infty}^{\infty} \big[e^{it\log(1+e^z)} - 1\big] e^{-w_{s,\gamma}\cdot z}\, dz. \tag{6.7}$$

In Section 6.3, we will discuss several issues about the laws $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$; it was validated in [26, Chapter 2] that both $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$ are in fact infinitely divisible.

We now discuss the case $0 < \gamma \le 1$, this case include Laplace ($\gamma = 1$) as a special case; the main result for this case is the following theorem:

Theorem 6.2. *For $0 < \gamma \le 1$ and $0 < s < 1$, let*

$$\mu_n = \mu_{n,s,\gamma} \equiv (\gamma s \log n)^{\frac{1}{\gamma}}, \qquad m^* = m^*(n,s,\gamma) \equiv \begin{cases} 2^{1/\gamma} \cdot n^{1-s}, & \gamma < 1, \\ (3/2) \cdot n^{1-s}, & \gamma = 1, \end{cases} \tag{6.8}$$

and $\mathcal{LR}_{n,m^*} \equiv \mathcal{LR}_{n,m^*,s,\gamma}$, then as $n \to \infty$:

$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N\Big(-\frac{1}{2}, 1\Big), \qquad \text{under } H_0,$$
$$\mathcal{LR}_{n,m^*} \stackrel{w}{\Longrightarrow} N\Big(\frac{1}{2}, 1\Big), \qquad \text{under } \mathcal{H}_1^{n,m^*}.$$

Theorems 6.1 and 6.2 are proved in [26, Chapter 3]. As $\gamma = 2$ corresponds to the Gaussian case, the study in Section 2 is a special case of Theorem 6.1; however, in comparison, technically we need much more subtle analysis to prove Theorem 6.1 than Theorem 2.1.

In this paper, we skip the proof for Theorem 6.1 and Theorem 6.2.

6.2. Detection boundary

Similar to the Gaussian case, Theorem 6.1 implies that there is a *threshold effect* for the detection problem of (6.2)–(6.3). Dropping some lower order term when necessary, m^* would be reduced into a clean form: $m^* = n^{\rho^*_\gamma(s)}$, where

$$\rho^*_\gamma(s) = 1 - s, \qquad 0 < \gamma \le 1,$$

$$\rho^*_\gamma(s) = \begin{cases} 1 - a_1(\gamma)\cdot s, & 0 < s \le s_0(\gamma), \\ s \cdot x_s^\gamma(\gamma), & s_0(\gamma) < s < 1, \end{cases} \qquad \gamma > 1.$$

Similarly, in the s-r plane, the curve $r = \rho^*_\gamma(s)$ separates the square $\{(s,r) : 0 < s < 1,\ 0 < r < 1\}$ into two regions: a detectable region above the curve, and an undetectable region below the curve; we called $r = \rho^*_\gamma(s)$ the *detection boundary*.

Theorem 6.3. *For $\gamma > 0$, let $\mu_n = \mu_{n,s,\gamma} = (\gamma \cdot s\log(n))^{1/\gamma}$, $m = n^r$, and $\mathcal{LR}_{n,m} \equiv \mathcal{LR}_{n,m,s,\gamma}$.*

1. *When $r > \rho^*_\gamma(s)$, consider the likelihood ratio test (LRT) that rejects H_0 when $\mathcal{LR}_{n,m} > 0$, then the sum of Type I and Type II errors tends to 0:*

$$P_{H_0}\{\textit{Reject } H_0\} + P_{\mathcal{H}_1^{(n,m)}}\{\textit{Accept } H_0\} \to 0, \qquad n \to \infty.$$

2. *When $r < \rho^*_\gamma(s)$,*

$$\lim_{n\to\infty} \left\| F_0^{(n,m)} - F_1^{(n,m)} \right\|_{KS} = 0,$$

where $F_0^{(n,m)}$ and $F_1^{(n,m)}$ are the cdf's of $LR_{n,m}$ under H_0 and $\mathcal{H}_1^{(n,m)}$ respectively. As a result, the sum of Type I and Type II errors for any test tends to 1:

$$P_{H_0}\{\textit{Reject } H_0\} + P_{\mathcal{H}_1^{(n,m)}}\{\textit{Accept } H_0\} \to 1, \qquad n \to \infty.$$

The proof of Theorem 6.3 is similar to that of Theorem 1.2, and we skip it.

In [11], we have studied in detail the performance of Higher Criticism statistic for Model (6.2)–(6.3), and showed the Higher Criticism is also optimal adaptive for Model (6.2)–(6.3) with any fixed $\gamma > 0$. It is interesting to notice that for any fixed $\gamma > 1$, the detection boundary is a partly linear ($0 < s < s_0(\gamma)$) and partly curved ($s_0(\gamma) < s < 1$). Again, this implies that the detection problem is essentially different for those parameters (s,r) near the linear part and those near the curved part. Asymptotically, when (s,r) is close to the curved part, statistics based on those a few largest observations would be able to effectively detect, while when (s,r) is close to the linear part, statistics based on a few largest observations will completely fail, and only the newly proposed statistic Higher Criticism, or the Berk–Jones statistic, which is asymptotically equivalent to the Higher Criticism in some sense [5, 11], is still able to efficiently detect. See [11] for more discussion.

Moreover, notice that when $\gamma > 1$ approaches 1, the curved part of the detection boundary continues to shrink and eventually vanishes, leaves only the linear part. So when $0 < \gamma \le 1$, statistics based on the largest a few observations would completely fail for all $0 < s < 1$. However, Higher Criticism and Berk-Jones statistics would still be efficient.

In Figure 5, we plot $r = \rho^*_\gamma(s)$ for $\gamma = 3$, 2, 1.5, and $\gamma \le 1$. Notice that $\gamma = 2$ corresponds to the Gaussian case and $\rho^*_2 \equiv \rho^*$.

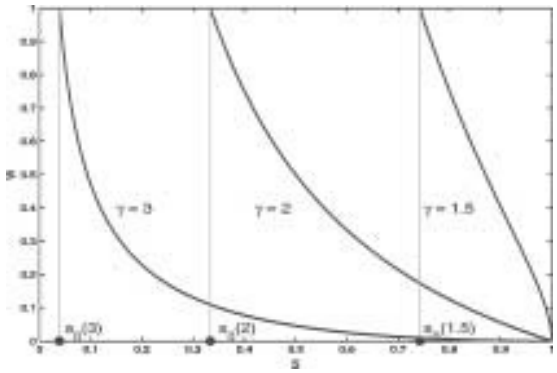


Figure 2: Illustration of $w_{s,\gamma}$ as a function of s for fixed γ. From left to right, three curves correspond to $w_{s,\gamma}$ over intervals $[s_0(\gamma), 1]$ for $\gamma = 3, 2$ and 1.5.

6.3. Remarks on the infinitely divisible laws

In this section, we addressed several issues about the infinitely divisible laws $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$.

The distribution of $\nu^0_{s,\gamma}$ or $\nu^1_{s,\gamma}$ is uniquely determined by the value of $w_{s,\gamma}$. By elementary analysis, for fixed γ, when s ranges between $s_0(\gamma)$ and 1, $w_{s,\gamma}$ strictly decreases from 1 to 0. In Figure 2, we graph $w_{s,\gamma}$ as a function of s with $\gamma = 1.5, 2, 3$. Notice that $\gamma = 2$ corresponds to the Gaussian case, and

$$w_{s,2} = (1-s)/(2s).$$

As $0 < w_{s,\gamma} < 1$, it is easy to check that $e^{\psi^0_{s,\gamma}}$ and $e^{\psi^1_{s,\gamma}}$ are absolute integrable; thus by the inversion formula ([12] for example), both $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$ have a bounded continuous density function. In Figure 3, we graph the density functions for $\nu^0_{s,\gamma}$ or $\nu^1_{s,\gamma}$, with $w_{s,\gamma} = 0.4, 0.5, 0.6$ separately; recall that the density function is uniquely determined by $w_{s,\gamma}$. Figure 3 suggests that, heuristically, the smaller the $w_{s,\gamma}$, the better separation between $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$, it would be interesting to validate this, but we skip further discussion. Notice here that the density functions correspond to $w_{s,\gamma} = 0.5$ are the same as those in Figure 1, where $w_{s,\gamma} = 0.5$ since we take $s = 1/2, \gamma = 2$.

Last, we claim that $\nu^0_{s,\gamma}$ has a finite first moment as well as a finite second moment, and so does $\nu^1_{s,\gamma}$. In fact, elementary analysis shows that the second derivatives of both $e^{\psi^0_{s,\gamma}}$ and $e^{\psi^1_{s,\gamma}}$ exist, so the claim follows directly from the well-known theorem, that the existence of the second derivatives of characteristic functions implies the existence of the second moments, see ([12, p. 104]). Moreover, the first moment of $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$ are:

$$\int \left[\log\left(1+e^z\right) - e^z\right] e^{-(1+w_{s,\gamma})z}\, dz, \qquad \int \left[\left(1+e^z\right)\cdot \log\left(1+e^z\right) - e^z\right] e^{-(1+w_{s,\gamma})z}\, dz,$$

and are negative and positive respectively; the second moment of them are:

$$\int \left[\log^2\left(1+e^z\right)\right] e^{-(1+w_{s,\gamma})z}\, dz, \qquad \int \left[\left(1+e^z\right)\cdot \log^2\left(1+e^z\right)\right] e^{-(1+w_{s,\gamma})z}\, dz.$$

It would be interesting to study that, whether higher order moments exist for $\nu^0_{s,\gamma}$ or $\nu^1_{s,\gamma}$. Here we skip further discussion.

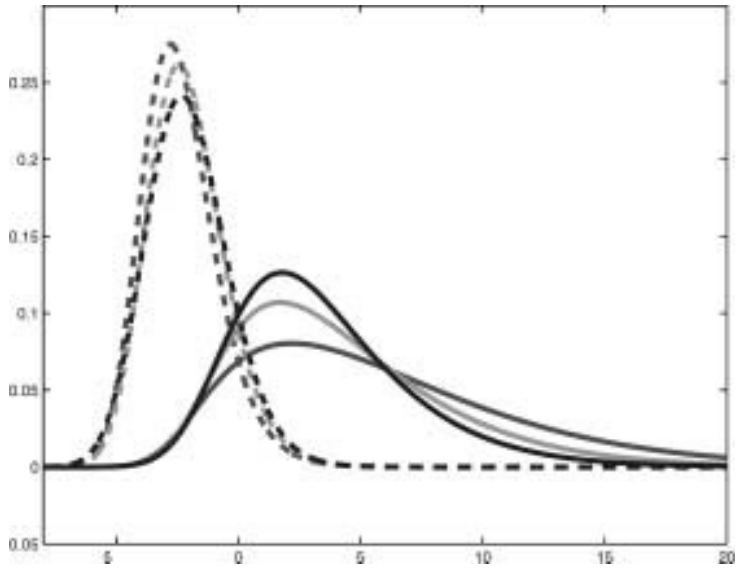


Figure 3: Density functions for $\nu^0_{\gamma,s}$ and $\nu^1_{s,\gamma}$. The distributions of $\nu^0_{s,\gamma}$ and $\nu^1_{s,\gamma}$ only depends on $w_{s,\gamma}$. Left: from top to bottom, density functions for $\nu^0_{s,\gamma}$ with $w_{s,\gamma} = 0.4,\ 0.5,\ 0.6$. Right: from bottom to top, density functions for $\nu^1_{s,\gamma}$ with $w_{s,\gamma} = 0.4,\ 0.5,\ 0.6$.

7. Discussions

7.1. Re-parametrization and detection boundary

In Section 6, we calibrated the amplitude of the signal μ and the number of frames m through parameters s and r by:

$$\mu_{n,s,\gamma} = (\gamma \cdot s \cdot \log n)^{1/\gamma}, \qquad m = n^r, \qquad 0 < s < 1,\ 0 < r < 1.$$

This particular calibration is very convenient for discussing the limit law of the log-likelihood ratio: in order to make the log-likelihood ratio converge to non-degenerate distribution, the critical value of $m = m^*$ may contain a log term, namely in the case $s > s_0(\gamma)$. When we attempt to develop a different (but equivalent) calibration, this log term may complicate the notation system quite a bit. However, the above calibration is not convenient for the discussion of the detection boundary. Recall that the detection boundary for the Generalized-Gaussian Mixture model (6.2)–(6.2) in the s–r plane is $r = \rho^*_\gamma(s)$, where:

$$\rho^*_\gamma(s) = 1 - s, \qquad 0 < \gamma \le 1,$$

$$\rho^*_\gamma(s) = \begin{cases} 1 - a_1(\gamma) \cdot s, & 0 < s \le s_0(\gamma), \\ s \cdot x^\gamma_s(\gamma), & s_0(\gamma) < s < 1, \end{cases} \qquad \gamma > 1;$$

unfortunately, here $x_s(\gamma)$ is the solution of $x^\gamma - x^{\gamma-1} = 1/s$, which doesn't have an explicit formula. In addition to providing a completely explicit formula for the detection boundary, the following calibration we will introduce might also be more familiar. As before, let $N = n \cdot m$ be the total number of observations, and ϵ_N denote

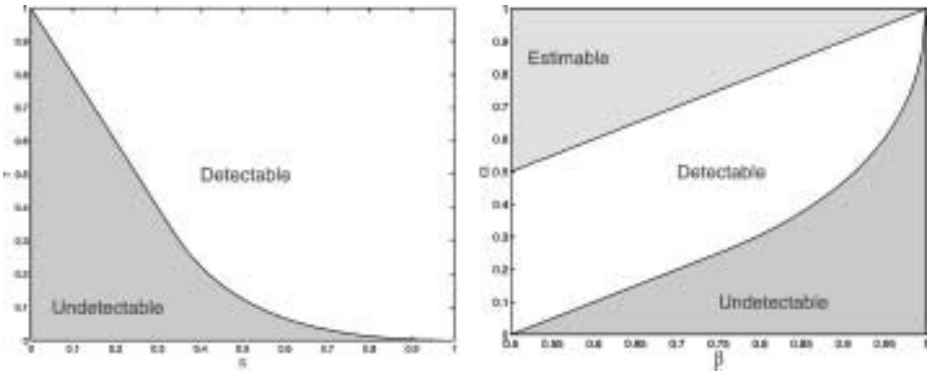


Figure 4: Left Panel: detection regions for the Model (1.4)–(1.5) as well as Gaussian mixture model (2.2)–(2.3), the detection boundary separates the detectable region (above) from the undetectable region (bottom). Right panel: detection regions in the $\beta - \alpha$ plane by the re-parametrization in Section 7.1. The detection boundary separates the detectable region from the undetectable region. The mapping of the re-parametrization maps the line segment $\{(s, r) : s = 1, 0 < r < 1\}$ in the left panel to the line segment $\{\alpha = \beta : 1/2 < \beta < 1\}$, which separates the estimable region (top) from the non-estimable region. When (α, β) falls into the estimable region, it is possible not only to detect the presence of nonzero means, but also to estimate those means.

the fraction of observations containing a signal, so $m = N \cdot \epsilon_N$, and $n = 1/\epsilon_N$; we now introduce parameters (β, α) and let:

$$\epsilon_N = N^{-\beta}, \qquad \mu_N = \mu_{N,\alpha} = (\gamma\alpha \log n)^{1/\gamma};$$

this re-parametrization is equivalent to a simple transformation:

$$\beta = 1/(1+r), \qquad \alpha = s/(1+r), \qquad 1/2 < \beta < 1, \qquad 0 < \alpha < 1; \tag{7.1}$$

elementary algebra enables us to rewrite the detection boundary $r = \rho^*_\gamma(s)$ as:

$$\alpha = \bar{\rho}^*_\gamma(\beta) \equiv \begin{cases} \left[2^{1/(\gamma-1)} - 1\right]^{\gamma-1} \cdot (\beta - 1/2), & 1/2 < \beta \le 1 - 2^{-\gamma/(\gamma-1)}, \\ \left(1 - (1-\beta)^{1/\gamma}\right)^{\gamma}, & 1 - 2^{-\gamma/(\gamma-1)} < \beta < 1. \end{cases}$$

Figure 4 can help to understand the re-parametrization. In fact, the above transform is a one-to-one mapping, which maps the squared region in the s - r plane $\{(s, r) : 0 < s < 1,\ 0 < r < 1\}$ (left panel) to the region in the β-α plane which formed by cutting the triangular region on the top off the square $\{(\beta, \alpha) : 0 < \alpha < 1,\ 0 < \beta < 1\}$ (right panel). Moreover, the new sub-regions above/below the curve $\alpha = \bar{\rho}^*_\gamma(\beta)$ is the image of the detectable/undetectable regions. See Figure 4 for more illustration. For Model (1.4)–(1.5), a problem closely related to the detection problem we have discussed in this paper is the estimation problem: with the same calibration, what is the critical value of m such that the signals can be reliably *estimated*? Surprisingly, though multiple-looks is helpful for the detection, it is not at all helpful for estimation; and in order that the signal be estimable, we have to set the parameter $s > 1$, or $\mu \ge \sqrt{2 \log n}$; this range of s is not showed in the left panel of Figure 4. But by (7.1), $s > 1 \Longleftrightarrow \alpha > \beta$, so in other

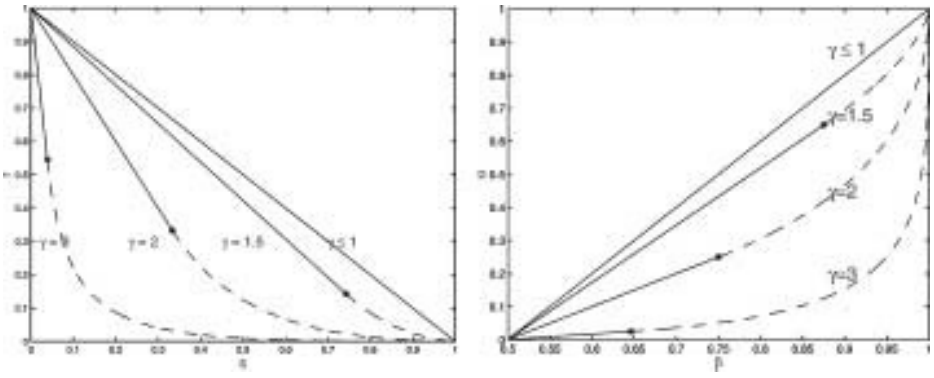


Figure 5: Left panel: Detection boundaries in the s-r plane for Model (6.2)–(6.3), with $\gamma \leq 1$, and $\gamma = 1.5,\ 2,\ 3$ from top to bottom. A small dot separates each curve into two parts, the solid part of the curves are line segments. Right panel: The same detection Boundaries in the β-α plane after the re-parametrization defined in (7.1).

words, in order that the signal be estimable, we need to pick (α, β) from the triangular region on the top of the right panel in Figure 4; we call this triangular region the *estimable* region. A similar problem was discussed in [1], with Model (2.2)–(2.3) instead of Model (1.4)–(1.5).

7.2. Discussions on Model (1.4)–(1.5)

We now address several issues about the multiple-looks model, Model (1.4)–(1.5).

First, in astronomy, there is a Poisson version of the multiple-looks model. As it is of interests to study directly the Poisson model rather than the Gaussian model in this paper, the Gaussian model is more convenient to study, and reveals insights about the Poisson model.

Second, in Model (1.4)–(1.5), we have assumed that each $X_j^{(k)}$ has equal variance either it contains a signal or not. It is interesting to consider a more general case, in which we assume that, the pixels containing signals have equal variances $\sigma^2 > 1$, while all other pixels have equal variance 1. Our study in this paper can be generalized to this case easily, and the parameter σ should have some scaling effect on the detection boundary $r = \rho^*(s)$.

Last, it is interesting to study what happens if we relax some assumptions of Model (1.4)–(1.5). For example, instead of assuming that exactly one pixel per frame possibly contains a signal, we could consider a harder problem that, in each frame, there is more than one pixel possibly containing a signal with equal mean, while the position of such pixels are (independently or not) sampled from $\{1, 2, \ldots, n\}$, but independently from frame to frame. Heuristically, if the number of those pixels containing a signal are relatively small, we should be able to show that, this model also can be converted approximately into a Gaussian mixture model by random shuffling; notice that the study of the resulting Gaussian mixture model should be similar to that in Section 2.

7.3. Relation to other work

There are two points of contact with earlier literature. The first one is with Burnashev and Begmatov [8], who studied the limit law of log-likelihood ratio with a

setting which can be translated into ours with large n but $m = 1$. They showed that, for n iid sample $z_i \sim N(0,1)$, with approximate normalization, $\mathrm{Ave}_j\{e^{\mu_n z_j - \mu_n^2/2}\}$ weakly converges to a stable distribution as $n \to \infty$. It is interesting to notice here that, the non-Gaussian weak limits in Theorems 2.1 and 6.1 are infinitely divisible, but not stable. It would be interesting to study whether the non-Gaussian limit in Theorem 4.1 is stable or not.

The second point of contact is with the beautiful series of papers by Ingster [19, 20], and [21]. Ingster studied extensively the Gaussian mixture model (2.2)–(2.3), ranging from the limit law of the log-likelihood ratio as well as the minimax estimation of signals lying in an ℓ_n^p ball. These papers revealed the same limiting behavior of log-likelihood ratio (and so the threshold effect) as discussed in Section 2. Our approach in Section 2 was developed independently.

In this paper, our starting point was the multiple-looks model (1.4)–(1.5), which is different than the model studied by Ingster. We found that we could treat the multiple-looks model by proving that, after a re-expression of the problem, we obtained convergence in variation norm to the Gaussian mixture model (2.2)–(2.3), which we then analyzed. Hence, although we obtained eventually the same results as Ingster, our application and motivation were different. We think the alternative viewpoint adds something to the discussion. Moreover, the extension to the studies on generalized-Gaussian mixtures in Section 6 has not been studied before, and various effects of the parameter γ are interesting.

8. Appendix

In this section, we will prove Theorem 4.1. Consider the following three sub-regions of the square $\{(s,\tau\} : 0 < s < 1, 0 < \tau < 1\}$.

ω_a**:** $0 < s \le 1/4$ and $0 < \tau < \rho^*(s)$, or $1/4 < s \le 1/3$ and $0 < \tau < 2 - 6s$,

ω_b**:** $1/4 < s < 1/3$ and $\tau = 2 - 6s$,

ω_c**:** $1/3 < s < 1$ and $0 < \tau < 2(1 - \sqrt{s})^2$, or $1/4 < s < 1/3$ and $\tau > 2 - 6s$;

recall $\mathcal{LR}_j^{(k)} = \log(1 - (1/n) + (1/n) \cdot e^{\mu_n \bar{X}_j^{(k)} - \mu_n^2/2})$, we have the following lemma:

Lemma 8.1. *If $\mu_n = \mu_{n,s} = \sqrt{2s \log n}$, $\ell_n = \ell_{n,\tau} = n^{\tau/2}$, and with $\tau = \tau(s,r)$ defined in Theorem* 4.1*, then when $n \to \infty$,*

$$E_0\big[e^{it \cdot \ell_n \cdot \mathcal{LR}_j^{(k)}}\big] = \begin{cases} 1 - n^{-(2-2s)+\tau} \cdot \dfrac{(t^2 + o(1))}{2}, & (s,\tau) \in \omega_a, \\ 1 - n^{-(2-2s)+\tau} \dfrac{(t^2 + o(1))}{4}, & (s,\tau) \in \omega_b, \\ 1 + \dfrac{1}{\mu_n \cdot \sqrt{2\pi}} \cdot n^{[\frac{1-\tau/4}{4s}\tau - \frac{(1+s)^2}{4s}] + \tau/4} \cdot \big(\tilde{\psi}_{s,\tau}^0(t) + o(1)\big), & (s,\tau) \in \omega_c, \end{cases}$$

and

$$E_1\big[e^{it \cdot \ell_n \cdot \mathcal{LR}_j^{(k)}}\big] = \begin{cases} 1 - n^{-(1-2s)+\tau/2} \cdot \dfrac{(t^2 + o(1))}{2}, & (s,\tau) \in \omega_a, \\ 1 - n^{-(1-2s)+\tau/2} \cdot \dfrac{(t^2 + o(1))}{4}, & (s,\tau) \in \omega_b, \\ 1 + \dfrac{1}{\mu_n \cdot \sqrt{2\pi}} n^{[\frac{1-\tau/4}{4s}\tau - \frac{(1-s)^2}{4s}] - \tau/4} \cdot \big(\tilde{\psi}_{s,\tau}^*(t) + o(1)\big), & (s,\tau) \in \omega_c, \end{cases}$$

with E_0 and E_1 denote the expectation with respect to the law of $z \sim N(0,1)$ and $z \sim N(\mu_n, 1)$ respectively; here $\tilde{\psi}^0_{s,\tau}(t)$ is defined in Theorem 4.1*, and $\tilde{\psi}^*_{s,\tau}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (e^{ite^z} - 1) e^{-\frac{1-s-\tau/2}{2s} z}\, dz$.*

Proof. As the proof for two equations are similar, we only prove the first one. Similar to that in Section 2.1, namely (2.9)–(2.14):

$$\begin{aligned} &E_0\big[e^{it\cdot \ell_n \cdot \mathcal{LR}_j^{(k)}}\big] \\ &\quad = 1 + \frac{1}{\mu_n} e^{-\frac{(1+s)^2}{8s^2}\mu_n^2} \int \big[e^{it\cdot \ell_n \cdot \log(1+e^z)} - 1 - it \cdot \ell_n \cdot e^z\big] \phi(z/\mu_n)\, dz \\ &\qquad + O\big(\ell_n^2/n^2\big); \end{aligned} \tag{8.1}$$

by substitution $e^{z'} = \ell_n \cdot e^z$, we rewrite

$$\int \big[e^{it\cdot \ell_n \cdot \log(1+e^z)} - it \cdot \ell_n \cdot e^z - 1\big] e^{-\frac{1+s}{2s} z} \phi\left(\frac{z}{\mu_n}\right) dz \tag{8.2}$$

$$= n^{\frac{1+s-\tau/4}{4s}\tau} \int \big[e^{it\cdot \ell_n \cdot \log(1+e^z/\ell_n)} - it \cdot e^z - 1\big] e^{-\frac{1+s-\tau/2}{2s} z} \cdot \phi\left(\frac{z}{\mu_n}\right) dz. \tag{8.3}$$

Observe that $(1+s-\tau/2)/(2s) > 1$ for $(s,\tau) \in \omega_a \cup \omega_b \cup \omega_c$, and moreover, according to (s,τ) in ω_a, ω_b, and ω_c, $(1+s-\tau/2)/(2s) > 2$, $= 2$ and < 2; by similar arguments as in the proof of Lemma 2.1, we derive:

$$\begin{aligned} &\int \big[e^{i\cdot \ell_n \cdot t \log(1+e^z/\ell_n)} - ite^z - 1\big] e^{-\frac{1+s-\tau/2}{2s} z} \phi\left(\frac{z}{\mu_n}\right) dz \\ &\quad = \begin{cases} -\big[\big(t^2 + o(1)\big)/2\big] \cdot \mu_n \cdot n^{-(1-3s-\tau/2)^2/(4s)}, & (s,\tau) \in \omega_a, \\ -\big[\big(t^2 + o(1)\big)/4\big] \cdot \mu_n, & (s,\tau) \in \omega_b, \\ \dfrac{1}{\sqrt{2\pi}}\big(\tilde{\psi}^0_{s,\tau}(t) + o(1)\big), & (s,\tau) \in \omega_c; \end{cases} \end{aligned}$$

inserting this back into (8.3), Lemma 8.1 follows. □

We now proceed to prove Theorem 4.1. With $\tau = \tau(s,r)$ as defined in Theorem 4.1, observe by the calibrations in Theorem 4.1, $(s,\tau) \in \omega_a \Leftrightarrow (s,r) \in \Omega_a$, $(s,\tau) \in \omega_b \Leftrightarrow (s,r) \in \Omega_b$, and $(s,\tau) \in \omega_c \Leftrightarrow (s,r) \in \Omega_c$, so by Lemma 8.1 and elementary analysis,

$$\ell_n \cdot \mathcal{LR}_{n,m} = \sum_{k=1}^{m} \left[\sum_{j=1}^{n} \big(\ell_n \cdot \mathcal{LR}_j^{(k)}\big)\right] \Longrightarrow \begin{cases} N(0,1), & (s,r) \in \Omega_a, \\ N(0,1/2), & (s,r) \in \Omega_b, \\ \tilde{\nu}^0_{s,\tau}, & (s,r) \in \Omega_c, \end{cases}$$

under the H_0 as well as under $\mathcal{H}_1^{(n,m)}$; moreover, with (s,r,τ) in such range, we argue in a similar way as the study in Section 3 that, there is only negligible difference between $\mathcal{LR}_{n,m}$ and $LR_{n,m}$; combining these gives Theorem 4.1. □

Acknowledgments

The author would like to thank David Donoho for guidance and inspiration, as well as extensive discussions, references and encouragement. We also would like to thank Iain Johnstone and Jian Sun for discussions, suggestions, and references. These results were reported in the author's Ph.D. Thesis; the author thanks committee

members for valuable suggestions and encouragement. The author would also like to thank the editor Anirban DasGupta for valuable suggestions, references, and pointers. This work has been partially supported by National Science Foundation grants DMS 00-77261 and DMS 95-05151.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000). Adapting to unkown sparsity by controlling the False Discovery Rate. accepted for publication pending revision, *Ann. Statist.*

[2] Bates, R. H. T. (1984). Optical astronomical speckle imaging, in *Indirect Imaging*. Edited by Roberts, J. A. Cambridge University Press, Cambridge.

[3] Begmatov, I. A. (1988). Asymptotic optimality for the likelihood ratio test in the case of minimax detection of a signal in Gaussian white noise, (Russian) *Izv. Akad. Nauk UzSSR Ser. Fiz.-Mat. Nauk* **3**, 6–11, 88. MR965397

[4] Begmatov, I. A. (1989). Optimality of the likelihood ratio test in the problem of detecting orthogonal signals, (Russian) *Izv. Akad. Nauk UzSSR Ser. Fiz.-Mat. Nauk* **2**, 3–6, 106. MR1023272

[5] Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit Test Statistics that Dominate the Kolmogorov Statistic. *Z. Wahrsch. verw. Geb.* **47**, 47–59. MR521531

[6] Brown, L. D. (1971). Non-local asymptotic optimality of appropriate likelihood ratio tests, *Ann. Statist.* **42**, 1206–1240. MR314167

[7] Bickel, P. J. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, 83–96. J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy, and B. L. S. Prakasa Rao, eds. Wiley Eastern Limited.

[8] Burnashev, M. V. and Begmatov, I. A. (1990). On a problem of detecting a signal that leads to stable distributions, (Russian) *Teor. Veroyatnost. i Primenen. 35* **3**, 557–560; translation in *Theory Probab. Appl.* **35** (1990), no. 3, 556–560 (1991). MR1091213

[9] Chen, H. and Chen, J. (2001). Large sample distribution of the likelihood ratio test for normal mixtures, *Statist. Probab. Lett.* **52**, 125–133. MR1841402

[10] Dembo, A. and Zeitouni, O. (1993). *Large deviation techniques and applications*, Jones and Bartlett Publishers. MR1202429

[11] Donoho, D. and Jin, J. (2004). Higher Criticism for detecting sparse heterogeneous mixtures. *To appear in Ann. Statist.* **32**, no. 3.

[12] Durrett, R. (1996). *Probability: Theory and Examples*, 2nd ed., Duxbury Press. MR1609153

[13] Eisenberg, B. (1971). Non-optimality of likelihood ratio tests for sequential detection of signals in Gaussian noise. *Statistical decision theory and related topics (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1970)*, pp. 377–383. Academic Press, New York. MR339417

[14] Efron, B. (1967). The power of the likelihood ratio test, *Ann. Statist.* **38**, 802–806. MR212935

[15] Feller, W. (1966). *An Introduction to Probability Theory and its Application*, Vol. 2, John Wiley & Sons. MR210154

[16] Garel, B. (2001). Likelihood ratio test for univariate Gaussian mixture, *J. Statist. Plann. Inference* **96**, 325–350. MR1842105

[17] Ghosh, J. K. and Sen, P. K (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results, *Preceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, 789–806, Wadworth Statist./Probab. Ser., Wadsworth, Belmont, CA. MR822065

[18] Hartigan, J. A. (1985). A failure of likelihood ratio asymptotics for normal mixtures, *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* Wadsworth Advanced Books, Monterey, CA and Institute of Mathematical Statistics, Hayward, CA. MR822066

[19] Ingster, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution, *Math. Methods Statist.* **6**, 47–69. MR1456646

[20] Ingster, Y. I. (1999). Minimax detection of a signal for l_n^p-balls, *Math. Methods Statist.* **7**, 401–428. MR1680087

[21] Ingster, Y. I. (2002). Adaptive detection of a signal of growing dimension I, II. *Math. Methods Statist.*, **10** 2001, 395–421, **11**, 37–68.

[22] Ingster, Yu. I. and Lepski, O. (2002). On multichannel signal detection, Preprint No. 764, WIAS-Berlin. MR2036348

[23] Ingster, Y. I. and Suslina, I. A. (2000). Minimax nonparametric hypothesis testing for ellipsoids and Besov bodies, *ESAIM Probab. Statist.*, (electronic) **4**, 53–135. MR1784903

[24] Ingster, Y. I. and Suslina, I. A. (2002). On multichannel detection of a signal of known shape. *Zap. Nauchn. Sem. POMI*, **294**, (to appear in Russian). MR1976749

[25] Jensen, G. J. and Kornberg, R. D. (1998). Single-particle selection and alignment with heavy atom cluster-antibody conjugates, *Proc. Natl. Acad. Sci.* **95**, 9262–9267.

[26] Jin, J. (2003). Detecting and estimating sparse mixtures, Ph.D. Thesis, Stanford University, May, 2003.

[27] Johnson, N. L. and Kotz, S. and Balakrishnan, N. (1987). *Continuous Univariate Distributions*, Vol. 2, 2nd ed., John Wiley & Sons. MR1326603

[28] Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed., John Wiley & Sons. MR852406

[29] Pastor, G. and Gay, R. and Groenenboom, A. (2002). A sharp upper bound for the probability of error of the likelihood ratio test for detecting signals in white Gaussian noise, *IEEE Trans. Inform. Theory* **48**, 228–238. MR1872176

[30] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, John Wiley & Sons. MR838963

[31] Subbotin, M. T. (1923). On the Law of frequency of errors, *Mathematicheskii Sbornik* **31**, 296–301.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 287–290
© Institute of Mathematical Statistics, 2004

r-scan extremal statistics of inhomogeneous Poisson processes

Samuel Karlin[1] **and Chingfer Chen**[1]

Stanford University

Abstract: Studies of inhomogeneities in long DNA sequences can be insightful to the organization of the human genome (or any genome). Questions about the spacings of a marker array and general issues of sequence heterogeneity in our studies of DNA and protein sequences led us to statistical considerations of r-scan lengths, the distances between marker i and marker $i+r$, $i = 1, 2, 3, \ldots$. It is interesting to characterize the r-scan lengths harboring clusters or indicating regions of over-dispersion of the markers along the sequence. Applications are reviewed for certain words in the *Haemophilus* genome and the *Cyanobacter* genome.

1. Introduction

We are happy to contribute this paper to the festscript volume in honor of Dr. H. Rubin. The paper is of pratical and theoretical application. I also had the pleasure to develop with Herman an extended analysis concerning a family of distributions in possession of a monotone likelihood ratio ([1, 2]).

Question about spacings of a marker array and general issues of sequence heterogeneity in our studies of DNA and protein sequences led us to statistical considerations of r-scan lengths, the distances between marker i and marker $i + r$, $i = 1, 2, 3, \ldots$. It is interesting to charaterize the r-scan lengths harboring clusters or indicating regions of over-dispersion of the markers along the sequence. Concretely, a typical objective is to determine the probability of successive $\{r + 1\}$ markers falling within a DNA sequence stretch under an appropriate stochastic model of the marker array. There are similar issues pertaining to sparseness of markers. Particular markers (in the language of DNA, e.g., specific restriction sites, nucleosome placements, locations of genes) are distributed over the genome along chromosomes. The r-scan analysis has been largely applied to the homogeneous Poisson processes for a marker array distributed over a long contig. It is known that the organization of mammalian genomes shows substantial inhomogeneities, including "isochores", regions dominated by either C + G or A + T DNA base content.

Here we consider an inhomogeneous Poisson process Π on the real axis $(0, \infty)$ with an intensity $\lambda(s)$, $0 \leq s < \infty$. The intensity function $\lambda(s)$ can be of different types, for example, periodic or constant in successive intervals, depending on different applications. In this context, we would like to determine the asymptotic distribution of the kth minimum among the r-scan lengths over the interval horizon $(0, t)$, as $t \to \infty$.

2. Preliminaries. Minimal r-scan lengths from a general distribution

In the paper [3], the asymptotic distribution of the kth minimum r-scan length from a general distribution function has been studied by applying the Chen-Stein

[1]Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA. e-mail: karlin@math.stanford.edu

Keywords and phrases: r-scan statistics, inhomogeneous Poisson marker array, asymptotic distributions.

AMS 2000 subject classifications: 92B05, 92D20.

method [4]. In that context, an r-scan process is generated following a piecewise constant function or continuous general density $f(x)$ with bounded support $(0, T]$. Thus let $V_1, V_2, \ldots, V_{n-1}$ be $n-1$ *i.i.d.* samples drawn from the density $f(x)$, and let $V_1^* \le V_2^* \le \cdots \le V_{n-1}^*$ be the order statistics corresponding to $\{V_i\}$. For convenience, let $V_0^* = 0$ and $V_n^* = T$. Then the associated r-scan fragments $R_i = V_{i+r-1}^* - V_{i-1}^*$, $i = 1, \ldots, n-r+1$, and their order statistics R_i^* are defined in the usual way such that $R_1^* \le R_2^* \le \cdots \le R_{n-r+1}^*$. For an extensive review of r-scan statistics, see the book [5].

From $\{R_i\}$, we define the Bernoulli random variables

$$\begin{aligned} U_i^-(a) &= 1, \quad \text{if } R_i \le a \\ &= 0, \quad \text{if } R_i > a \end{aligned}$$

and their sum

$$N_{n-r+1}^-(a) = \sum_{i=1}^{n-r+1} U_i^-(a).$$

Denote by $m_{n,k} = R_k^*$. The asymptotic distribution (as $n \to \infty$) for $m_{n,k}$ is as follows.

Lemma 1. *For a given positive constant μ, let a_n be determined to satisfy*

$$\mu = \frac{(na_n)^r}{r!} n \int [f(x)]^{r+1}\, dx. \tag{1}$$

Then we have the Poisson approximation

$$\lim_{n\to\infty} d\big(N_{n-r+1}^-(a_n), Po(\mu)\big) = 0,$$

for $Po(\mu)$ the Poisson distribution with parameter μ. Here $d(\ ,\ )$ is the total variational distance between two random variables defined by

$$d(U, V) = \sup_A \big[Pr\{U \in A\} - Pr\{V \in A\}\big].$$

Moreover, the kth minimal r-scan length, $m_{n,k}$, possesses the asymptotic distribution

$$\lim_{n\to\infty} \Pr\{m_{n,k} > a_n\} = \sum_{i=0}^{k-1} e^{-\mu} \frac{\mu^i}{i!}. \tag{2}$$

Proof of the above lemma is given in [3], Section 8. Here, by adapting the foregoing result, we will determine the asymptotic distribution of the kth minimal r-scan length corresponding to an inhomogeneous Poisson process in $(0, t)$, as $t \to \infty$.

3. Minimal r-scan lengths for an inhomogeneous Poisson process

The asymptotic theorem for the minimal r-scan length will be derived from the distributional property of $\widetilde{N}_t^-(a)$, where $\widetilde{N}_t^-(a)$ is the number of r-scan segments of lengths $\le a$ over the interval horizon $(0, t)$. It is clear that if $\widetilde{N}_t^-(a) < k$, the kth minimal r-scan length $m_{t,k}$ exceeds the level a. Thus if the Poisson approximation holds for $\widetilde{N}_t^-(a)$, we can access the asymptotic law for $m_{t,k}$. Here the r-scan process of interest is generated from an inhomogeneous Poisson process Π with an intensity function $\lambda(s)$, $0 \le s < \infty$. The main theorem is as follows.

Theorem 1. *Assume $\lambda(s)$ defined for $s \geq 0$ satisfies*

$$\int_0^t \lambda(s)\,ds \to \infty, \quad \text{as } t \to \infty. \tag{3}$$

For a given positive constant μ, let a_t be determined to satisfy the equation

$$\mu = \frac{a_t^r}{r!} \int_0^t \lambda^{r+1}(s)\,ds. \tag{4}$$

Then we have the Poisson approximation

$$\lim_{t\to\infty} d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) = 0.$$

Moreover, the kth minimal r-scan length, $m_{t,k}$, possesses the asymptotic distribution

$$\lim_{t\to\infty} \Pr\{m_{t,k} > a_t\} = \sum_{i=0}^{k-1} e^{-\mu}\frac{\mu^i}{i!}.$$

Proof of Theorem 1. If n_t denotes the point count of the Poisson Process Π in (0, t), then

$$E[n_t] = \int_0^t \lambda(s)ds, \qquad Var(n_t) = \int_0^t \lambda(s)\,ds.$$

For convenience, let $\bar{n}_t = \lfloor E[n_t] \rfloor$. Thus the Berry–Esseen estimate assures

$$\Pr\left\{\left|\frac{n_t}{\bar{n}_t} - 1\right| > \sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}}\right\} = O\left(\frac{1}{\sqrt{\bar{n}_t}}\right).$$

Therefore

$$d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) \leq d\big(N^-_{\bar{n}_t - r+1}(a_t), Po(\mu)\big) + O\left(\sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}}\right).$$

If $n_t = \bar{n}_t$, the $\bar{n}_t$ points in $(0, t)$ are distributed independently according to $g(x)$, with

$$g(x) = \frac{\lambda(x)}{\int_0^t \lambda(x)dx}, \quad 0 \leq x \leq t. \tag{5}$$

Following the result of Lemma 1, we have

$$\lim_{n\to\infty} d\big(N^-_{n-r+1}(a_n), Po(\mu)\big) = 0$$

for

$$a_n = \sqrt[r]{\frac{r!\mu}{n^{r+1}\int_0^t g^{r+1}(x)\,dx}}. \tag{6}$$

Since

$$\bar{n}_t = \int_0^t \lambda(s)\,ds \to \infty,$$

we replace n with $\bar{n}_t$ in formula (6) and $g(x)$ with $\frac{\lambda(x)}{\int_0^t \lambda(x)dx}$ to verify equation (4). On this bases, we obtain

$$\lim_{t\to\infty} d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) = 0,$$

with

$$a_t = \sqrt[r]{\frac{r!\mu}{\int_0^t \lambda^{r+1}(s)\,ds}}.$$

This completes the proof of Theorem 1.

4. Examples

Haemophilus influenza is a bacterium which engenders an infection in the lungs of humans [6]. The study of the USSs (uptake signal sequences) AAGTGCGGT (USS+) and its inverted complement (USS-) in the *H. influenza* genome (length of 1.83 Mb, Rd strain) provides opportunities for characterizing global genomic inhomogeneities. The result of homogeneous r-scan tests for $r = 1, 2, \ldots, 6$ shows a significant even spacings between the markers such that the USS+ and USS- are remarkably evenly spaced around the genome such that both USS+ positions and USS- positions have respective minimum spacings higher than expected by chance with probability 0.001. This rare possibility may suggest that the homogeniety assumption doesn't fit the real distributions of the markers and an inhomogeneous r-san test should be applied for the marker array.

Another example is the distribution of the palindrome GGCGATCGCC labeled HIP1 (highly iterated palindrome) in the genome of the organism *Synechocystis* (3.6 Mb). *Synechocystis* is thought to be the evolutionary precursor of vascular plant plastids [7]. The photosynthetic endosymbiont became dependent on host genetic information for maintenance and evolved into an organelle specialized for CO_2 fixation. The r-scan analysis of the genome shows in this case a significantly even distribution. The observed minimal *1*-scan spacing is 52 bp (base pair) which is much larger than the threshold of 9 bp with the probability of 0.001. Similar conclusions apply to the r-scan tests of $r = 2, ..., 6$. The even spacing of HIP1 in *Synechocystis* is more dramatic than the situation of USSs in *H. influenza*.

References

[1] Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Annals of Mathematical Statistics.* **27** 272–299. MR81593

[2] Karlin, S. and Rubin, H. (1956). Distributions possessing a monotone likelihood ratio. *Journal of American Statistical Association.* **51** 637–643. MR104303

[3] Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Probab.* **2** 329–357. MR1161058

[4] Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5** 403–434. MR1092983

[5] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics.* Springer Verlag. MR1869112

[6] Karlin, S., Mrázek, J. and Campbell, A. (1996). Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Research.* **24** No. 21, 4263–4272.

[7] Mrázek, J., Bhaya, D., Grossman, A. R. and Karlin, S. (2001). Highly expressed and alien genes of the *Synechocystis* genome. *Nucleic Acids Research.* **29**, No. 7, 1590–1601.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 291–308
© Institute of Mathematical Statistics, 2004

On the strong consistency, weak limits and practical performance of the ML estimate and Bayesian estimates of a symmetric domain in R^k

Wen-Chi Tsai[1] **and Anirban DasGupta**[2]

National Cheng-Chi University and Purdue University

Abstract: This paper considers a problem of estimating an unknown symmetric region in R^k based on n points randomly drawn from it. The domain of interest is characterized by two parameters: size parameter r and shape parameter p. Three methods are investigated which are the maximum likelihood, Bayesian procedures, and a composition of these two. A modification of Wald's theorem as well as a Bayesian version of it are given in this paper to demonstrate the strong consistency of these estimates. We use the measures of symmetric differences and the Hausdorff distance to assess the performance of the estimates. The results reveal that the composite method does the best. Discussion on the convergence in distribution is also given.

1. Introduction

It is a pleasure to write this article for Professor Rubin's Festschrift. I cannot begin to enumerate the things I have learned from him, and the number of times I walked into his office or he walked into mine, drew up a chair, and started a conversation, and opened my eyes. This paper itself is a prime example of how much I benefitted from him in my student days at Purdue.

In biology, the size and shape of home range within a community of a species of animal are often a starting point for the analysis of a social system. In forestry, estimating the geographical edge of a rare species of plant based on sighting of individuals is an important issue as well. The need to estimate an unknown domain by using a set of points sampled randomly from it can also be seen in many other disciplines. See Macdonald et al. (1979), Seber (1986, 1992), and Worton (1987).

If one considers the shape of the unknown domain an infinite-dimensional parameter, the convex hull of the sample will be the maximum likelihood solution. Most of the literature hence focuses on the studies of the convex hull and the results are all for one dimensional and two dimensional regions. Refer to Ripley et al. (1977), Moore (1984), and Bräker et al. (1998).

However, if we use these results in some other applications, for example, recognizing the valid region of predictor variables, which usually involves more than two dimensions, we will then encounter some difficulties in implementation. As the dimensionality rises to higher than three dimensions, where a simple visual illustration is impossible, describing the convex hull of a sample becomes much more

[1]National Cheng-Chi University, 64, Sec. 2, Zhi-nan Rd., Wenshan, Taipei 116, Taiwan, Republic of China. e-mail: wctsai@nccu.edu.tw

[2]Purdue University, Department of Statistics, 150 N. University Street, West Lafayette, IN 47907-2068, USA. e-mail: dasgupta@stat.purdue.edu

Keywords and phrases: ball, Bayes, convex, domain, Euclidean space, extreme value, maximum likelihood, simulation, strong consistency, weak limit.

AMS 2000 subject classifications: 62F10, 62F12, 62F15.

difficult. Hence, a more practical approach for estimating a higher dimensional domain is necessary. Due to this consideration, we would like to characterize the shape of a domain by a finite-dimensional parameter rather than using a non-parametric model to which most literature is devoted. Besides, it is easier to establish properties of estimates of the set of interest under parametric modelling. This would make us more comfortable using these estimates.

Since the configuration of a roughly spherical object is easier characterized, we would like to start our investigation with a particular family of sets, the L_p balls, because of their richness in fitting roughly rounded objects and in deriving pilot theoretical inference.

Let $B_{p,r}$ denote the centered L_p ball with radius r with respect to the metric induced from p-norm in the k-dimensional Euclidean space, R^k; namely

$$B_{p,r} = \{\underset{\sim}{x} \in R^k : \|\underset{\sim}{x}\|_p \leq r\}, \tag{1}$$

here

$$\|\underset{\sim}{x}\|_p = \begin{cases} \left(|x_1|^p + \cdots + |x_k|^p\right)^{1/p} & \text{when } p \text{ is finite,} \\ \max(|x_1|, \ldots, |x_k|) & \text{when } p \text{ is infinite.} \end{cases}$$

We call $\|\cdot\|_p$ the p-norm operator. The unknown set S we wish to estimate will be assumed to be an L_p ball; namely $S = B_{p_0,r_0}$ for some $0 < p_0 \leq \infty$ and $0 < r_0 < \infty$.

Notice that in our approach, the center of symmetry of the domain S is assumed to be known. This will not be exactly true in practice. A short discussion is given in the last section.

Also notice that, when the dimension k equals one, the family of L_p balls becomes the family of closed intervals $[-r, r]$ in the real line. Our one dimensional version of estimating an L_p ball can be viewed as the well known "end-point" problem: estimating the end points a and b by using points randomly selected from $[a, b]$. Also, p does not play any role in characterizing the set S which we wish to estimate when $k = 1$. Therefore throughout this paper, we will take $k \geq 2$. However, the one dimensional case often lends much intuition to the case of higher dimensions.

Now let $\underset{\sim}{x}_1 = (x_{11}, \ldots, x_{1k})', \ldots, \underset{\sim}{x}_n = (x_{n1}, \ldots, x_{nk})'$ denote a realization of n points from the domain S. We would like to estimate S by using these observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$. We will assume that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are independently uniformly drawn from S. It is possible in practice that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are independently drawn from S not uniformly but following a measure μ on R^k other than the Lebesgue measure, truncated to S with finite $\mu(S)$; i.e. $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n \overset{i.i.d.}{\sim} \frac{\mu(\cdot)}{\mu(S)}$. There will be no problem in deriving similar results which we establish in this article if μ is known and for which $B_{p,r}$ is identifiable. However, if μ is unknown, estimating S becomes much more difficult. The reason is that we will be unable to distinguish between a rare event (e.g. the density with respect to μ at a point $\underset{\sim}{x}$ is small) and a null event (e.g. point $\underset{\sim}{x}$ is not in the support S); see Hall (1982).

To summarize, we have taken an interesting problem and analyzed an interesting parametric model. We have given two very general results on strong consistency, and additional results on weak convergence as well as practical evaluation by very detailed numerics. We have indicated how to possibly address more general cases and commented on application. These are the main contributions.

2. Estimation

As the domain S which we wish to estimate is characterized by parameters p and r, a plug-in method can be used to estimate S. We will consider three natural methods

of estimation for p and r: maximum likelihood method, a Bayesian approach, and a combination of these two methods.

The maximum likelihood estimates have a drawback that they underestimate the volume of the true set with probability one and the magnitude of this bias is difficult to evaluate. The Bayesian approach does not have this underestimating problem. However, they are hard to calculate. That is not uncommon in a Bayesian analysis. An alternative approach which combines the maximum likelihood estimate and the Bayesian approach is therefore proposed. This approach treats the volume of the true set as a parameter and estimates it using a Bayesian method. Then it corrects the maximum likelihood estimates for their biases accordingly. We are excited about this approach.

Let us now look at the maximum likelihood method in detail first. Recall that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are uniformly drawn from S. Thus the likelihood function of p and r is

$$L(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) = \frac{1}{\lambda(B_{p,r})^n} \mathbf{1}_{\{(p,r):\underset{\sim}{x}_i \in B_{p,r}\ \forall\ i=1,\ldots,n\}}; \tag{2}$$

here λ is the Lebesgue measure. The formula for the Lebesgue volume of $B_{p,r}$ is

$$\lambda(B_{p,r}) = 2^k r^k \frac{\Gamma(1+\frac{1}{p})^k}{\Gamma(1+\frac{k}{p})}; \tag{3}$$

(see Gradshteyn and Ryzhik (1994), p. 647). If we denote the maximum likelihood estimate of (p, r) by $(\hat{p}_{mle}, \hat{r}_{mle})$, then we have

$$(\hat{p}_{mle}, \hat{r}_{mle}) = \arg\max_{(p,r)} L(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) = \arg \min_{\{(p,r):\underset{\sim}{x}_i \in B_{p,r} \forall i=1,\ldots,n\}} \lambda(B_{p,r}). \tag{4}$$

Moreover, as $\lambda(B_{p,r})$ is an increasing function of r for any fixed p, $(\hat{p}_{mle}, \hat{r}_{mle})$ must satisfy

$$\hat{r}_{mle} = \max_{1 \le i \le n} \|\underset{\sim}{x}_i\|_{\hat{p}_{mle}},$$

and hence

$$\hat{p}_{mle} = \arg\max_p \left(2^k \Big(\max_{1 \le i \le n} \|\underset{\sim}{x}_i\|_p \Big)^k \frac{\Gamma(1+\frac{1}{p})^k}{\Gamma(1+\frac{k}{p})} \right)^{-1}.$$

The profile likelihood of p mostly appears to be unimodal and therefore it is usually not difficult to obtain $\hat{p}_{mle}$ and $\hat{r}_{mle}$ numerically.

Despite this easy characterization of the maximum likelihood estimate, there is a disadvantage in using this estimate. Consider the end-point problem. Suppose $x_1, \ldots, x_n$ are iid Unif([a,b]). It is well known that the maximum likelihood set estimate of $[a, b]$, $[x_{(1)}, x_{(n)}]$, is always contained in the true interval. And therefore the length of the estimated support $[x_{(1)}, x_{(n)}]$ is always shorter than the true length $b - a$. Similarly, when the dimension $k > 1$, the volume of the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ is always smaller than the true volume $\lambda(B_{p_0,r_0})$. The reason is that the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ is the L_p ball which possesses the smallest volume among L_p balls containing all the observations. On the other hand, the true domain evidently contains all the observations. Therefore, we have $\lambda(B_{\hat{p}_{mle}, \hat{r}_{mle}}) \le \lambda(B_{\hat{p}_0, \hat{r}_0})$.

Here we would like to point out that unlike the end-point problem (or the nonparametric setting) where the maximum likelihood interval estimate is always contained in the true interval, the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ does not need to be inside the true set all the time.

Now let us move to the Bayesian approach. We will choose the loss function being

$$l_\lambda(S, \widehat{S}) = \lambda(S \triangle \widehat{S}) \tag{5}$$

where $\triangle$ denotes the symmetric difference operator. If we denote the prior of (p, r) by π, the posterior of (p, r) after observing $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ is

$$\pi(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) \propto \pi(p, r)\frac{1}{\lambda(B_{p,r})^n}\mathbf{1}_{\{(p,r)\,:\,\underset{\sim}{x}_i \in B_{p,r} \forall 1 \le i \le n\}}.$$

Thus the Bayesian estimate based on the loss function (5) is

$$(\hat{p}_{bayes}, \hat{r}_{bayes}) = \arg\min_{(\hat{p},\hat{r})} E_{\pi(p,r|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(\lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})\big). \tag{6}$$

Though we are able to show theoretically that $(\hat{p}_{bayes}, \hat{r}_{bayes})$ is strongly consistent and does not have the underestimating problem like $(\hat{p}_{mle}, \hat{r}_{mle})$ does, however the computation of $(\hat{p}_{bayes}, \hat{r}_{bayes})$ is difficult. The reason is that we do not have a formula of $\lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ for any two general $B_{p,r}$ and $B_{\hat{p},\hat{r}}$ unless $B_{p,r} \subset B_{\hat{p},\hat{r}}$ or $B_{\hat{p},\hat{r}} \subset B_{p,r}$. So, in general, it seems we have to approximate numerically the Bayesian estimate. This is a formidable numerical problem and indeed we are not sure that a minimizer reported by the computer can be trusted.

Therefore an alternative approach is introduced to fix the drawback of the maximum likelihood set estimates which always underestimate the true volume and the disadvantage of the Bayesian estimates which have computational difficulty. The alternative approach tries to estimate the true volume using the Bayesian method, and then corrects the maximum likelihood estimate for bias, based on the estimated volume.

If we consider the loss function

$$l_{vol}(S, \widehat{S}) = \big|\lambda(S) - \lambda(\widehat{S})\big|, \tag{7}$$

it can be analyzed easily. One notes that it only gives a penalty for inaccuracy of volume estimation. Therefore it provides us with only a decision on the volume of S. The following proposition characterizes the class of all Bayesian estimates in this situation.

Proposition 1. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Define the transformation $v(p, r) = \lambda(B_{p,r})$ and denote a median of posterior of $v(p, r)$ by v_m. Then all the L_p balls with volume v_m are Bayesian estimates under the loss* (7).

Proof. Let us denote by $\pi(v|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$ the distribution of $v = v(p, r) = \lambda(B_{p,r})$ with (p, r) having distribution $\pi(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$. The risk

$$\rho(\hat{p}, \hat{r}) = E_{\pi(p,r|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(|\lambda(B_{p,r}) - \lambda(B_{\hat{p},\hat{r}})|\big) = E_{\pi(v|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(|v - v(\hat{p}, \hat{r})|\big) \tag{8}$$

which depends only on $v(\hat{p}, \hat{r})$ and is minimized when $v(\hat{p}, \hat{r})$ equals v_m. Namely $B_{\hat{p},\hat{r}}$ is a Bayes estimate with respect to loss (7) for any $(\hat{p}, \hat{r})$ for which $\lambda(B_{\hat{p},\hat{r}}) = v_m$. □

As there are infinitely many L_p balls with volume v_m, we need a criterion to help us to choose one among these as the estimate of S. A reasonable way to choose a specific L_p ball as an estimate of S could be the pair (p, r) that has the smallest Euclidean distance from $(\hat{p}_{mle}, \hat{r}_{mle})$ among the infinitely many pairs implied in Proposition 1. Thus, this composite approach is to find

$$(\hat{p}_{comb}, \hat{r}_{comb}) = \arg\min_{\{(p,r):\lambda(B_{p,r})=v_m\}} (p - \hat{p}_{mle})^2 + (r - \hat{r}_{mle})^2. \tag{9}$$

We characterize $(\hat{p}_{comb}, \hat{r}_{comb})$ below. It is nice that the characterization is as explicit as it turned out to be.

Proposition 2. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Then $(\hat{p}_{comb}, \hat{r}_{comb})$ in* (9) *exists. Furthermore, $\hat{p}_{comb}$ is the unique root of*

$$p^2(p - \hat{p}_{mle}) - r(p)\left(\psi\left(1 + \frac{k}{p}\right) - \psi\left(1 + \frac{1}{p}\right)\right)\big(r(p) - \hat{r}_{mle}\big) = 0 \tag{10}$$

and $\hat{r}_{comb} = r(\hat{p}_{comb})$. Here ψ is the digamma function and

$$r(p) = \frac{\hat{v}_m^{1/k}}{2} \frac{\Gamma(1 + \frac{k}{p})^{1/k}}{\Gamma(1 + \frac{1}{p})}. \tag{11}$$

Proof. It is clear from (3) that for any fixed $0 < p \leq \infty$, $r(p)$ is the unique solution in r of $\lambda(B_{p,r}) = v_m$. If we can show that $(p - \hat{p}_{mle})^2 + (r(p) - \hat{r}_{mle})^2$ has a unique minimum at some $p = \tilde{p}$, then $(\hat{p}_{comb}, \hat{r}_{comb}) = (\tilde{p}, r(\tilde{p}))$.

This follows on observing that $\lambda(B_{(\hat{p}_{mle}, \hat{r}_{mle})}) = \hat{v}_{mle} < v_m$ which implies that the point $(\hat{p}_{mle}, \hat{r}_{mle})$ is under the curve $(p, r(p))$ in the (p, r) plane. Furthermore, $r(p)$ is strictly convex and differentiable, therefore, we have the existence and uniqueness of $\tilde{p}$, and it must satisfy

$$(\tilde{p} - \hat{p}_{mle}) + \big(r(\tilde{p}) - \hat{r}_{mle}\big) r'(\tilde{p}) = 0. \tag{12}$$

By some further calculations, we obtain $r'(p) = r(p)(\psi(1 + \frac{k}{p}) - \psi(1 + \frac{1}{p}))\frac{1}{p^2}$. From (12) it now follows that $\tilde{p}$ is the unique root of (10). □

3. Strong consistency of the estimates

Maximum likelihood and Bayesian estimates are the most widely used methods of estimation and there is an enormous amount of literature on it. However, a lot of the well known asymptotic theory applies only to those distributions satisfying certain "regularity" conditions. See Lehman and Casella (1998), Le Cam (1953), Huber (1967), and Perlman (1972). One of the conditions requires that the distributions have common support. Apparently, we cannot look for answers in these theories for our problem, as the support is the parameter itself. Consequently, a more direct approach would be necessary and the Wald theorem would be the core key.

3.1. Strong consistency of ML estimate

Let us consider the maximum likelihood estimate first. The most popular strong consistency theorem for the maximum likelihood estimate is due to Wald (1949). It can be applied to the non-regular case. In his paper, Wald gave several conditions to prove a main theorem first. Then he established, essentially through this main theorem, the strong consistency of the maximum likelihood estimate (in fact, of a more general family of estimates) provided that the distributions admit those conditions. Though our problem does not satisfy Wald's conditions, the main theorem, however, holds for our problem. Therefore, here we will try to combine his main theorem and his strong consistency theorem for our maximum likelihood estimate. For completeness, we provide the proof.

Theorem 1 (Wald). *Let P_θ be a distribution with density $f(\underset{\sim}{x}; \theta)$, where $\theta \in \Theta$. Suppose the realizations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ come from $P_{\underset{\sim}{\theta}_0}$ independently for some $\underset{\sim}{\theta}_0 \in \Theta$. Let $\hat{\underset{\sim}{\theta}}_n$ be a function of $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ satisfying*

$$\frac{f(\underset{\sim}{x}_1; \hat{\underset{\sim}{\theta}}_n) \cdots f(\underset{\sim}{x}_n; \hat{\underset{\sim}{\theta}}_n)}{f(\underset{\sim}{x}_1; \underset{\sim}{\theta}_0) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta}_0)} \geq c > 0 \quad \textit{for all } n \quad \textit{and} \quad \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n \textit{ for some positive } c. \tag{13}$$

If for any given neighborhood of θ_0, say U, it also holds that

$$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty}\frac{\sup_{\theta\in\Theta\setminus U} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{f(\underset{\sim}{x}_1;\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta}_0)}=0\right\}=1, \tag{14}$$

then we have

$$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty}\hat{\underset{\sim}{\theta}}_n=\underset{\sim}{\theta}_0\right\}=1. \tag{15}$$

This theorem basically states that if the likelihood ratio of $\underset{\sim}{\theta}$ to $\underset{\sim}{\theta}_0$ is uniformly small as $\underset{\sim}{\theta}$ falls outside any given neighborhood of the true parameter $\underset{\sim}{\theta}_0$, then the estimate $\hat{\underset{\sim}{\theta}}_n$ must be close to $\underset{\sim}{\theta}_0$ since by assumption its likelihood ratio to $\underset{\sim}{\theta}_0$ is always greater than or equal to c (which is greater than 0).

Proof. This theorem does not require that the coordinates of $\underset{\sim}{\theta}_0$ are finite (note that the shape parameter p in our problem can be infinity). But we will give the proof for θ_0 having finite coordinates only to avoid redundancy since the proofs are similar.

To prove (15), it suffices to show that for any neighborhood of $\underset{\sim}{\theta}_0$, say U, $\hat{\underset{\sim}{\theta}}_n$ will fall inside U eventually with probability one. But from (14), one sees that, with probability one, there exists N, which may depend on $\{\underset{\sim}{x}_i\}_{i=1}^{\infty}$, such that

$$\frac{\sup_{\theta\in\Theta\setminus U} f(\underset{\sim}{x}_1,\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta})}{f(\underset{\sim}{x}_1,\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta}_0)}<\frac{c}{2}\ \forall\ n\geq N.$$

However, (13) claims that

$$\frac{f(\underset{\sim}{x}_1,\hat{\underset{\sim}{\theta}}_n)\cdots f(\underset{\sim}{x}_n,\hat{\underset{\sim}{\theta}}_n)}{f(\underset{\sim}{x}_1,\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta}_0)}\geq c>\frac{c}{2}\ \forall\ n\quad\text{and}\quad \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n.$$

Thus, $\hat{\underset{\sim}{\theta}}_n\notin\Theta\setminus U$ when $n\geq N$. Therefore $\hat{\underset{\sim}{\theta}}_n$ belongs to U eventually with probability one, as claimed. □

Since a maximum likelihood estimate, if it exists, obviously satisfies (13) with $c=1$, this theorem also proves the strong consistency of the maximum likelihood estimate provided (14) holds. Fortunately, our family of distributions $\{\text{Unif}(B_{p,r})\}_{\{0<p\leq\infty,0<r<\infty\}}$ satisfies (14).

Lemma 1. *Let P_θ denote* Unif($\mathbf{B}_{p,r}$)*, where $\underset{\sim}{\theta}=(p,r)$ and $\underset{\sim}{\theta}\in\Theta=\{(p,r):0<p\leq\infty,0<r<\infty\}$. Then $\{P_{\underset{\sim}{\theta}}\}_{\underset{\sim}{\theta}\in\Theta}$ satisfies* (14).

Proof. The proof is extremely lengthy and involved. To maintain the flow of this paper, we will only give a rough sketch here and refer the rigorous proof to Tsai (2000).

The basic idea of this proof is as follows. For any given $(p,r)\neq(p_0,r_0)$, one has either $B_{p_0,r_0}\subset_\lambda B_{p,r}$ or $\lambda(B_{p_0,r_0}\setminus B_{p,r})>0$, here $A\subset_\lambda B$ means A is contained in B properly in the Lebesgue measure; i.e. $A\subset B$ and $\lambda(B\setminus A)>0$. In the first situation, we will have the likelihood ratio equal to $(\frac{\lambda(B_{p_0,r_0})}{\lambda(B_{p,r})})^n$ which goes to 0 as n goes to ∞ since $\lambda(B_{p_0,r_0})<\lambda(B_{p,r})$. For the second case, we will, eventually, observe some $\underset{\sim}{x}_i$ not belonging to $B_{p,r}$, which results in the zero value of the likelihood ratio. As a result, (14) shall hold. □

Now by Theorem 1, and Lemma 1, we have the strong consistency of $(\hat{p}_{mle},\hat{r}_{mle})$.

Corollary 1. *Let $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Then the maximum likelihood estimate $(\hat{p}_{mle},\hat{r}_{mle})$ is strongly consistent.*

3.2. Strong consistency of Bayesian estimate

Let us now move to the consistency of the Bayesian estimate. The following is a general result on the strong consistency of the Bayesian estimate under a general assumption on the distribution family and the loss function. Basically, this theorem and its proof are very similar to the Wald Theorem given in the previous section except that we have to include the prior and the loss which are the other elementary components for Bayesian analysis. The generality of this theorem makes it an attractive result of independent interest.

Theorem 2. *Suppose P_θ denotes a distribution with density $f(\underset{\sim}{x};\underset{\sim}{\theta})$, where $\underset{\sim}{\theta} \in \Theta$. Assume the observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are iid with probability $P_{\underset{\sim}{\theta}_0}$ for some $\underset{\sim}{\theta}_0 \in \Theta$. Let $\pi(\underset{\sim}{\theta})$ be a prior of $\underset{\sim}{\theta}$ and $l(\underset{\sim}{\theta}, \underset{\sim}{\hat{\theta}})$ be a loss function such that*

$$\int_\Theta \pi(\underset{\sim}{\theta})\, \mathrm{d}\underset{\sim}{\theta} < \infty \quad \textit{and} \quad \int_\Theta l(\underset{\sim}{\theta}, \underset{\sim}{\theta}_0)\pi(\underset{\sim}{\theta})\, \mathrm{d}\underset{\sim}{\theta} < \infty. \tag{16}$$

Then the Bayesian estimate will converge to $\underset{\sim}{\theta}_0$ with probability one (under $P_{\underset{\sim}{\theta}_0}$) provided that for any neighborhood of $\underset{\sim}{\theta}_0$, say $\underset{\sim}{U}$, there exist sets $W \subset V \subset \underset{\sim}{U}$ satisfying

(i) $$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty} \frac{\sup_{\theta\in\Theta\setminus V} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})} = 0\right\} = 1,$$

(ii) $$\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta} > 0, \quad \textit{and}$$

(iii) $$\inf_{\underset{\sim}{\hat{\theta}}\in U^c, \underset{\sim}{\theta}\in V} l(\underset{\sim}{\theta}, \underset{\sim}{\hat{\theta}}) - l(\underset{\sim}{\theta}, \underset{\sim}{\theta}_0) \geq \epsilon \quad \textit{for some } \epsilon > 0.$$

Remark 1. In this theorem, there is a condition on all components of the problem (likelihood, prior, and loss). Condition (i) states that the likelihood ratio for $\underset{\sim}{\theta}$ far away from $\underset{\sim}{\theta}_0$ versus $\underset{\sim}{\theta}$ near $\underset{\sim}{\theta}_0$ is uniformly small. Condition (ii) requires that the prior puts a positive mass around the true $\underset{\sim}{\theta}_0$. Condition (iii) says that the loss function does punish for bad decisions. These conditions are all quite mild.

Proof. We divide the proof into several steps for clarity and ease of understanding.
Step 1: Let us denote the posterior of $\underset{\sim}{\theta}$ given $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ by

$$\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) \propto \Pi_{i=1}^n f(\underset{\sim}{x}_i, \underset{\sim}{\theta})\, \pi(\underset{\sim}{\theta}).$$

Then the posterior expected loss for decision $\underset{\sim}{\hat{\theta}}$ is $\rho(\underset{\sim}{\hat{\theta}}) = E_{\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\left(l(\underset{\sim}{\theta}, \underset{\sim}{\hat{\theta}})\right)$, and the Bayesian estimate is $\underset{\sim}{\hat{\theta}}_{bayes} = \arg\min_{\underset{\sim}{\hat{\theta}}\in\Theta} \rho(\underset{\sim}{\hat{\theta}})$.

To prove the strong consistency of $\underset{\sim}{\hat{\theta}}_{bayes}$, it suffices to show that for any neighborhood of $\underset{\sim}{\theta}_0$, say U, $\underset{\sim}{\hat{\theta}}_{bayes}$ will fall inside U eventually with probability one (under $P_{\underset{\sim}{\theta}_0}$). Now, let V, W, and ϵ be as defined in condition (i), (ii), and (iii). We will show that

$$P_{\underset{\sim}{\theta}_0}\left\{\inf_{\underset{\sim}{\hat{\theta}}\in U^c} \rho(\underset{\sim}{\hat{\theta}}) \geq \rho(\underset{\sim}{\theta}_0) + \frac{1}{4}\epsilon \quad \text{eventually}\right\} = 1. \tag{17}$$

This will imply

$$P_{\underset{\sim}{\theta}_0}\left\{\arg\min_{\underset{\sim}{\hat{\theta}}} \rho(\underset{\sim}{\hat{\theta}}) \in U \ \text{ eventually}\right\} = 1,$$

proving this theorem.

Step 2: In this step, we will break $\rho(\hat{\underset{\sim}{\theta}}) - \rho(\underset{\sim}{\theta}_0)$ into several terms whose magnitudes are easier to investigate. Note that

$$\begin{aligned}
&\rho(\hat{\underset{\sim}{\theta}}) - \rho(\underset{\sim}{\theta}_0) \\
&\quad = E_{\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}(l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0)) \\
&\quad = \frac{\int_\Theta (l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0))\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \\
&\quad = \frac{\int_V (l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0))\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \cdot \frac{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \\
&\qquad + \frac{\int_{V^c} l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}})\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} - \frac{\int_{V^c} l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0)\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \\
&\quad = (I)\cdot(II) + ((III) - (IV))\,. \qquad (18)
\end{aligned}$$

Step 3: In this step, we will show that (I) is always greater than or equal to ϵ. From condition (iii), it is easy to see that

$$(I) \geq \frac{\int_V \epsilon \cdot \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} = \epsilon \ \text{ for all } \ \hat{\underset{\sim}{\theta}} \in U^c. \qquad (19)$$

Step 4: Now, we claim

$$P_{\underset{\sim}{\theta}_0}\{(II) \longrightarrow 1\} = 1. \qquad (20)$$

Note that

$$\begin{aligned}
|1-(II)| &= \frac{\int_{V^c} \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \leq \frac{\int_{V^c} \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \\
&\leq \frac{\sup_{\underset{\sim}{\theta}\in V^c} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})} \frac{\int_{V^c} \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}. \qquad (21)
\end{aligned}$$

From condition (i), together with condition (ii) and (16), we get that the upper bound (21) converges to 0 with probability one. Consequently, claim (20) is proved.
Step 5: Now let us look at the term $(III) - (IV)$. We would like to show that

$$P_{\underset{\sim}{\theta}_0}\left\{\inf_{\hat{\underset{\sim}{\theta}}\in U^c}\{(III)-(IV)\} \geq -\frac{1}{4}\epsilon \ \text{ eventually}\right\} = 1. \qquad (22)$$

Since (III) is nonnegative, we have $(III) - (IV) \geq -(IV)$ which does not depend on $\hat{\underset{\sim}{\theta}}$. Moreover

$$0 \leq (IV) \leq \frac{\sup_{\underset{\sim}{\theta}\in V^c} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})} \frac{\int_{V^c} l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0)\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}$$

Again from conditions (i) and (ii), and (16), we get that (IV) converges to 0 with probability one. Therefore (22) is true.
Step 6: Finally, as an immediate consequence of (18), (19), (20), and (22) together, we obtain (17). This theorem therefore follows. □

Now we would like to apply Theorem 2 to our problem. The following lemma says that the distribution family $\mathrm{Unif}(B_{p,r})$, and the loss function $l((p,r),(\hat{p},\hat{r})) = \lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ satisfy condition (i) and (iii) of Theorem 2.

Lemma 2. *Let P_θ denote the distribution* Unif$(B_{p,r})$*, where $\theta = (p, r)$ and $\theta \in \Theta = \{(p, r) : 0 < p \le \infty, 0 < r < \infty\}$. Let $l((p, r), (\hat{p}, \hat{r})) = \lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ be the loss function and let π be the prior on θ. Suppose $\theta_0 = (p_0, r_0)$ is a fixed point in Θ and π is positive in some neighborhood of (p_0, r_0). Then for any neighborhood of (p_0, r_0), say U, there exist sets $W \subset V \subset U$ such that the conditions* (i), (ii), *and* (iii) *in Theorem* 2 *hold.*

Proof. The idea of the proof is not difficult. However, the proof is very lengthy. Refer to Tsai (2000). □

Now, as an application of Theorem 2, we have the strong consistency of $(\hat{p}_{bayes}, \hat{r}_{bayes})$ as follows:

Corollary 2. *Let $x_1, \ldots, x_n$ be iid with distribution* Unif$(B_{p,r})$*. Suppose the true value of (p, r) is denoted by (p_0, r_0). Let π be a proper prior on (p, r) such that π is positive in some neighborhood of (p_0, r_0). Assume also that $E_{\pi(p,r)}(\lambda(B_{p,r}))$ is finite. Then the Bayesian estimate under the loss $l((p, r), (\hat{p}, \hat{r})) = \lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ converges to (p_0, r_0) with probability one.*

Proof. From the assumption on π, one has

$$E_{\pi(p,r)}\left[l((p, r), (\hat{p}, \hat{r}))\right] \le E_{\pi(p,r)}\left[\lambda(B_{p,r}) + \lambda(B_{p_0,r_0})\right] < \infty.$$

Thus, the corollary follows from Theorem 2 and Lemma 2 immediately. □

3.3. Strong consistency of combined estimate

Now we discuss the strong consistency of a combined estimate $(\hat{p}_{comb}, \hat{r}_{comb})$. Recall that it is the pair (p, r) closest to the initial guess $(\hat{p}_{mle}, \hat{r}_{mle})$ with $\lambda(B_{p,r})$ equal to v_m, the posterior median of $v = \lambda(B_{p,r})$. From Corollary 1 and Corollary 3 below, $(\hat{p}_{mle}, \hat{r}_{mle})$ and v_m are both strongly consistent in the respective parameters. One may expect, therefore, that the combined estimate will be strongly consistent as well. We give a general theorem in this direction below. Again the generality makes it an appealing theorem of independent interest.

Theorem 3. *Let $x_1, \ldots, x_n$ be a sample from a distribution P_θ, $\theta \in \Theta$. Let Θ be a metric space with a metric d. Let $\hat{\theta}_n$ and $\hat{\beta}_n$ be functions of the observations $x_1, \ldots, x_n$ such that $\hat{\theta}_n$ and $\hat{\beta}_n$ converge almost surely to θ and $\beta(\theta)$, respectively, under P_θ, where $\beta(\theta)$ is a function of θ. Suppose $\tilde{\theta}_n \stackrel{def}{=} \arg\min_{\{\theta:\beta(\theta)=\hat{\beta}_n\}} d(\hat{\theta}_n, \theta)$ exists and is unique. Then $\tilde{\theta}_n$ converges to θ with probability one if for any $\epsilon > 0$, there exists a neighborhood of $\beta(\theta)$ contained in $\beta(B_d(\theta, \epsilon))$, where $B_d(\theta, \epsilon)$ is the ϵ-ball centered at θ with respect to the metric d.*

Proof. To prove the strong consistency of $\tilde{\theta}_n$, it is enough to show that for any $\epsilon > 0$,

$$P_\theta\left\{ d(\theta, \tilde{\theta}_n) < 3\epsilon \text{ eventually} \right\} = 1. \tag{23}$$

By assumption, there exists a neighborhood of $\beta(\theta)$, say B, contained in $\beta(B_d(\theta, \epsilon))$; so, if $\hat{\beta}_n \in B$, there exists $\tilde{\theta}$ within ϵ distance of θ such that $\beta(\tilde{\theta}) = \hat{\beta}_n$. Then, one has

$$d(\hat{\theta}_n, \tilde{\theta}_n) = \min_{\{\theta:\beta(\theta)=\hat{\beta}_n\}} d(\hat{\theta}_n, \theta) \le d(\hat{\theta}_n, \tilde{\theta}) \le d(\hat{\theta}_n, \theta) + d(\theta, \tilde{\theta}) \le d(\theta, \hat{\theta}_n) + \epsilon,$$

which implies

$$d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}_n) \le d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) + d(\hat{\underset{\sim}{\theta}}_n, \tilde{\underset{\sim}{\theta}}_n) \le 2d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) + \epsilon.$$

Furthermore, if $d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) < \epsilon$, then we have $d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}_n) < 3\epsilon$. On the other hand, $\hat{\underset{\sim}{\theta}}_n$ and $\hat{\underset{\sim}{\beta}}_n$ are strongly consistent for $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ respectively. This implies

$$P_{\underset{\sim}{\theta}}\left\{ d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) < \epsilon \text{ and } \hat{\underset{\sim}{\beta}}_n \in B \text{ eventually} \right\} = 1.$$

This proves (23) and hence the theorem. □

To apply the above general theorem to our problem, we need the strong consistency of v_m. This will be implied by the following theorem which generalizes Theorem 2.

Theorem 4. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a sample from P_θ with density $f(\underset{\sim}{x}; \underset{\sim}{\theta})$, where $\underset{\sim}{\theta} \in \Theta$. Suppose we are interested in estimating a function $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ (rather than $\underset{\sim}{\theta}$) itself and the loss is a function of $\underset{\sim}{\theta}$ through $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, say $l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}})$. Denote the true value of $\underset{\sim}{\theta}$ by $\underset{\sim}{\theta}_0$ and the prior of $\underset{\sim}{\theta}$ by π. Assume $\int \pi(\underset{\sim}{\theta}) \mathrm{d}\underset{\sim}{\theta} < \infty$ and $\int l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \underset{\sim}{\beta}(\underset{\sim}{\theta}_0))\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta} < \infty$. Then the Bayesian estimate of $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, $\arg\min_{\underset{\sim}{\beta}} E_{\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)}(l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}}))$, converges to $\underset{\sim}{\beta}_0 \equiv \underset{\sim}{\beta}(\underset{\sim}{\theta}_0)$ with probability one under $P_{\underset{\sim}{\theta}_0}$ provided that for any neighborhood of $\underset{\sim}{\beta}_0$, say B, there exists sets $W \subset V \subset \underset{\sim}{\beta}^{-1}(B)$ satisfying*

(i) $$P_{\underset{\sim}{\theta}_0}\left\{ \lim_{n\to\infty} \frac{\sup_{\underset{\sim}{\theta} \in \Theta \setminus V} f(\underset{\sim}{x}_1; \underset{\sim}{\theta}) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta} \in W} f(\underset{\sim}{x}_1; \underset{\sim}{\theta}) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta})} = 0 \right\} = 1,$$

(ii) $$\int_W \pi(\underset{\sim}{\theta}) \mathrm{d}\underset{\sim}{\theta} > 0, \quad \textit{and}$$

(iii) $$\inf_{\hat{\underset{\sim}{\beta}} \in B^c, \underset{\sim}{\theta} \in V} l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}}) - l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \underset{\sim}{\beta}(\underset{\sim}{\theta}_0)) \ge \epsilon \quad \textit{for some } \epsilon > 0.$$

Remark 2. Theorem 2 is a special case of Theorem 4 when we take $\underset{\sim}{\beta}(\underset{\sim}{\theta}) = \underset{\sim}{\theta}$. Moreover, in this theorem, $\underset{\sim}{\beta}$ does not have to be one-to-one and $\underset{\sim}{\beta}^{-1}(B)$ is defined as $\{ \underset{\sim}{\theta} : \underset{\sim}{\beta}(\underset{\sim}{\theta}) \in B \}$.

Proof. The proof is exactly the same as that of Theorem 2. □

We now apply Theorem 4 to prove the strong consistency of v_m.

Corollary 3. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Define $v = v(p, r) = \lambda(B_{p,r})$. Let π be a prior on (p, r) and v_m the posterior median of v. Let also (p_0, r_0) denote the true value of (p, r). If π is positive in a neighborhood of (p_0, r_0), then v_m converges to $v(p_0, r_0)$ with probability one.*

Proof. Denote $v(p_0, r_0)$ by v_0. Let B be a neighborhood of v_0. Without loss of generality, we can assume $B = (v_0 - \delta, v_0 + \delta)$ for some $\delta > 0$. Since $v(p, r)$ is a continuous function of (p, r), there exists a neighborhood of (p_0, r_0), say U, such that $U \subset v^{-1}(B')$, where $B' = (v_0 - \frac{\delta}{3}, v_0 + \frac{\delta}{3})$. Then by Lemma 2, there exist sets $W \subset V \subset U$ such that conditions (i) and (ii) in Theorem 4 hold.

Furthermore, if $(p, r) \in V$, one has $v(p, r) \in B'$, which implies

$$|v(p, r) - v(p_0, r_0)| < \frac{\delta}{3} \quad \text{and} \quad |v(p, r) - \hat{v}| > \frac{2\delta}{3} \quad \text{for all } \hat{v} \in B.$$

This gives us condition (iii) of Theorem 4.

The corollary, therefore, follows from Theorem 4 immediately. □

Corollary 3 endows us with the strong consistency of v_m needed to apply the general result of Theorem 3. We are now ready to prove the strong consistency of $(\hat{p}_{comb}, \hat{r}_{comb})$.

Corollary 4. *The estimator* $(\hat{p}_{comb}, \hat{r}_{comb})$ *defined in Section* 2.3 *is strongly consistent.*

Proof. To prove this proposition, we will apply Theorem 3 for the case when the true value, p_0, of p is finite. When p_0 is infinity, we will prove this proposition directly. Recall that

$$(\hat{p}_{comb}, \hat{r}_{comb}) = \arg \min_{\{(p,r):\lambda(B_{p,r})=v_m\}} \left[(p-\hat{p}_{mle})^2 + (r-\hat{r}_{mle})^2\right].$$

Also Corollary 1 and Corollary 3 give us the strong consistency of $(\hat{p}_{mle}, \hat{r}_{mle})$ and v_m respectively.

Case 1 $p_0 = \infty$: By (10) and the fact that $v_m \geq \lambda(B_{\hat{p}_{mle},\hat{r}_{mle}})$, $\hat{p}_{comb}$ must be greater than $\hat{p}_{mle}$. As $\hat{p}_{mle}$ converges to $p_0 = \infty$ with probability one, so does $\hat{p}_{comb}$. Furthermore, Proposition 2 also gives $\hat{r}_{comb} = \frac{v_m}{2}^{1/k} \frac{\Gamma^{1/k}(1+\frac{k}{\hat{p}_{comb}})}{\Gamma(1+\frac{1}{\hat{p}_{comb}})}$. Thus the strong consistency of $\hat{r}_{comb}$ follows from the strong consistency of $\hat{p}_{comb}$ and v_m immediately.

Case 2 $p_0 < \infty$: We will prove this case as an application of Theorem 3. For any given $\epsilon > 0$, let us take $B = (\lambda(B_{p_0,(r_0-\epsilon)^+}), \lambda(B_{p_0,r_0+\epsilon}))$. It is easy to see that for any $b \in B$, there exists $(r_0-\epsilon)^+ < r < r_0+\epsilon$ such that $v(p_0, r) = b$ and certainly the distance between (p_0, r) and (p_0, r_0) is smaller than ϵ. Therefore, the assumptions in Theorem 3 are all satisfied. This proposition for the case when $p_0 < \infty$ follows. □

4. Discussion

This section will first compare the performance of the maximum likelihood estimate with the combined estimate, especially when the sample size is small. Recall that the calculation of Bayes estimate is difficult. Then, some simulation and conjectures on the asymptotic distribution of the estimates will be given as, unfortunately, they are very hard. We end with a brief discussion for the case when the center of symmetry of the true set is unknown.

4.1. Comparison of $(\hat{p}_{mle}, \hat{r}_{mle})$ *and* $(\hat{p}_{comb}, \hat{r}_{comb})$

We remarked that the combined estimate can be principally considered as a dilation of the maximum likelihood estimate. Our simulation will try to examine: (i) in what fashion the combined estimate dilates the maximum likelihood estimate, (ii) if it indeed helps with regard to underestimation of the volume of the true set, and (iii) if the choice of the prior on p and r affects the performance of the combined estimate.

The tables and figures referenced below are based on a simulation of size 750 with true $(p, r) = (2, 1)$, dimension $k = 2$, and sample size $n = 10$. We consider three respective priors on (p, r). They are $\pi_1(p, r) = pe^{-p}re^{-r}$, $\pi_2(p, r) = \frac{1}{2}p^2e^{-p}re^{-r}$, and $\pi_3(p, r) = \frac{2}{\pi(1+p^2)}re^{-r}$ respectively. We denote each of the corresponding combined estimates by $(\hat{p}_{comb1}, \hat{r}_{comb1})$, $(\hat{p}_{comb2}, \hat{r}_{comb2})$, and $(\hat{p}_{comb3}, \hat{r}_{comb3})$, respectively.

Table 1 gives the mean and the standard error of the volume of $B_{\hat{p}_{mle},\hat{r}_{mle}}$, $B_{\hat{p}_{comb1},\hat{r}_{comb1}}$, $B_{\hat{p}_{comb2},\hat{r}_{comb2}}$, and $B_{\hat{p}_{comb3},\hat{r}_{comb3}}$, and their symmetric difference as well as their Hausdorff distances to the true set. This table shows that the volumes of the

Table 1: The mean and standard error (in parentheses) from a size 750 simulation of the volume of the maximum likelihood estimate and the combined estimates with respect to three different priors on (p, r) and the symmetric difference distances and the Hausdorff distances to the true set.

true $(p,r) = (2,1)$	$k = 2$, $n = 10$		
$(\hat{p}, \hat{r}) =$	$\lambda(B_{\hat{p},\hat{r}})$	$d_\lambda(B_{p,r}, B_{\hat{p},\hat{r}})$	$d_H(B_{p,r}, B_{\hat{p},\hat{r}})$
$(\hat{p}_{mle}, \hat{r}_{mle})$	2.70519(0.294328)	0.466593(0.290792)	0.118659(0.077068)
$(\hat{p}_{comb1}, \hat{r}_{comb1})$	3.10145(0.321972)	0.336655(0.200423)	0.095912(0.065939)
$(\hat{p}_{comb2}, \hat{r}_{comb2})$	3.10321(0.332331)	0.342807(0.201751)	0.097780(0.066906)
$(\hat{p}_{comb3}, \hat{r}_{comb3})$	3.13368(0.344959)	0.351907(0.202675)	0.098770(0.065463)

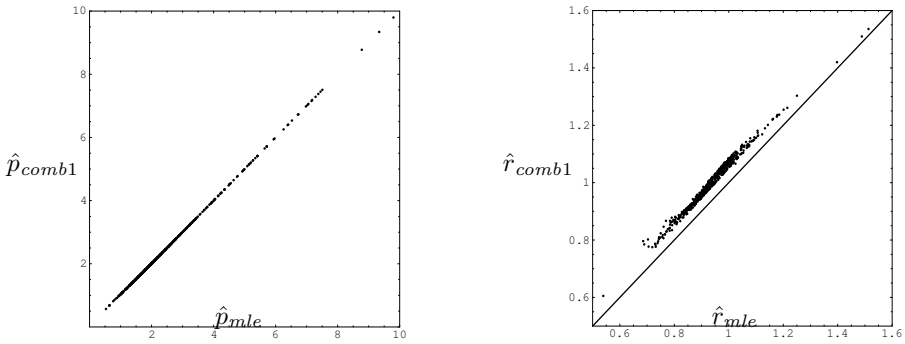


Figure 1: Scatter plots of $(\hat{p}_{mle}, \hat{p}_{comb})$ and $(\hat{r}_{mle}, \hat{r}_{comb})$.

combined estimates are much closer to the true volume (which is $\pi = 3.14159$), but with a higher variance, than that of the maximum likelihood estimate. Moreover, the distances, either one, of the combined estimates to the true set are about 20% to 30% less compared to the maximum likelihood estimate. It also appears that the selection of the prior does not affect the performance of the combined estimate very much.

Figure 1 plots $\hat{p}_{comb1}$ against $\hat{p}_{mle}$ and $\hat{r}_{comb1}$ against $\hat{r}_{mle}$. We see that the scatter plot of $(\hat{p}_{mle}, \hat{p}_{comb1})$ is virtually the 45 degree line; $(\hat{r}_{mle}, \hat{r}_{comb1})$s', on the other hand, all fall above the 45 degree line. We have similar results for the other two combined estimators. So, $B_{\hat{p}_{comb},\hat{r}_{comb}}$ may indeed be considered as if it was dilated from $B_{\hat{p}_{mle},\hat{r}_{mle}}$ by enlarging only the radius r while keeping p essentially fixed at $\hat{p}_{mle}$. This is interesting.

4.2. Convergence in distribution

In this section, some simulation and conjectures on the asymptotic distribution of the maximum likelihood estimate will be given. Figure 2 shows several scatter plots of $(n(\hat{p}_{mle} - p), n(\hat{r}_{mle} - r))$ with $p = 2$, $r = 1$, and various sample sizes. We believe that when the true value of p is finite, $(n(\hat{p}_{mle} - p), n(\hat{r}_{mle} - r))$ converges to some nondegenerate distribution which puts all its mass in the half plane: $\{(x, y) : y \leq$

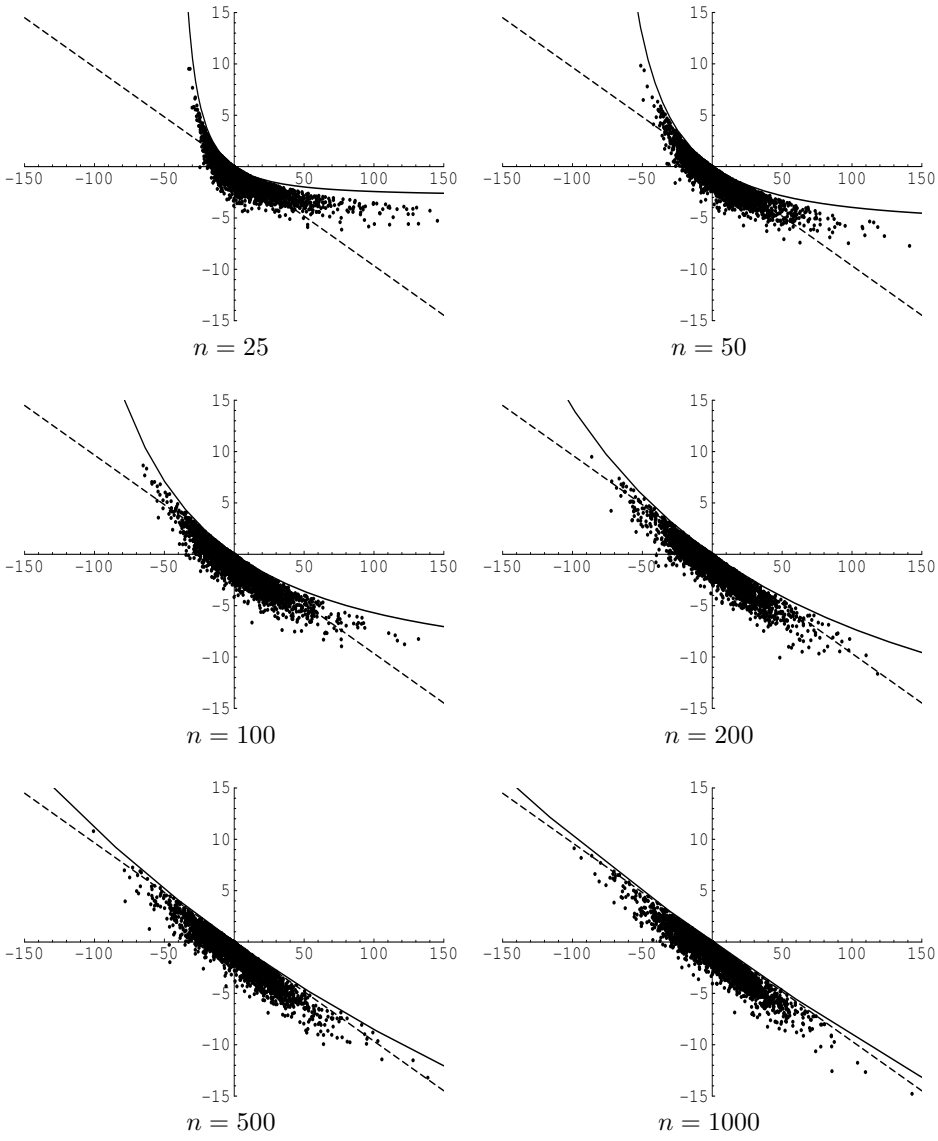


Figure 2: Scatter plots of $(n(\hat{p}_{mle}-p), n(\hat{r}_{mle}-r))$ with $(p,r)=(2,1)$ and $k=2$. Solid line is $(n(t-p), n(\frac{\Gamma^{1/k}(1+\frac{k}{t})}{\Gamma(1+\frac{1}{t})}\frac{\Gamma(1+\frac{1}{p})}{\Gamma^{1/k}(1+\frac{k}{p})}-1)r)$, where t ranges from 0 to ∞. Broken line is the straight line through the origin with slope $\frac{-1}{p^2}(\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p}))$.

$\frac{-r}{p^2}\{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})x\}$. It is also obvious that the correlation of $\hat{p}_{mle}$ and $\hat{r}_{mle}$ is negative. When $\hat{p}_{mle}$ overestimates the true p, the corresponding $\hat{r}_{mle}$ will then likely underestimate the true r, and vice versa.

Figure 3 gives scatter plots of $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle}-r))$ for the case where $(p,r)=(\infty,1)$. It seems that $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle}-r))$ converges to some nondegenerate distribution having support in the fourth quadrant. Interestingly, the convergence rates seem dependent on the true value of p.

In fact, these conjectures were inspired by the case when one of the parameters (p or r) is known. A summary for the behavior of $\hat{p}_{mle}$ when r is assumed to be known is given below. Similar results can also be derived for the case when p is assumed to be known. See Tsai (2000) for details.

If we assume r is known, say, $r=r_0$, the characterization of the maximum likelihood estimate of p becomes very simple. We are in fact able to give the exact distribution of $\hat{p}_{mle}$ and therefore the weak convergence result for $\hat{p}_{mle}$. The idea of getting this result is very simple. Indeed this problem can be converted to an endpoint problem if we consider the new random variables $z_i=\lambda(B_{p_{r_0}(\underset{\sim}{x}_i),r_0})$, where $B_{p_{r_0}(x_i),r_0}$ is the smallest L_p ball containing $\underset{\sim}{x}_i$ with radius r_0. It can be easily shown that Z_i's are independently and identically distributed with value between 0 and the volume of the true domain and $\lambda(B_{\hat{p}_{mle},r_0})=\max_{1\le i\le n} z_i$, whose asymptotic distribution is well known. Thus we have the following weak convergence result for $\hat{p}_{mle}$ when the true r is known.

Proposition 3. *Suppose* $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ *are iid from* $\mathrm{Unif}(B_{p,r_0})$*, where* $0<r_0<\infty$ *is known. Let* G *denote an exponential random variable with mean* 1*. Then*
(I) *when* $p<\infty$*,*

$$n(\hat{p}_{mle}-p)\xrightarrow{\mathcal{D}}-\frac{p^2}{k}\left(\frac{1}{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})}\right)G, \tag{24}$$

where ψ *is the digamma function, and*
(II) *when* $p=\infty$*,*

$$\sqrt{n}\frac{1}{\hat{p}_{mle}}\xrightarrow{\mathcal{D}}\sqrt{\frac{12}{\pi^2k(k-1)}}\sqrt{G}. \tag{25}$$

Remark 3. Note that when $p<\infty$, interestingly, the asymptotic variance, $(\frac{p^2}{k}(\frac{1}{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})}))^2$, is a decreasing function of the dimension k. It appears that the curse of dimensionality does not show up in this problem. To the contrary, for estimation of the single shape parameter p, it is beneficial to have a large k!

Remark 4. In fact, $n(\lambda(B_{\hat{p}_{mle},r_0})-\lambda(B_{p,r_0}))\xrightarrow{\mathcal{D}}-\lambda(B_{p,r_0})G$. If we divide both sides by the true volume, this expression tells us that the proportion of the uncommon part between the estimate and the true set (to the true set) converges with the rate n to an exponential distribution. It does not relate to the true set. The convergence rates of $\hat{p}_{mle}$, however, do depend on p. It is interesting that the speed of convergence of $\hat{p}_{mle}$ slows down from n to $\sqrt{n}$ discontinuously as p changes from finite to infinite. We believe that this phenomenon is caused by the difficulty of "catching the corners" of a square, for example. This is also interesting.

4.3. Unknown center of symmetry

In practice, the center of symmetry of the object usually would not be known. It then has to be estimated. In this section, we will have a brief examination of this situation.

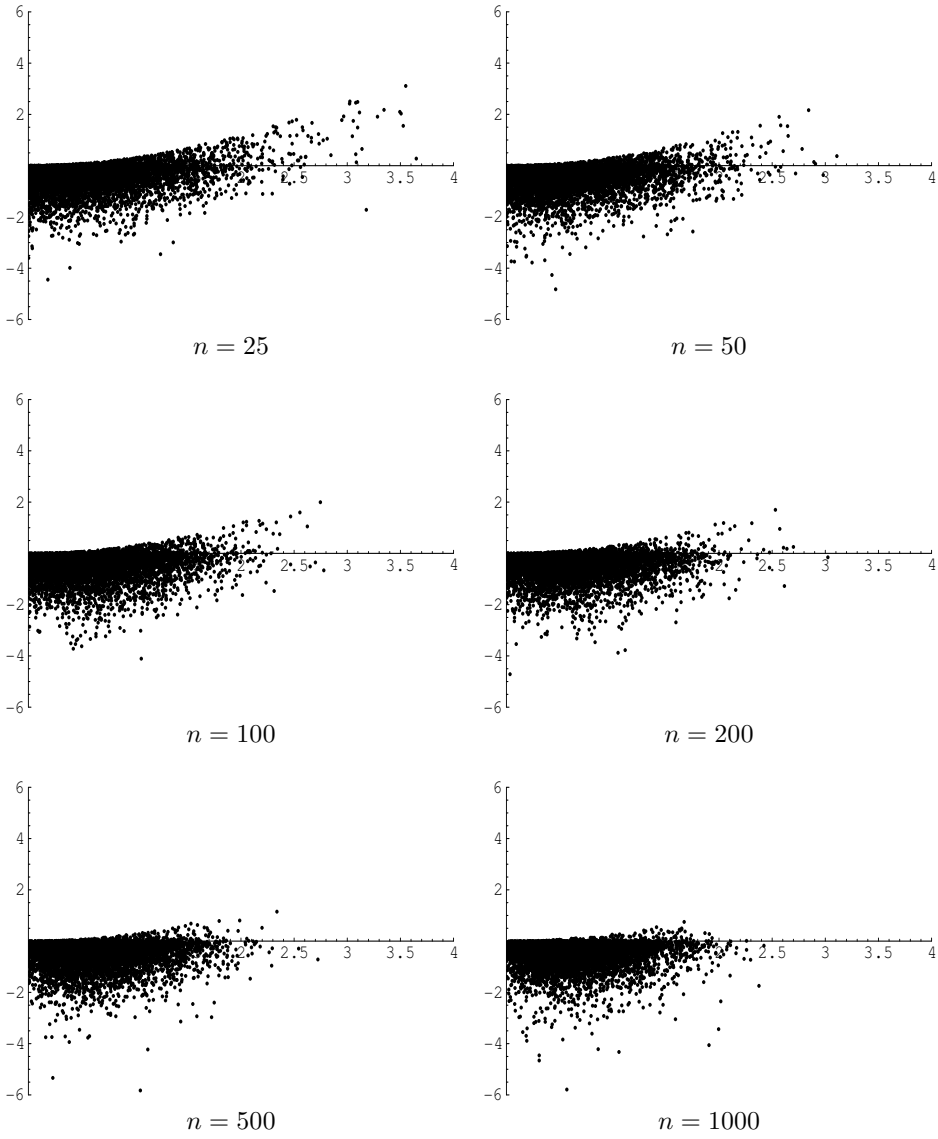


Figure 3: Scatter plots of $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle} - r))$ with $(p, r) = (\infty, 1)$ and $k = 2$ for different sample sizes.

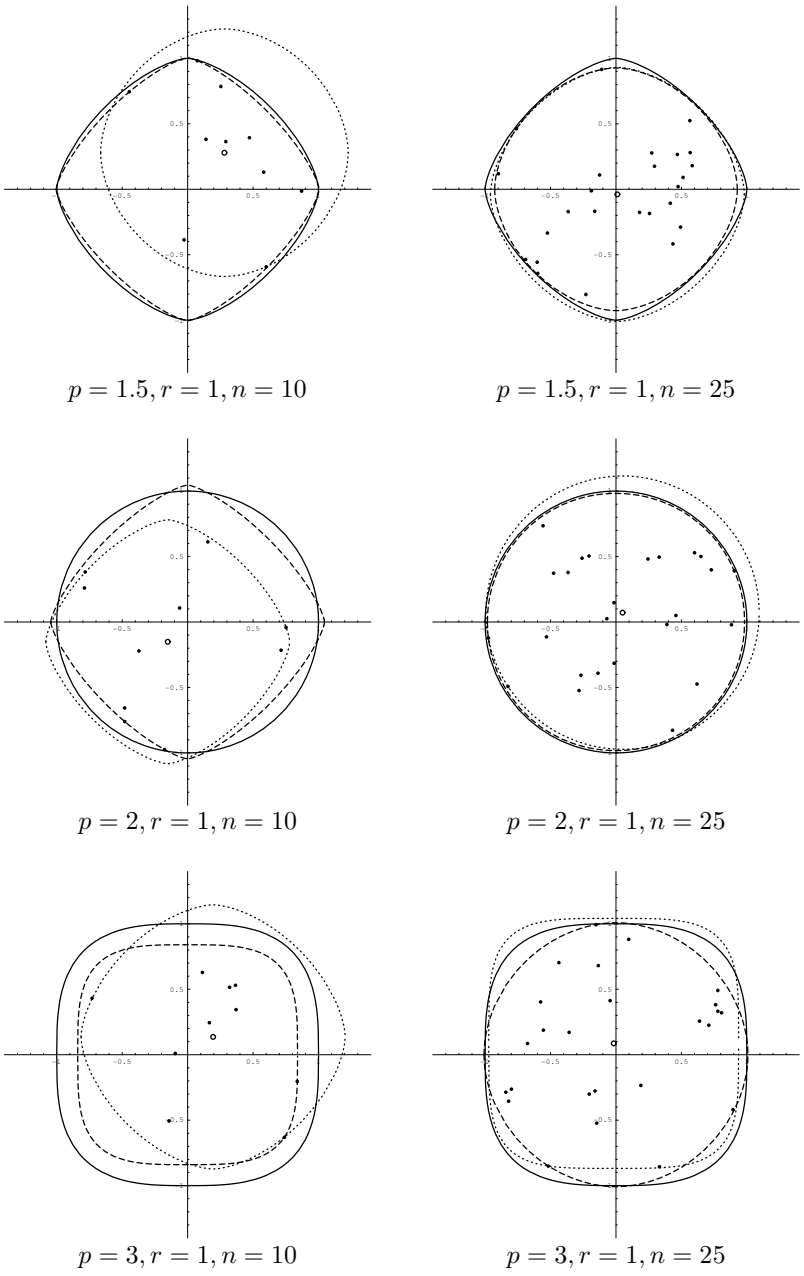


Figure 4: Visual display of the set estimate when the center is unknown. The region bounded by the solid curves is the true set, by the broken or the dotted curve is the maximum likelihood estimate with the center assumed to be known or estimated by the mean of the observations, respectively. The conspicuous circle is the estimated center and the dots are the observations.

Apparently, it is not easy to estimate the center together with the shape parameter p and the size parameter r by using the maximum likelihood method. See Amey et al. (1991) for some calculations. Besides, the problem of underestimating of the volume of the maximum likelihood estimate in this situation will be more serious. Therefore, it may be preferable to estimate the center by some other external methods. We tried the mean of the observations, and the L_2 median (spatial median) (which minimizes $\Sigma_{1\le i\le n}\|\underset{\sim}{x}_i - \underset{\sim}{u}\|_2$ over $\underset{\sim}{u}$). It turns out that the mean of the observations performs better than the L_2 median. Therefore, here we attempt to check how the estimate may be influenced if the center is unknown and is estimated by the mean of the observations. Figure 4 gives a visual comparison between the maximum likelihood estimates with center treated to be known and with center estimated by the mean of the observations. It can be seen that the shape of the estimates can vary very much depending on whether the center is known or estimated. But the estimate of the size parameter does not differ that much. Moreover, the volume of the maximum likelihood estimate with the center estimated by the mean of the observations can exceed the volume of the true set. When the realizations cluster to one side with some observations appearing in the far opposite direction, apparently the estimate can miss the true set badly. Therefore, constructing a better estimate for the center of symmetry of the true set is important.

References

[1] Amey, A., Gupta, A. K., Norton, V., and Székeley, G. J. (1991). On Maximum Likelihood Estimation of the Center of a Centrally Symmetric Convex Domain, *Jour. Stat. Comp. and Simu.*, **39**, 1–7. MR1109862

[2] Bräker, H., Hsing, T., and Bingham, N. H. (1998). On the Hausdorff Distance Between a Convex Set and an Interior Random Convex Hull, *Adv. Appl. Probab.*, **30**, 295–316. MR1642840

[3] (1994). Gradshteyn, I. S. and Ryzhik, I. M. Table of Integrals, Series, and Products, Academic Press, New York,

[4] Hall, P. (1982). ON Estimating the Endpoint of a Distribution, *Ann. Statist.*, **10**(2), 556–568. MR653530

[5] Huber, P. J. (1967). The Behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. Le Cam, L. M. and Neyman, J., Vol. 1, Univ. Calif. Press, Berkeley and Los Angeles, pp. 221–233. MR216620

[6] Le Cam, L. M. (1953). *On some Asymptotic Properties of Maximum Likelihood Estimate and Related Bayes Estimates*, Univ. Calif. Publ. Statist., Vol. 1, pp. 277–330. MR54913

[7] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, Springer, New York. MR1639875

[8] Macdonald, D. W. and Ball, F. G. and Hough, N. G. (1979). The evaluation of Home Range Size and Configuration Using Radio Tracking Data, *A Handbook on Biotelemetry and Radio Tracking*, eds. Amlaner, C. J. and Macdonald, D. W., Oxford, New York, pp. 405–424.

[9] Moore, M. (1984). On the Estimation of a Convex Set, *Ann. Statist.*, **12** 3, 1090–1099. MR751296

[10] Perlman, M. D. (1972). On the Strong Consistency of Approximate Maximum Likelihood Estimators, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, eds. Le Cam, L. M. and Neyman, J. and Scott, E., Vol. 1, Univ. Calif. Press, Berkeley and Los Angeles, pp. 263–281. MR405668

[11] Ripley, B. D. and Rasson, J.-P. (1977). Finding the Edge of a Poisson Forest, *J. Appl. Probab.*, **14** 3, 483–491. MR451339

[12] Seber, G. A. F. (1986). A Review of Estimating Animal Aboundance, *Biometrics*, **42**, 267–292. MR872967

[13] Seber, G. A. F. (1992). A Review of Estimating Animal Aboundance II, *Internat. Statist. Rev.*, **60** 2, 129–166.

[14] Tsai, W.-C. (2000). Some Contribution to Stochastic Geometry, Department of Statistics, Purdue University.

[15] Wald, A. (1949). Note on the Consistency of the maximum likelihood, *Ann. Math. Statist.*, **20**, 596. MR32169

[16] Worton, B. J. (1987). A Review of Models of Home Range for Animal Movement, *Ecol. Modelling*, **38**, 277–298.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 309–318
© Institute of Mathematical Statistics, 2004

Maximum likelihood estimation for the contact process

Marta Fiocco[1] and Willem R. van Zwet[2]

Leiden University Medical Centre and University of Leiden

Abstract: The contact process—and more generally interacting particle systems—are useful and interesting models for a variety of statistical problems. This paper is a report on past, present and future of research by the authors concerning the problem of estimating the parameters of the contact process. A brief review of published work on an ad-hoc estimator for the case where the process is observed at a single (large) time t is given in Section 1. In Section 2 we discuss maximum likelihood estimation for the case where the process is observed during a long time interval $[0, t]$. We construct the estimator and state its asymptotic properties as $t \to \infty$, but spare the reader the long and tedious proof. In Section 3 we return to the case where the process is observed at a single time t and obtain the likelihood equation for the estimator. Much work remains to be done to find a workable approximation to the estimator and study its properties. Our prime interest is to find out whether it is significantly better than the ad-hoc estimator in Section 1.

It was a joy to write this paper for Herman Rubin's festschrift. To this is added the bonus that Herman will doubtless solve our remaining problems immediately.

1. Introduction

The contact process was introduced and first studied by Harris (1974). It is described as follows. At every time $t \geq 0$, every point (or site) x in the d-dimensional integer lattice Z^d is in one of two possible states that we shall call infected and healthy. The process starts at time $t = 0$ with a non-empty set $A \subset Z^d$ of infected sites. At time $t \geq 0$, the state of the site $x \in Z^d$ will be indicated by a random variable $\xi_t^A(x)$, given by

$$\xi_t^A(x) = \begin{cases} 1, & \text{if site x is infected at time t} \\ 0, & \text{if site x is healthy at time t .} \end{cases} \tag{1.1}$$

The function $\xi_t^A : Z^d \to \{0, 1\}$ describes the state of the process at time t and $\xi_0^A = 1_A$, the indicator function of the set A.

The evolution of this $\{0, 1\}$-valued random field is described by the following dynamics. A healthy site is infected independently and at rate $\lambda > 0$ by each of its $2d$ immediate neighbors that is itself infected. An infected site recovers at

[1]Department of Medical Statistics, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands. e-mail: m.fiocco@lumc.nl

[2]Mathematical Institute, University of Leiden, PO Box 9512, 2300 RA Leiden, The Netherlands. e-mail: vanzwet@math.leidenuniv.nl

Keywords and phrases: contact process, supercritical contact process, maximum likehood, counting process.

AMS 2000 subject classifications: 62M30.

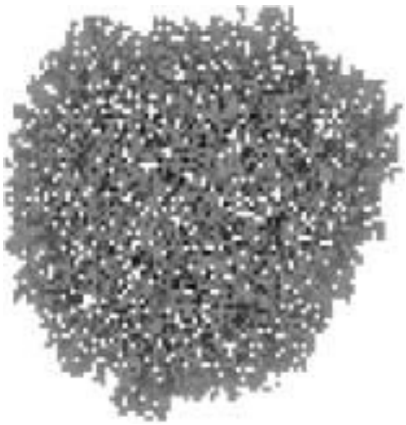


Figure 1: The process $\xi_t^{\{0\}}$ for $\lambda = 3$ and $\mu = 1$ after $30,000$ steps. Infected sites are represented by gray 1×1 squares. A darker gray level indicates a longer duration of the present infection.

rate $\mu > 0$. Given the configuration ξ_t^A at time t, the processes involved are independent until a change occurs. For $d = 2$ the contact process is a simplified model for the spread of an infection or, more generally, of a biological species in the plane. The growth of a forest is an example if diseased and healthy are interpreted as presence and absence of a tree in a square centered at the lattice site.

In Figure 1 we show the process that started with a single infected site at the origin with $\lambda = 3$ and $\mu = 1$ after 30,000 steps, i.e. 30,000 infections and recoveries. Infected sites are indicated by gray 1×1 squares. An additional feature of this figure is that for each infected site we have kept track of the number of steps since it was last infected and have indicated this by the gray level at that site: the darker the gray level, the older the present infection at a site. If we view the process as a model for the growth of a forest, then the gray level indicates the age of the tree. Obviously, the older trees are in the center of the picture away from the boundary.

It is sometimes convenient to represent the state of the contact process at time t by the set of infected sites rather than by the function $\xi_t^A : Z^d \to \{0, 1\}$. Usually, this set is also denoted by ξ_t^A. Thus, by an abuse of notation, we write

$$\xi_t^A = \big\{x \in Z^d : \xi_t^A(x) = 1\big\}. \tag{1.2}$$

Let

$$\tau^A = \inf\big\{t : \xi_t^A = \emptyset\big\} \tag{1.3}$$

denote the time the infection dies out with the convention that $\tau^A = \infty$ if the infection survives forever. For a set $C \subset R^d$ and $a > 0$, we write $aC = \{ax : x \in C\}$. For sets C and D in R^d, $C \oplus D = \{x+y : x \in C, y \in D\}$ will denote their Minkowski sum and we define

$$H_t^A = \bigcup_{0 \le s \le t} \xi_s^A \oplus Q[-1/2, 1/2]^d \,. \tag{1.4}$$

Thus H_t^A is obtained from the set of sites that have been infected up to and including time t by replacing each site by a hypercube with sides of length 1 centered at this site.

The contact process has been the subject of extensive studies during the past decades. We list a few of its basic properties.

Property 1. If $\rho = \lambda/\mu$ exceeds a certain critical value ρ_d , then the infection will continue forever (i.e. $\tau^A = \infty$) with positive probability depending on the dimension d and the initial set A. This is called the supercritical case. On the other hand, if $\rho \leq \rho_d$, then the infection will eventually die out (i.e. $\tau^A < \infty$) with probability 1. We shall restrict attention to the supercritical case.

Property 2. In the supercritical case, there exist positive constants C and γ such that for every $t > 0$ and $A \subset Z^d$ with cardinality $|A|$,

$$\mathbb{P}\big(t < \tau^A < \infty\big) \leq Ce^{-\gamma t}, \qquad \mathbb{P}\big(\tau^A < \infty\big) \leq e^{-\gamma|A|}. \tag{1.5}$$

In particular, if A is infinite, then in the supercritical case the infection will survive forever.

Property 3. The distribution of the set ξ_t^A converges weakly to a limit distribution

$$\mathbb{P}\big(\tau^A < \infty\big)\delta_\emptyset + \mathbb{P}\big(\tau^A = \infty\big)\nu, \tag{1.6}$$

where $\delta_\emptyset$ denotes the measure that assigns probability 1 to the empty set and ν is the equilibrium measure depending only on ρ and the dimension d. Thus, given that the process survives forever—which is possible only in the supercritical case—it tends in distribution to ν. Here weak convergence coincides with convergence in distribution of the finite dimensional projections $\xi_t^A \cap F$, (i.e. $\{\xi_t^A(x) : x \in F\}$) for finite $F \subset Z^d$.

Property 4. There exists a bounded convex set $U \subset R^d$ with the origin as an interior point such that for every bounded $A \subset Z^d$, $\epsilon > 0$ and $t \to \infty$,

$$(1-\varepsilon)tU \subset \ H_t^A \ \subset (1+\varepsilon)tU, \tag{1.7}$$

eventually almost surely on the set $\{\tau^A = \infty\}$ where ξ_t^A survives forever. Thus if the infection persists, then for large t, $\mathrm{H}^A{}_t$ will grow linearly in t in every direction and $\mathrm{t}^{-1}\ \mathrm{H}^A{}_t$ will assume the shape of U. Moreover, on the set $\{\tau^A = \infty\}$ and for large t, the distribution of $\xi_t^A \cup (1-\epsilon)tU$ will approach its asymptotic distribution under the equilibrium measure ν in a sense that we shall not make precise here.

For these facts and other related matters the reader may consult Liggett (1985 & 1999).

The contact process and its many possible generalizations provide an interesting class of models for problems in spatial statistics and image analysis. In Fiocco & van Zwet (2003a & b) we began a statistical study of the supercritical contact process $\xi_t^{\{0\}}$ that starts with a single infected site at the origin and is conditioned on survival, i.e. on $\{\tau^{\{0\}} = \infty\}$. For this process we considered the simplest possible statistical problem, that is, to estimate the parameters of the contact process based on observing the set of infected sites at a single (large but unknown) time t. This corresponds to the realistic situation when one observes a large forest that has obviously been there for a long time without any knowledge when it began. On the

basis of such an observation it is clear that one can only estimate $\rho = \lambda/\mu$ but not λ and μ individually, as without knowing t, one cannot distinguish between observing the processes with parameters $c\lambda$ and $c\mu$ at time t/c for different values of $c > 0$. Equivalently, one may set $\mu = 1$ arbitrarily and estimate λ.

For any $x, y \in Z^d$ and $C \subset R^d$, let $|x - y| = \sum_{1 \le i \le d} |x_i - y_i|$ denote the L_1-distance of x and y, and define

$$k_t^{\{0\}}(x) = \big(1 - \xi_t(x)\big) \sum_{|x-y|=1} \xi_t^{\{0\}}(y), \tag{1.8}$$

$$n_t^{\{0\}}(C) = \sum_{x \in C \cap Z^d} \xi_t^{\{0\}}(x), \quad k_t^{\{0\}}(C) = \sum_{x \in C \cap Z^d} k^{\{0\}}(x). \tag{1.9}$$

Notice that $n_t^{\{0\}}(C)$ is simply the number of infected sites in C and $k_t^{\{0\}}(C)$ equals the number of neighboring pairs of infected and healthy sites, with the healthy site in C. For $x \in Z^d$, the flip rates at time t equal $\lambda k_t^{\{0\}}(x)$ and $\mu \xi_t^{\{0\}}(x)$ for the transitions $0 \to 1$ and $1 \to 0$ respectively and hence the number $n_t^{\{0\}}(C)$ of infected sites increases by 1 at time t with rate $k_t^{\{0\}}(C)$ and decreases by 1 with rate $n_t^{\{0\}}(C)$. In Property 4 above, we explained that on $\{\tau^{\{0\}} = \infty\}$ and at a large time t, the process will have progressed past the set $(1-\epsilon)tU$ and will be close to equilibrium there. This implies that the rate of increase of $n_t^{\{0\}}((1-\epsilon)tU)$ should approximately equal its rate of decrease, so that $\lambda k_t^{\{0\}}((1-\epsilon)tU) \approx \mu n_t^{\{0\}}((1-\epsilon)tU)$. Hence on $\{\tau^{\{0\}} = \infty\}$, $n_t^{\{0\}}((1-\epsilon)tU)/k_t^{\{0\}}((1-\epsilon)tU)$ should be a plausible estimator of $\rho = \lambda/\mu$, or of λ if one assumes $\mu = 1$, where it not for the fact that U is unknown. However, one can show that for every $\epsilon > 0$, the convex hull $\mathcal{C}(\xi_t^{\{0\}})$ of the set of infected sites $\xi_t^{\{0\}}$ satisfies

$$(1-\epsilon)tU \subset \mathcal{C}\Big(\xi_t^{\{0\}}\Big) \subset (1+\epsilon)tU, \tag{1.10}$$

eventually almost surely on $\{\tau^{\{0\}} = \infty\}$, so that $\mathcal{C}(\xi_t^{\{0\}})$ apparently approximates tU. If, for any $\delta > 0$, we define

$$C_t = (1-\delta)\mathcal{C}\Big(\xi_t^{\{0\}}\Big), \tag{1.11}$$

then for some $\epsilon > 0$, (1.10) ensures that $C_t \subset (1-\epsilon)tU$ eventually a.s. on $\{\tau^{\{0\}} = \infty\}$. Hence

$$\tilde{\rho}_t = \frac{n_t^{\{0\}}(C_t)}{k_t^{\{0\}}(C_t)}, \tag{1.12}$$

would seem to be a sensible estimator of ρ, given that the process $\xi_t^{\{0\}}$ will survive forever. Indeed we prove in Fiocco & van Zwet (2003b) that conditional on $\{\tau^{\{0\}} = \infty\}$, $\tilde{\rho}_t$ is a strongly consistent and asymptotically normal estimator of ρ, that is, as $t \to \infty$,

$$\tilde{\rho}_t \to \rho \qquad a.s. \qquad |C_t|_d^{1/2}\big(\tilde{\rho}_t - \rho\big) \xrightarrow{D} N\big(0, \tau^2\big) \tag{1.13}$$

Here $|C_t|_d$ denotes the cardinality of $C_t \cap Z^d$, or alternatively, the Lebesgue measure of C_t , and an explicit expression for τ^2 is available. For our purposes we merely note

that this implies that $(\tilde{\rho}_t - \rho) = \mathcal{O}_P(t^{-d/2})$ on $\{\tau^{\{0\}} = \infty\}$. Simulation confirms that the estimator behaves as predicted by the asymptotics (Fiocco (1997)).

For the estimator $\tilde{\rho}_t$ to perform well asymptotically as well as in simulations, it is essential that δ should indeed be positive in (1.11). At time t the process has spread approximately to tU, but beyond $(1-\epsilon)tU$ it is not yet in equilibrium and our argument fails. This is also intuitively obvious: having just reached the boundary of tU, the infected sites beyond $(1-\epsilon)tU$ should be less dense than they are closer to the origin where the infection arrived earlier and had time to achieve equilibrium. Beyond $(1-\epsilon)tU$ the fraction of infected sites should be too small, but among the infected sites the fraction with healthy neighbors should be too large. As a result $\tilde{\rho}_t$ should systematically underestimate ρ if δ is taken to be zero and simulation not only confirms this, but shows that in this case the estimator is bad. This effect also shows up asymptotically as $t \to \infty$. If $\delta = 0$, we can still prove consistency but no longer asymptotic normality. Shrinking the convex hull $\mathcal{C}(\xi_t^{\{0\}})$ to obtain the mask C_t for the estimator is essential for obtaining a satisfactory estimator.

Two minor problems are left. First, shrinking $\mathcal{C}(\xi_t^{\{0\}})$ towards the origin to obtain C_t is possible only if one knows where the origin is, i.e. where the infection has started. Generally this is not known: one sees the forest today, but not when or where it began. Of course one can estimate the origin in many different ways, for instance by averaging the locations of the infected sites. Shrinking towards this estimated origin will not influence the asymptotic behavior of the estimator. A more elegant solution is to replace the shrinking of $\mathcal{C}(\xi_t^{\{0\}})$ by another operation that removes the sites near the boundary of this set. Such operation is called peeling, where one removes layer after layer of sites on the boundary of the convex hull. In general, almost any reasonable type of shrinking will leave the asymptotic behavior of the estimator unchanged as long as the same fraction of sites is removed. Simulation suggests that this fraction should be around 20-30%, decreasing with increasing t.

Second, our analysis refers only to the behavior of the process - and hence of the estimator - on the set where $\tau^{\{0\}} = \infty$. Obviously, if $\tau^{\{0\}} < \infty$, there is not much to observe for sufficiently large t, since the infection will have died out. On the other hand, we can not know with certainty at any finite time t that we are indeed in the case where $\tau^{\{0\}} = \infty$, so one may wonder whether asymptotic results for $t \to \infty$ that are valid only on the set $\tau^{\{0\}} = \infty$ have any statistical significance. However, (1.5) ensures that having survived until a large time t, the infection will survive forever with overwhelming probability. Asymptotic results conditional on $\tau^{\{0\}} = \infty$ are therefore the same as those conditioned on $\tau^{\{0\}} \geq t$, that is, on the infection being present when observed.

2. Maximum likelihood for the fully observed process

Having briefly described the statistical results obtained for the contact process observed at a single time t, we now turn to the case where this process is observed continuously on the interval $[0,t]$ for a known (large) $t > 0$. In this case it should be possible to estimate λ and μ separately, rather than just their ratio $\rho = \lambda/\mu$. In fact we shall derive the maximum likelihood estimators of these parameters.

Let $0 < T_1 < \ldots < T_N$ denote the times when the contact process undergoes a change in the time interval $[0,t]$ and, for $i = 1, 2, \ldots, N$, let x_i denote the site at which the change occurs at time T_i. It will be convenient to write $T_0 = 0$ and

$T_{N+1} = t$ and $\xi_i = \xi_{T_i}^{\{0\}}$ for the configuration of the process at time T_i. Given the configuration ξ_{i-1} at time T_{i-1}, the rate of change at site x equals

$$r_i(x) = \begin{cases} \lambda \sum_{|x-y|} \xi_{i-1}(y) & \text{if } \xi_{i-1}(x) = 0 \\ \mu & \text{if } \xi_{i-1}(x) = 1, \end{cases} \tag{2.1}$$

and the total rate of change at any site is given by

$$R_i = \sum_{x \in Z^d} r_i(x) = \lambda k_{T_{i-1}}^{\{0\}}\big(Z^d\big) + \mu n_{T_{i-1}}^{\{0\}}\big(Z^d\big) \tag{2.2}$$

It follows that the likelihood of the observed process on $[0, t]$ is given by

$$L(\lambda, \mu) = \prod_{1 \le i \le N} R_i \exp\big\{-R_i[T_i - T_{i-1}]\big\}[r_i(x_i)/R_i] \exp\big\{-R_{N+1}[t - T_N]\big\}.$$

Hence

$$\log L(\lambda, \mu) = -\sum_{1 \le i \le N+1} R_i[T_i - T_{i-1}] + U_t \log \lambda + D_t \log \mu + h\big(\xi^{\{0\}}\big)$$

where U_t and D_t are the number of upward and downward jumps of the process on $[0, t]$, i.e.

$$U_t = \#\big\{0 \le i \le N-1 : \xi_{i-1}(x_i) = 0\big\} = \#\big\{1 \le i \le N : \xi_i(x_i) = 1\big\}, \tag{2.3}$$

$$D_t = \#\big\{0 \le i \le N-1 : \xi_{i-1}(x_i) = 1\big\} = \#\big\{1 \le i \le N : \xi_i(x_i) = 0\big\}, \tag{2.4}$$

and $h(\xi^{\{0\}})$ depends on the process $\{\xi_s^{\{0\}} : 0 \le s \le t\}$, but not on the parameters λ and μ.

Define

$$A_t = \int_0^t k_s^{\{0\}}\big(Z^d\big)\, ds, \qquad B_t = \int_0^t n_s^{\{0\}}\big(Z^d\big)\, ds. \tag{2.5}$$

As $n_s^{\{0\}}(Z^d)$ and $k_s^{\{0\}}(Z^d)$ are constant for $s \in [T_{i-1}, T_i)$ and $T_{N+1} = t$, (2.2) implies that

$$\sum_{1 \le i \le N+1} R_i[T_i - T_{i-1}] = \lambda A_t + \lambda B_t$$

and hence

$$\log L(\lambda, \mu) = -\lambda A_t - \mu B_t + U_t \log \lambda + D_t \log \mu + h\big(\xi^{\{0\}}\big). \tag{2.6}$$

Differentiating with respect to λ, and μ we find that the maximum likelihood estimators $\hat{\lambda}_t$ and $\hat{\mu}_t$ of λ, and μ are given by

$$\hat{\lambda}_t = \frac{U_t}{A_t}, \qquad \hat{\mu}_t = \frac{D_t}{B_t}. \tag{2.7}$$

The maximum likelihood estimator of $\rho = \lambda/\mu$ therefore equals

$$\hat{\rho}_t = \frac{U_t B_t}{D_t A_t}. \tag{2.8}$$

As in the previous section we can prove that conditional on $\{\tau^{\{0\}} = \infty\}$, these estimators are strongly consistent and asymptotically normal, but converge to the parameter to be estimated at the faster rate $\mathcal{O}(t^{-(d+1)/2})$. Thus conditional on $\{\tau^{\{0\}} = \infty\}$ and as $t \to \infty$,

$$\hat{\lambda}_t \to \lambda \quad a.s., \quad \hat{\mu}_t \to \mu \quad a.s. \quad \hat{\rho}_t \to \rho \quad a.s., \tag{2.9}$$

$$\begin{aligned} t^{(d+1)/2}(\hat{\lambda}_t - \lambda) &\to N(0, \sigma_\lambda^2), \\ t^{(d+1)/2}(\hat{\mu}_t - \mu) &\to N(0, \sigma_\mu^2), \\ t^{(d+1)/2}(\hat{\rho}_t - \rho) &\to N(0, \sigma_\rho^2), \end{aligned} \tag{2.10}$$

again with explicit expressions for the variances being available. The proof is long and involved and will be given elsewhere.

There are two different ways of looking at these maximum likelihood estimators heuristically. First we may observe that the counting process U_t has compensator λA_t and since $A_t \to \infty$ if $\tau^{\{0\}} = \infty$, $\hat{\lambda}_t = U_t/A_t$ should approximate λ. Similarly, μB_t is the compensator of D_t and $\hat{\mu}_t = D_t/B_t$ should approximate μ on $\{\tau^{\{0\}} = \infty\}$. Hence $\hat{\lambda}_t$, $\hat{\mu}_t$ and $\hat{\rho}_t$ are plausible estimators of λ, μ and ρ.

However, one may also be interested in a comparison of the maximum likelihood estimator $\hat{\rho}_t$ based on the fully observed process $\{\xi_s^{\{0\}} : 0 \le s \le t\}$, and the ad-hoc estimator $\tilde{\rho}_t$ of Section 1, which is based on observing $\xi_t^{\{0\}}$ at the single time t. We assume throughout that $\tau^{\{0\}} = \infty$. First of all, (B_t/A_t) in (2.8) estimates the same quantity ρ as $\tilde{\rho}_t = n_t^{\{0\}}(C_t)/k_t^{\{0\}}(C_t)$ in (1.12). On the one hand, B_t/A_t averages information over the interval $[0, t]$ and should therefore have a variance of a smaller order than $n_t^{\{0\}}(C_t)/k_t^{\{0\}}(C_t)$. On the other hand B_t/A_t uses the entire set of infected points and its healthy neighbors, and we have argued in Section 1, that without shrinking this set, this will lead to underestimating ρ. The factor U_t/D_t in (2.8) now serves to correct this negative bias. In equilibrium, the number of upward and downward jumps should approximately cancel out, but near the boundary of the set of infected points, equilibrium has not yet set in. In fact, the number of infected sites $U_t - D_t + 1$ grows roughly as a constant factor times the Lebesgue measure of tU, that is, at the rate of t^d. Individually, both U_t and D_t are counting processes and easily seen to be of order t^{d+1}. Hence $(U_t/D_t) - 1$ is positive and decreases at the rate t^{-1}, so that the factor U_t/D_t in (2.8) does serve to correct the negative bias which does indeed decrease like t^{-1}.

The asymptotic results (1.13) and (2.10) imply that the estimators $\tilde{\rho}_t$ and $\hat{\rho}_t$ of ρ have random errors of orders $\mathcal{O}(t^{-d/2})$ and $\mathcal{O}(t^{-(d+1)/2})$ respectively. Hence the maximum likelihood estimator $\hat{\rho}_t$ based on observing the entire process $\{\rho_s^{\{0\}} : 0 \le s \le t\}$, is asymptotically an order of magnitude better than the ad-hoc estimator $\tilde{\rho}_t$ based on a single observation of $\xi_t^{\{0\}}$. In Figure 2 we show a single run of simulated values of both estimators after $500, 1,000, 1,500, \ldots, 20,000$ jumps of the process for $\lambda = 0.8$ $\mu = 1$, and hence $\rho = 0.8$. For the ad-hoc estimator, the shrinking of the convex hull of infected sites $\mathcal{C}(\xi_t^{\{0\}})$ to obtain the mask C_t has been achieved by peeling rather than multiplication by $(1 - \delta)$ as is done in (1.11). Peeling fractions of 30%, 50% and 70% were used. It appears that the maximum likelihood estimator is indeed superior.

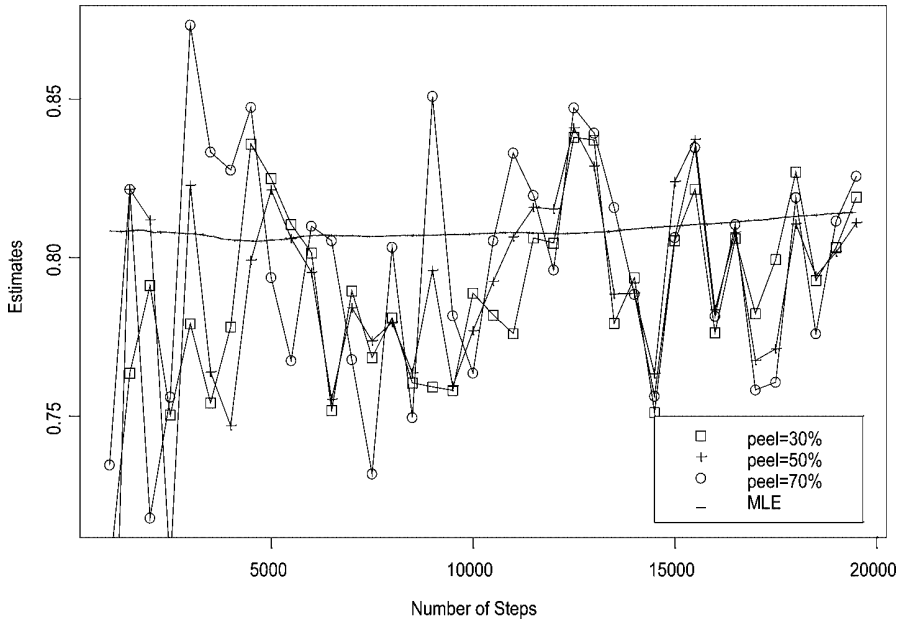


Figure 2: Maximum likelihood estimator $\hat{\rho}_t$ and the ad-hoc estimator $\tilde{\rho}_t$.

3. Maximum likelihood for the singly observed process

As we pointed out in Section 1, one will rarely have the opportunity to observe the process throughout a time interval $[0, t]$. In most cases one will have to be content with a single observation of the process at a (large but unknown) time t. For the latter situation we reported on the study of an ad-hoc estimator $\tilde{\rho}_t = n_t^{\{0\}}(C_t)/k_t^{\{0\}}(C_t)$ of ρ, and noted that it is essential to choose the mask C_t well inside the convex hull $\mathcal{C}(\xi_t^{\{0\}})$ of the set of infected points in order to avoid underestimating ρ. Of course, we are still interested in finding and studying the maximum likelihood estimator for this case, if only to see whether or not it will improve substantially on the ad-hoc estimator.

Obviously this is going to be a difficult assignment. In Section 2 we studied the maximum likelihood estimator for the fully observed process and discovered two things. First of all this estimator uses the ratio of (the integrals of) $n_s^{\{0\}}(Z^d)$ and $k_s^{\{0\}}(Z^d)$ and we conclude that the use $n_t^{\{0\}}/k_t^{\{0\}}$ in the ad-hoc estimator was a good idea. Second, the bias correction was achieved by the correction factor U_t/D_t, which is a rather more subtle way to achieve this than by discarding a sizeable fraction of the data, as is done for the ad-hoc estimator. It therefore seems plausible that the maximum likelihood estimator for the singly observed process will also depend on $n_t^{\{0\}}/k_t^{\{0\}}$, and that conditional expectations of the numbers of upward and downward jumps in $[0, t]$ given $\xi_t^{\{0\}}$ will also play a part.

Since we observe $\xi_t^{\{0\}}$ at an unknown time t and have no information about the times of any of the jumps, we may discard the time element entirely and view the process as a sequence of configurations $\xi_1^{\{0\}}, \xi_2^{\{0\}}, \ldots, \xi_{n-1+2k}^{\{0\}}$ after the first, second,..., $(n-1+2k)$th jumps that take place consecutively at sites $x_1, x_2, \ldots, x_{n-1+2k}$ during the time interval $[0, t]$. The final configuration $\xi_{n-1+2k}^{\{0\}}$ equals

the observed configuration $\xi_t^{\{0\}}$. For some k, $(n-1+k)$ of the jumps are upward (i.e. $n_t^{\{0\}}$ increases by 1 at this jump) and k are downward. Hence the total increase $n_t^{\{0\}}(Z^d) - n_0^{\{0\}}(Z^d) = n_t^{\{0\}}(Z^d) - 1$ of the number of infected points must equal $(n-1+k)-k = n-1$, so that we must have $n = n_t^{\{0\}}(Z^d)$. Finally, we write $n_{i-1}^{\{0\}}(x_i)$ and $k_{i-1}^{\{0\}}(x_i)$ for the values of $n_t^{\{0\}}$ and $k_t^{\{0\}}$ after the time of the $(i-1)st$ jump at the site x_i where the next jump will occur, and $n_i^{\{0\}}(Z^d)$ and $k_i^{\{0\}}(Z^d)$ for the values of $n_t^{\{0\}}(Z^d)$ and $k_t^{\{0\}}(Z^d)$ immediately after the $i-th$ jump. The probability of $(n-1+k)$ upward and k downward jumps consecutively at sites $x_1, x_2, \ldots, x_{n-1+2k}$ equals

$$\prod_{1\le i\le n-1+2k} \frac{\lambda k_{i-1}^{\{0\}}(x_i) + \mu n_{i-1}^{\{0\}}(x_i)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}$$
$$= \lambda^{n-1-k}\mu^k \prod_{1\le i\le n-1+2k} \frac{k_{i-1}^{\{0\}}(x_i) + n_{i-1}^{\{0\}}(x_i)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}$$

because either $k_{i-1}^{\{0\}}(x_i)$ or $n_{i-1}^{\{0\}}(x_i)$ vanishes. It follows that the likelihood is given by

$$L^*(\lambda,\mu) = \sum_{0\le k<\infty} {\sum}^* \lambda^{n-1+k}\mu^k \prod_{1\le i\le n-1+2k} \frac{k_{i-1}^{\{0\}}(x_i) + n_{i-1}^{\{0\}}(x_i)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}, \tag{3.1}$$

where $\sum^*$ denotes summation over all possible sequences $\xi_1^{\{0\}}, \xi_2^{\{0\}}, \ldots, \xi_{n-1+2k}^{\{0\}}$ for which $\xi_{n-1+2k}^{\{0\}}$ is the first configuration equaling $\xi_t^{\{0\}}$, and $n = n_t^{\{0\}}(Z^d)$. As we noted in Section 1 we can only estimate $\rho = \lambda/\mu$, but not λ and μ separately as t is unknown. However, we can still maximize the likelihood L^* as a function of λ and μ, but we shall find that both likelihood equations are identical. If U and D denote the number of upward and downward jumps until the configuration equals $\xi_t^{\{0\}}$ for the first time, then differentiation with respect to λ and μ yields the likelihood equations

$$E(U|\xi_t^{\{0\}}) = E\left[\sum_{1\le i<U+D}\left[\frac{\lambda k_{i-1}^{\{0\}}(Z^d)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}\right]|\xi_t^{\{0\}}\right], \tag{3.2}$$

$$E(D|\xi_t^{\{0\}}) = E\left[\sum_{1\le i<U+D}\left[\frac{\mu n_{i-1}^{\{0\}}(Z^d)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}\right]|\xi_t^{\{0\}}\right]. \tag{3.3}$$

Adding these two equations yields the identity $E(U+D|\xi_t^{\{0\}}) = E(U+D|\xi_t^{\{0\}})$, so (3.2) and (3.3) are equivalent to the difference

$$E(U-D|\xi_t^{\{0\}}) = E\left[\sum_{1\le i<U+D}\left[\frac{\lambda k_{i-1}^{\{0\}}(Z^d) - \mu n_{i-1}^{\{0\}}(Z^d)}{\lambda k_{i-1}^{\{0\}}(Z^d) + \mu n_{i-1}^{\{0\}}(Z^d)}\right]|\xi_t^{\{0\}}\right],$$

and since $U - D = n_t^{\{0\}}(Z^d) - 1$, this reduces to

$$E\left[\sum_{1\le i<U+D}\left[\frac{\lambda k_{i-1}^{\{0\}}(Z^d)-\mu n_{i-1}^{\{0\}}(Z^d)}{\lambda k_{i-1}^{\{0\}}(Z^d)+\mu n_{i-1}^{\{0\}}(Z^d)}\right]|\xi_t^{\{0\}}\right]=n_t^{\{0\}}(Z^d)-1. \tag{3.4}$$

Even though this last step removes the dependence on the conditional expectation of $U-D$, this is no great help since the conditional behavior of $U+D$ still enters through the range of the summation.

Thus, as expected the maximum likelihood estimator of $\rho=\lambda/\mu$ presumably depends on both $n_t^{\{0\}}(Z^d)/k_t^{\{0\}}(Z^d)$ and the conditional behavior of U given $\xi_t^{\{0\}}$. Obviously there are two different possibilities to study the maximum likelihood estimator, namely asymptotic approximation of the estimator and simulation. Work in the former direction is in progress.

References

[1] Fiocco, M. (1997). *Statistical Estimation for the Supercritical Contact Process.* PhD thesis, University of Leiden, The Netherlands.

[2] Fiocco, M. and van Zwet, W. R. (2003a). Decaying correlations for the supercritical contact process conditioned on survival. *Bernoulli*, **9**(5), 763–781.

MR2047685

[3] Fiocco, M. and van Zwet, W. R. (2003b). Parameter estimation for the supercritical contact process *Bernoulli*, **9**(6), 1071–1092. MR2046818

[4] Liggett, T. (1985). *Interacting Particle Systems.* Springer-Verlag, New York. MR776231

[5] Liggett, T. (1999). *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes.* Springer-Verlag, New York. MR1717346

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 319–331
© Institute of Mathematical Statistics, 2004

On the "Poisson boundaries" of the family of weighted Kolmogorov statistics

Leah Jager[1] **and Jon A. Wellner**[1]

University of Washington

Abstract: Berk and Jones (1979) introduced a goodness of fit test statistic R_n which is the supremum of pointwise likelihood ratio tests for testing $H_0 : F(x) = F_0(x)$ versus $H_1 : F(x) \neq F_0(x)$. They showed that their statistic does not always converge almost surely to a constant under alternatives F, and, in fact that there exists an alternative distribution function F such $R_n \to_d \sup_{t>0} \mathbb{N}(t)/t$ where $\mathbb{N}$ is a standard Poisson process on $[0, \infty)$. We call the particular distribution function F which leads to this limiting Poisson behavior the *Poisson boundary distribution function for* R_n. We investigate Poisson boundaries for weighted Kolmogorov statistics $D_n(\psi)$ for various weight functions ψ and comment briefly on the history of results concerning Bahadur efficiency of these statistics. One result of note is that the logarithmically weighted Kolmogorov statistic of Groeneboom and Shorack (1981) has the same Poisson boundary as the statistic of Berk and Jones (1979).

1. Introduction

Suppose that $X_1, \ldots, X_n$ are i.i.d. F on $\mathbb{R}$ and we want to test the null hypothesis

$$H : F(x) = F_0(x) \qquad \text{for all} \quad x \in \mathbb{R}$$

where F_0 is continuous, versus the alternative hypothesis

$$K : F(x) \neq F_0(x) \qquad \text{for some} \quad x \in \mathbb{R}.$$

As usual, we can reduce to the case when F_0 is the Uniform$(0, 1)$ distribution on $[0, 1]$; i.e. $F_0(x) = x$ for $0 \leq x \leq 1$.

Berk and Jones (1979) introduced the test statistic R_n, which is defined as

$$R_n = \sup_{-\infty<x<\infty} K\big(\mathbb{F}_n(x), F_0(x)\big), \tag{1.1}$$

where

$$K(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}, \tag{1.2}$$

and $\mathbb{F}_n$ is the empirical distribution functions of the X_i's, given by

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \leq x]}. \tag{1.3}$$

[1]University of Washington, Statistics, Box 354322, Seattle, WA 98195-4322, USA. e-mail: leah@stat.washington.edu; jaw@stat.washington.edu

Keywords and phrases: Bahadur efficiency, Berk–Jones statistic, consistency, fixed alternatives, goodness of fit, Kolmogorov statistic, Poisson process, power, weighted Kolmogorov statistic.

AMS 2000 subject classifications: primary 60G15, 60G99; secondary 60E05.

Define

$$K^+(x,y) = \begin{cases} K(x,y), & 0 < y < x < 1, \\ 0, & 0 \le x \le y \le 1, \\ \infty, & \text{otherwise} \end{cases}, \tag{1.4}$$

and

$$K^-(x,y) = \begin{cases} K(x,y), & 0 < x < y < 1, \\ 0, & 0 \le y \le x \le 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Berk and Jones also studied the one-sided statistics R_n^+ and R_n^- defined by

$$R_n^+ = \sup_x K^+\big(\mathbb{F}_n(x), x\big), \qquad R_n^- = \sup_x K^-\big(\mathbb{F}_n(x), x\big).$$

Berk and Jones (1979) discussed the optimality properties of the statistics R_n^+ and R_n. They showed, in particular, that they have greater Bahadur efficiency than the corresponding Kolmogorov statistics. Berk and Jones (1979) also extended this comparison to weighted Kolmogorov statistics via the results of Abrahamson (1967). In view of the results of Groeneboom and Shorack (1981), these comparisons are trivial for any weight funtion ψ of the form $\psi(x) = [x(1-x)]^{-b}$ for any positive b since Groeneboom and Shorack show that the limiting efficacy of the weighted Kolmogorov statistics with power function weighting is in fact zero for any alternative for which the efficacy makes sense. Moreover, as we show here the efficacies of the weighted Kolmogorov statistics are not well-defined (and the Bahadur efficiency comparison is not meaningful) for fixed alternatives at or beyond certain "Poisson boundaries" which we describe below. Thus it seems to us that the assertion by Owen (1995), at the end of his section 1, that the statistics of Berk and Jones (1979) have "increased efficiency over any weighted Kolmogorov–Smirnov method at any alternative distribution" is an over-interpretation of the results of Berk and Jones (1979).

Wellner and Koltchinskii (2003) present a proof of the limiting null distribution of the Berk-Jones statistic, and Owen (1995) computes exact quantiles under the null distribution for finite n; see also Owen (2001). Using these quantiles, Owen constructed confidence bands for F by inverting the Berk and Jones test, and then calculates the power associated with the Berk-Jones test statistic for fixed alternatives of the form $F(x) = F_0(x)^\alpha$. See Jager and Wellner (2004) for some corrections of the results of Owen (1995).

One of the interesting results for the statistic R_n proved in Berk and Jones (1979) is the following limit behavior under a rather extreme alternative distribution.

Theorem 1 (Berk and Jones (1979)). *Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F given by*

$$F(x) = \frac{1}{1 + \log(1/x)}, \qquad 0 < x < 1 \qquad \textit{and} \qquad 0 < b < 1. \tag{1.5}$$

Then

$$R_n^+ \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} \frac{1}{U},$$

$$R_n \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} \frac{1}{U}$$

where $\mathbb{N}$ *is a standard Poisson process on* $[0,\infty)$ *and* U *is a Uniform*$[0,1]$ *random variable.*

Because of the Poisson nature of the limiting distribution in Theorem 1, we call the corresponding alternative distribution function F a "Poisson boundary" for the test statistic R_n. The fact that $\sup_{t>0} \mathbb{N}(t)/t \stackrel{d}{=} 1/U$ follows from results of Pyke (1959), page 571, and elementary manipulations, or, alternatively from the classical result of Daniels (1945) that

$$P\left(\sup_{0<t\le 1} \mathbb{G}_n(t)/t \ge x\right) = 1/x \qquad \text{for} \qquad x \ge 1$$

where $\mathbb{G}_n$ is the empirical distribution function of n i.i.d. Uniform$(0,1)$ random variables (see e.g. Shorack and Wellner (1986), page 404) together with the Poisson convergence results of Wellner (1977b).

For alternatives F that are "less extreme" than the F given in Theorem 1, Berk and Jones (1979) give sufficient conditions under which following more usual or "expected" behavior holds:

$$R_n^+ \underset{a.s.}{\to} \sup_x K^+\big(F(x),x\big), \qquad \text{and} \qquad R_n \underset{a.s.}{\to} \sup_x K\big(F(x),x\big).$$

Some questions related to this type of result are discussed further in Section 4.

Our main purpose here is to note that the phenomena of a Poisson boundary is not unique to the Berk–Jones statistic R_n, but that in fact this type of behavior holds for a general class of "weighted" type statistics. Indeed we will show that the Poisson boundary for the weighted Kolmogorov statistics is a much less extreme alternative than the Poisson boundary distribution function F (given in (1.5)) found by Berk and Jones (1979) for their statistic.

2. "Poisson boundaries" for weighted Kolmogorov statistics

Consider the family of weighted Kolmogorov–Smirnov statistics given by

$$D_n(b) \equiv \sup_{0<x<1} \frac{|\mathbb{F}_n(x)-x|}{(x(1-x))^b} \tag{2.6}$$

where $\mathbb{F}_n$ is the empirical distribution function of the X_i's and $0 < b < 1$. The asymptotic behavior of $D_n(b)$ under the null hypothesis H is well-known: for $0 < b < 1/2$

$$n^{1/2} D_n(b) \underset{d}{\to} \sup_{0<t<1} \frac{|\mathbb{U}(t)|}{(t(1-t))^b}$$

where $\mathbb{U}$ is a standard Brownian bridge process, while for $1/2 < b \le 1$

$$n^{1-b} D_n(b) \underset{d}{\to} \max\left\{ \sup_{0<t<\infty} \frac{|\mathbb{N}(t)-t|}{t^b}, \sup_{0<t<\infty} \frac{|\tilde{\mathbb{N}}(t)-t|}{t^b} \right\}$$

where $\mathbb{N}$ and $\tilde{\mathbb{N}}$ are independent standard Poisson processes. The case $0 < b < 1/2$ follows from Chibisov (1964) and O'Reilly (1974); see e.g. Shorack and Wellner (1986), pages 461–466, or Csörgő and Horváth (1993), Theorem 3.2, page 217. The case $1/2 < b < 1$ follows from Mason (1983); see also Csörgő and Horváth (1993), Theorem 1.2, page 265. When $b = 1/2$ the limit behavior is due to Jaeschke (1979)

and Eicker (1979), which in turn rely on the classical results of Darling and Erdös (1956):

$$b_n n^{1/2} D_n(b) - c_n \underset{d}{\to} E_v^4$$

where $b_n = (2 \log\log n)^{1/2}$, $c_n = 2\log\log n + (1/2)\log\log\log n - (1/2)\log(4\pi)$, and $P(E_v^4 \le x) = \exp(-4e^{-x})$; see e.g. Shorack and Wellner (1986), page 600.

Our goal here is to prove the following theorems concerning particular fixed alternative hypotheses.

Theorem 2. *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. F where $F(x) = x^b$ for $0 \le x \le 1$. Then*

$$D_n(b) \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} \frac{1}{U} \tag{2.7}$$

where $U \sim$ Uniform$(0,1)$.

Theorem 2 does not cover the interesting special case $b = 1$. For $b = 1$ we have the following (more special) result.

Theorem 2A. *Suppose that $c > 1$ and that $X_1, X_2, \ldots, X_n$ are i.i.d. F where*

$$F(x) = \begin{cases} 0, & -\infty < x < 0, \\ cx, & 0 \le x \le 1/c, \\ 1, & 1/c \le x < \infty. \end{cases}$$

Then

$$D_n(1) \underset{d}{\to} \left(c \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} - 1 \right) \bigvee c \overset{d}{=} \left(c\frac{1}{U} - 1 \right) \bigvee c \equiv Y_c$$

where $U \sim$ Uniform$(0,1)$ and

$$P(Y_c \le x) = \begin{cases} 0, & x < c, \\ 1 - c/(x+1), & x \ge c. \end{cases} \tag{2.8}$$

Theorems 2 and 2A do not cover the case of (very light) logarithmic weights which are of interest because of their connection to the results of Groeneboom and Shorack 1981. These authors showed that with $\psi = \psi_2$ where $\psi_2(x) \equiv -\log(x(1-x))$, the ψ-weighted Kolmogorov statistics

$$D_n(\psi) \equiv \sup_{0<x<1} |\mathbb{F}_n(x) - F(x)|\psi(x), \qquad D_n^+(\psi) \equiv \sup_{0<x<1} (\mathbb{F}_n(x) - F(x))\psi(x) \tag{2.9}$$

have non-trivial large deviation behavior under the null hypothesis and hence have non-trivial Bahadur slopes as long as

$$D_n(\psi) \to_{a.s.} d(\psi, F), \qquad D_n^+(\psi) \to_{a.s.} d^+(\psi, F) \tag{2.10}$$

respectively under the alternative hypothesis F. Thus it is of interest to determine under what conditions (for what F's) (2.10) holds. A step in this direction is to find the Poisson boundary for $D_n(\psi_2)$. As it turns out, $D_n(\psi_2)$ has the same Poisson boundary distribution function as the Berk-Jones statistic R_n.

Theorem 2B. *Let F be the distribution function given by* (1.5)*. If $X_1, \ldots, X_n$ are i.i.d. F, then*

$$D_n^+(\psi_2) \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} \frac{1}{U},$$
$$D_n(\psi_2) \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} \frac{1}{U}$$

where $\mathbb{N}$ is a standard Poisson process and $U \sim$ Uniform$(0,1)$.

An alternative test statistic, $\tilde{R}_n$, which we have called the *reversed Berk–Jones statistic* in Jager and Wellner (2004), is defined by

$$\tilde{R}_n = \sup_{X_{(1)} \le x < X_{(n)}} K\big(F_0(x), \mathbb{F}_n(x)\big) \tag{2.11}$$

where $X_{(1)}$ and $X_{(n)}$ are the first and last order statistics, respectively.

The motivation behind this statistic comes from examination of the functions $K(F_0(x), F(x))$ and $K(F(x), F_0(x))$, for an alternative distribution function F. When F is stochastically smaller than F_0, we expect the Berk-Jones test to be more powerful than the reversed Berk-Jones statistic, since $\sup_x K(F(x), F_0(x)) > \sup_x K(F_0(x), F(x))$ in this case. However, in the case where F is stochastically larger than F_0, we have $\sup_x K(F(x), F_0(x)) < \sup_x K(F_0(x), F(x))$, and so we expect the reversed statistic to be more powerful.

We do not yet know if $\tilde{R}_n$ has a "Poisson boundary". The question is: does there exist an alternative distribution function F such that when sampling from F we have

$$\tilde{R}_n \underset{d}{\to} g(\mathbb{N})$$

for some functional g of a (standard?) Poisson process $\mathbb{N}$?

Before giving the proofs we state two results that will be used repeatedly in the proofs: the weighted Glivenko–Cantelli theorem of Lai (1974) (see also Wellner (1977a) and Shorack and Wellner (1986), page 410), and bounds for the sup of ratios given by Wellner (1978) and Berk and Jones (1979) (see also Shorack and Wellner (1986), Inequality 10.3.2, pages 415 and 416) that will be used several times in the proofs. Let $\mathbb{G}_n(t) = n^{-1} \sum_{i=1}^n 1_{[0,t]}(\xi_i)$ where $\xi_1, \ldots, \xi_n, \ldots$ are i.i.d. Uniform$(0,1)$ random variables, and let I be the identity function on $[0,1]$.

Theorem W-GC (Lai (1974); Wellner (1977a)). *Suppose that ψ is positive on $(0,1)$, decreasing on $(0,1/2]$, and symmetric about $1/2$. Then*

$$\limsup_{n\to\infty} \|(\mathbb{G}_n - I)\psi\| = \begin{cases} 0 & a.s. \\ \infty & a.s. \end{cases} \quad \text{according as} \quad \int_0^1 \psi(t)dt \begin{cases} < \infty \\ = \infty. \end{cases}$$

Theorem (Ratio bounds). (Wellner (1978), Berk and Jones (1979)). *For all $x \ge 1$ and $0 < \epsilon \le 1$*

$$P\left(\sup_{\epsilon \le t \le 1} \frac{\mathbb{G}_n(t)}{t} \ge x\right) \le \begin{cases} \exp(-n\epsilon h(x)) \\ \exp(-nK^+(\epsilon x, \epsilon)) \end{cases} \tag{2.12}$$

and

$$P\left(\sup_{\epsilon \le t \le 1} \frac{t}{\mathbb{G}_n(t)} \ge x\right) \le \begin{cases} \exp(-n\epsilon h(1/x)) \\ \exp(-nK^+(1-\epsilon/x, 1-\epsilon)) \end{cases} \tag{2.13}$$

where $h(x) \equiv x(\log x - 1) + 1$ and where K^+ is as defined in (1.4).

Now we provide proofs for Theorems 2, 2A, and 2B.

Proof of Theorem 2. Let $0 < \alpha < 1$. We write

$$\begin{aligned} D_n(b) &= \sup_{0<x<1} \frac{|\mathbb{F}_n(x) - x|}{(x(1-x))^b} \\ &= \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{(x(1-x))^b} \bigvee \sup_{x:F(x)\geq n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{(x(1-x))^b} \\ &= \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)(1-F(x)^{1/b})^b} \bigvee \sup_{x:F(x)\geq n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)(1-F(x)^{1/b})^b} \\ &\equiv D_n^{(1)}(b) \bigvee D_n^{(2)}(b). \end{aligned}$$

Now

$$\begin{aligned} D_n^{(1)}(b) &- \sup_{x:F(x)<n^{-\alpha}\frac{\mathbb{F}_n(x)}{F(x)}} \\ &= \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)(1-F(x)^{1/b})^b} - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\ &= \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x) - x}{F(x)(1-F(x)^{1/b})^b} \bigvee \sup_{x:F(x)<n^{-\alpha}} \frac{x - \mathbb{F}_n(x)}{F(x)(1-F(x)^{1/b})^b} \\ &\quad - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\ &\leq \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)(1-F(x)^{1/b})^b} - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\ &\quad + \sup_{x:F(x)<n^{-\alpha}} \frac{x}{F(x)(1-F(x)^{1/b})^b} \\ &\leq \left| \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)(1-F(x)^{1/b})^b} - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \right| \\ &\quad + 2 \sup_{x:\, x<n^{-\alpha/b}} \frac{x}{x^b(1-x)^b} \\ &\leq \sup_{x:F(x)<n^{-\alpha}} \left| \frac{\mathbb{F}_n(x)}{F(x)(1-F(x)^{1/b})^b} - \frac{\mathbb{F}_n(x)}{F(x)} \right| + o(1) \\ &\leq \sup_{x:F(x)<n^{-\alpha}} \left| \frac{\mathbb{F}_n(x)}{F(x)} \left(\frac{1}{(1-x)^b} - 1 \right) \right| + o(1) \\ &\leq \sup_{x:F(x)<n^{-\alpha}} \left| \frac{\mathbb{F}_n(x)}{F(x)} \right| \sup_{x:F(x)<n^{-\alpha}} \left| \left(\frac{1}{(1-x)^b} - 1 \right) \right| + o(1) \\ &\leq O_p(1)o(1) + o(1) = o_p(1). \end{aligned}$$

On the other hand,

$$\begin{aligned} &\sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} - D_n^{(1)}(b) \\ &\quad = \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} - \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)(1-F(x)^{1/b})^b} \\ &\quad \leq \sup_{x:F(x)<n^{-\alpha}} \frac{x}{x^b(1-x)^b} = o(1) \end{aligned}$$

since

$$\begin{aligned}
&\sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x)-x|}{F(x)(1-F(x)^{1/b})^b} \\
&\quad\geq \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)-x}{F(x)(1-F(x)^{1/b})^b} \\
&\quad\geq \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)(1-F(x)^{1/b})^b} - \sup_{x:F(x)<n^{-\alpha}} \frac{x}{x^b(1-x)^b} \\
&\quad\geq \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} - o(1).
\end{aligned}$$

Concerning $D_n^{(2)}(b)$ we have

$$\begin{aligned}
D_n^{(2)}(b) &= \sup_{x:F(x)\geq n^{-\alpha}} \frac{|\mathbb{F}_n(x)-x|}{F(x)(1-F(x)^{1/b})^b} \\
&\leq \sup_{x:F(x)\geq n^{-\alpha}} \frac{|\mathbb{F}_n(x)-F(x)|}{F(x)(1-F(x)^{1/b})^b} \\
&\qquad + \sup_{x:F(x)\geq n^{-\alpha}} \frac{|F(x)-x|}{F(x)(1-F(x)^{1/b})^b} \\
&\leq \sup_{x:n^{-\alpha}\leq F(x)\leq 1/2} \frac{|\mathbb{F}_n(x)-F(x)|}{F(x)(1-F(x)^{1/b})^b} \\
&\qquad + \sup_{x:1/2\leq F(x)<1} \frac{|\mathbb{F}_n(x)-F(x)|}{F(x)(1-F(x)^{1/b})^b} + 1 \\
&\leq \frac{1}{(1-(1/2)^{1/b})^b} \sup_{x:n^{-\alpha}\leq F(x)\leq 1/2} \frac{|\mathbb{F}_n(x)-F(x)|}{F(x)} \\
&\qquad + 2 \sup_{x:1/2\leq F(x)<1} \frac{|\mathbb{F}_n(x)-F(x)|}{(1-F(x)^{1/b})^b} + 1 \\
&= o(1)+o(1)+1
\end{aligned}$$

almost surely by Lemma 4.3 of Berk and Jones (1979) for the first term, and by the weighted Glivenko–Cantelli Theorem W-GC for the second term since

$$\int_0^1 \frac{1}{(1-x^{1/b})^b}\,dx = \int_0^1 (1-u)^{-b} b u^{b-1}\,du = b\Gamma(1-b)\Gamma(b) < \infty$$

for $b \in (0,1)$. Hence it follows that $\limsup_{n\to\infty} D_n^{(2)}(b) \leq 1$ almost surely. Putting all this together with the fact that

$$\sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \underset{d}{\to} \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} \overset{d}{=} 1/U$$

finishes the proof of Theorem 2. □

Proof of Theorem 2A. Since $\mathbb{F}_n \overset{d}{=} \mathbb{G}_n(F)$ where $\mathbb{G}_n$ is the empirical distribution function of i.i.d. Uniform$(0,1)$ random variables $\xi_1,\ldots,\xi_n$, we can write

$$
\begin{aligned}
D_n(1) &\stackrel{d}{=} \sup_{0<x<1} \frac{|\mathbb{G}_n(F(x)) - x|}{x(1-x)} \\
&= \sup_{0<x\le 1/c} \frac{|\mathbb{G}_n(cx) - x|}{x(1-x)} \bigvee \sup_{1/c<x\le 1} \frac{|1-x|}{x(1-x)} \\
&= \sup_{0<t\le n} \frac{|n\mathbb{G}_n(t/n) - t/c|}{(t/c)(1-t/(cn))} \bigvee c \\
&\underset{d}{\to} \sup_{0<t<\infty} \frac{|\mathbb{N}(t) - t/c|}{t/c} \bigvee c \\
&= \left(c \sup_{0<t<\infty} \frac{\mathbb{N}(t)}{t} - 1 \right) \bigvee 1 \bigvee c \\
&\stackrel{d}{=} \left(c\frac{1}{U} - 1 \right) \bigvee c \equiv Y_c.
\end{aligned}
$$

since $c > 1$ and since the process $\{n\mathbb{G}_n(t/n) : 0 < t \le n\}$ converges weakly to the standard Poisson process $\mathbb{N}$ in a topology that makes the weighted supremum functional in the last display continuous; see e.g. Wellner (1977b), Theorem 7, page 1007. Computation of the distribution of Y_c is straightforward. (Note that this distribution has a jump at c of height $1/(1+c)$.) □

Proof of Theorem 2B. Let $0 < \alpha < 1$. We write

$$
\begin{aligned}
D_n(\psi_2) &= \sup_{0<x<1} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \\
&= \sup_{x:F(x)<n^{-\alpha}} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \bigvee \sup_{x:F(x)\ge n^{-\alpha}} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \\
&= \sup_{x:F(x)<n^{-\alpha}} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \bigvee \sup_{x:F(x)\ge n^{-\alpha}} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \\
&\equiv D_n^{(1)}(\psi_2) \bigvee D_n^{(2)}(\psi_2).
\end{aligned}
$$

We first deal with $D_n^{(2)}(\psi_2)$. Note that

$$
\begin{aligned}
D_n^{(2)}(\psi_2) &= \sup_{x:F(x)\ge n^{-\alpha}} \big|\mathbb{F}_n(x) - x\big|\psi_2(x) \\
&\le \sup_{x:F(x)\ge n^{-\alpha}} \big|\mathbb{F}_n(x) - F(x)\big|\psi_2(x) \\
&\qquad + \sup_{x:F(x)\ge n^{-\alpha}} \big|F(x) - x\big|\psi_2(x) \\
&\le \sup_{x:n^{-\alpha}\le F(x)\le 1/2} \frac{|\mathbb{F}_n(x) - F(x)|}{F(x)} F(x)\psi_2(x) \\
&\qquad + \sup_{x:1/2\le F(x)<1} \frac{|\mathbb{F}_n(x) - F(x)|}{(1-F(x))^{3/4}} (1-F(x))^{3/4}\psi_2(x) + 1 \\
&\le \sup_{x:n^{-\alpha}\le F(x)\le 1/2} \frac{|\mathbb{F}_n(x) - F(x)|}{F(x)} \\
&\qquad + \sup_{x:1/2\le F(x)<1} \frac{|\mathbb{F}_n(x) - F(x)|}{(1-F(x))^{3/4}} + 1 \\
&= o(1) + o(1) + 1
\end{aligned}
$$

almost surely by Lemma 4.3 of Berk and Jones (1979) or Wellner (1978) for the first term, and Theorem W-GC for the second term. Here we also used $\psi_2(x)F(x) \le 1$ for $0 < x \le 1/2$, and $(1 - F(x))^{3/4}\psi_2(x) \le 1$ for $1/2 \le x < 1$.

To handle $D_n^{(1)}(\psi_2)$, note that

$$
\begin{aligned}
&D_n^{(1)}(\psi_2) - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\
&\quad = \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)} F(x)\psi_2(x) - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\
&\quad = \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x) - x}{F(x)} F(x)\psi_2(x) \bigvee \sup_{x:F(x)<n^{-\alpha}} \frac{x - \mathbb{F}_n(x)}{F(x)} F(x)\psi_2(x) \\
&\qquad - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \\
&\quad \le \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} F(x)\psi_2(x) - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} + \sup_{x:F(x)<n^{-\alpha}} x\psi_2(x) \\
&\quad \le \left| \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} F(x)\psi_2(x) - \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} \right| + o(1) \\
&\quad \le \sup_{x:F(x)<n^{-\alpha}} \left| \frac{\mathbb{F}_n(x)}{F(x)} \left(F(x)\psi_2(x) - 1\right) \right| + o(1) \\
&\quad \le \sup_{x:F(x)<n^{-\alpha}} \left| \frac{\mathbb{F}_n(x)}{F(x)} \right| \sup_{x:F(x)<n^{-\alpha}} \left| F(x)\psi_2(x) - 1 \right| + o(1) \\
&\quad \le O_p(1)o(1) + o(1) = o_p(1).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} - D_n^{(1)}(\psi_2) \\
&\quad = \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} - \sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)} F(x)\psi_2(x) \\
&\quad \le \sup_{x:F(x)<n^{-\alpha}} x\psi(x) = o(1)
\end{aligned}
$$

since

$$
\begin{aligned}
&\sup_{x:F(x)<n^{-\alpha}} \frac{|\mathbb{F}_n(x) - x|}{F(x)} F(x)\psi_2(x) \\
&\quad \ge \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x) - x}{F(x)} F(x)\psi_2(x) \\
&\quad \ge \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} F(x)\psi_2(x) - \sup_{x:F(x)<n^{-\alpha}} x\psi_2(x) \\
&\quad \ge \sup_{x:F(x)<n^{-\alpha}} \frac{\mathbb{F}_n(x)}{F(x)} (1 - o(1)) - o(1).
\end{aligned}
$$

Combining these pieces as in the proof of Theorem 2 completes the proof for $D_n(\psi_2)$. The proof for $D_n^+(\psi_2)$ is similar (and easier). □

3. A consistency result

Theorems 2, 2A, 2B suggest that we might expect classical behavior for the weighted Kolmogorov statistics under fixed alternatives F sufficiently "inside" their respective Poisson boundaries. Here are two of the expected consistency results. They are, in fact, corollaries the weighted Glivenko–Cantelli Theorem W-GC in Section 2, or of general Glivenko–Cantelli theory (see e.g. Dudley (1999) or Vaart and Wellner 1996).

Theorem 3. *Suppose that $X_1, X_2, \ldots$ are i.i.d. F on $[0,1]$ and $0 < b < 1$.*
(i) *If $E[(X(1-X))^{-b}] < \infty$, then*

$$D_n(b) \equiv \sup_{0<x<1} \frac{|\mathbb{F}_n(x) - x|}{(x(1-x))^b} \to_{a.s.} \sup_{0<x<1} \frac{|F(x) - x|}{(x(1-x))^b} \equiv d(b, F) < \infty.$$

(ii) *If $E[(X(1-X))^{-b}] = \infty$, then $\limsup_{n\to\infty} D_n(b) = +\infty$ a.s.*

Theorem 3B. *Suppose that $X_1, X_2, \ldots$ are i.i.d. F on $[0,1]$ and $\psi_2(x) \equiv -\log(x(1-x))$.*
(i) *If $E[\psi_2(X)] < \infty$, then*

$$D_n(\psi) \equiv \sup_{0<x<1} |\mathbb{F}_n(x) - x|\psi_2(x) \to_{a.s.} \sup_{0<x<1} |F(x) - x|\psi_2(x) \equiv d(\psi_2, F) < \infty.$$

(ii) *If $E[\psi_2(X)] = \infty$ then $\limsup_{n\to\infty} D_n(\psi_2) = +\infty$ almost surely.*

Proof of Theorem 3. Note that

$$\begin{aligned} |D_n(b) - d(b,F)| &\le \sup_{0<x<1} \frac{|\mathbb{F}_n(x) - F(x)|}{(x(1-x))^b} \\ &= \sup_{0<x<1} \frac{|\mathbb{G}_n(F(x)) - F(x)|}{(x(1-x))^b} \\ &= \sup_{0<u<1} \frac{|\mathbb{G}_n(u) - u|}{(F^{-1}(u)(1-F^{-1}(u)))^b} \\ &\underset{a.s.}{\to} 0 \end{aligned}$$

if

$$\int_0^1 \frac{1}{(F^{-1}(u)(1-F^{-1}(u)))^b}\, du < \infty \tag{3.14}$$

by Theorem W-GC, or by part A, of Wellner (1977a) and remark 1 on page 475. But (3.14) holds if and only if the stated hypothesis holds by the fact that $F^{-1}(U) \overset{d}{=} X \sim F$ for $U \sim U(0,1)$. □

Remark 1. Note that for the "Poisson boundary" distribution function $F(x) = x^b$ for $D_n(b)$

$$E[(X(1-X))^{-b}] = \int_0^1 \frac{bx^{b-1}}{(x(1-x))^b}\, dx = b\int_0^1 \frac{1}{x(1-x)^b}\, dx = \infty,$$

so the hypothesis of Theorem 3 part (i) (just) fails. On the other hand, if $F(x) = x^c$ with $b < c < 1$, then

$$E[(X(1-X))^{-b}] = \int_0^1 \frac{cx^{c-1}}{(x(1-x))^b}\, dx = c\int_0^1 \frac{1}{x^{1+b-c}(1-x)^b}\, dx < \infty,$$

so the hypothesis of Theorem 3(i) holds and $D_n(b) \to_{a.s.} d(b,F)$.

Remark 2. Note that for the "Poisson boundary" distribution function $F(x) = (1 + \log(1/x))^{-1}$ for the statistic $D_n(\psi_2)$,

$$E_F[\psi_2(X)] = \int_0^1 \log\left(\frac{1}{x(1-x)}\right) \frac{1}{x(1+\log(1/x))^2}\, dx = \infty$$

so the hypothesis of Theorem 3B part (i) (just) fails.

4. Further problems

Here is a partial list of open problems in connection with the statistics discussed here and in Jager and Wellner (2004).

Question 1. What are the theorems corresponding to Theorem 3 in the case of R_n and $\tilde{R}_n$? In other words, for exactly which alternative distribution functions F does it hold that

$$R_n \to_{a.s.} \sup_x K\big(F(x), F_0(x)\big) \equiv r(F, F_0)? \tag{4.15}$$

For exactly which alternative distribution functions F does it hold that

$$\tilde{R}_n \to_{a.s.} \sup_x K\big(F_0(x), F(x)\big) \equiv \tilde{r}(F, F_0)\,? \tag{4.16}$$

Question 2. For alternative distribution functions F such that (4.15) holds, can we obtain useful approximations to the power of R_n via limit theorems for

$$\sqrt{n}\big(R_n - r(F, F_0)\big)$$

along the lines of Raghavachari (1973)? Similarly for F's for which (4.16) holds for $\tilde{R}_n$?

Question 3. Donoho and Jin (2004) consider testing $H_0: F = N(0,1) = \Phi$ versus $H_1: F = (1-\epsilon)N(0,1) + \epsilon N(\mu, 1)$ where $\epsilon_n = n^{-\beta}$ and $\mu = \mu_n = \sqrt{2r\log n}$ for $\beta > 1/2$ and $r > 0$. They show that a natural "detection boundary" is given by

$$r^*(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \le 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases}$$

How do the statistics R_n, $\tilde{R}_n$, and $K_n(1/2)$ compare along the "detection boundary" of Donoho and Jin (2004) Note that Donoho and Jin (2004) find that $D_n(1/2)$ and R_n have quite comparable power behavior for their testing problem, but they show that $D_n(1/2)$ has better power in the region $r > r^*(\beta)$ and $3/4 < \beta < 1$.

Question 4. What is the limiting null distribution of $\tilde{R}_n$?

References

Abrahamson, I. (1967). Exact Bahadur efficiencies for the Kolmogorov–Smirnov and Kuiper one- and two- sample statistics. *Ann. Math. Statist.* **38**, 1475–1490. MR214192

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 47–59.

Chibisov, D. M. (1964). Some theorems on the limiting behavior of empirical distribution functions. *Selected Transl. Math. Statist. Prob.* **6**, 147–156.

Csörgő, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley, New York.

Daniels, H. E. (1945). The statistical theory of the strength of bundles of thread. *Proc. Roy. Soc. London Ser. A* **183**, 405–435. MR12388

Darling, D. A. and Erdös, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J.* **23**, 143–155. MR74712

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. Technical Report 2002-12, Department of Statistics, Stanford University. *Ann. Statist.* **32**, to appear.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical process. *Ann. Statist.* **7**, 116–138. MR515688

Groeneboom, P. and Shorack, G. R. (1981). Large deviations of goodness of fit statistics and linear combinations of order statistics. *Ann. Probab.* **9**, 971–987. MR632970

Jaeschke, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distibution function on subintervals. *Ann. Statist.* **7**, 108–115. MR515687

Jager, L. and Wellner, J. A. (2004). A new goodness of fit test: the reversed Berk-Jones statistic. *Technical Report* **443**, Department of Statistics, University of Washington. *http://www.stat.washington.edu/www/research/reports/2004/tr443.ps*.

Lai, T. L. (1974). Convergence rates in the strong law of large number for random variables taking values in Banach spaces. *Bull. Inst. Math. Acad. Sinica* **2**, 67–85. MR358901

Mason, David M. (1983). The asymptotic distribution of weighted empirical distribution functions. *Stochastic Process. Appl.* **15**, 99–109. MR694539

Mason, D. M. and Schuenemeyer, J. H. (1983). A modified Kolmogorov–Smirnov test sensitive to tail alternatives. *Ann. Statist.* **11**, 933–946. MR707943

Noé, M. (1972). The calculation of distributions of two-sided Kolmogorov–Smirnov type statistics. *Annals of Mathematical Statistics* **43**, 58–64. MR300379

O'Reilly, N. (1974). On the weak convergence of empirical processes in sup-norm metrics. *Ann. Probabilty* **2**, 642–651. MR383486

Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association* **90**, 516–521. MR1340504

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton.

Pyke, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Statist.* **30**, 568–576. MR107315

Raghavachari, M. (1973). Limiting distributions of Kolmogorov–Smirnov statistics under the alternative. *Ann. Statist* **1**, 67–73. MR346976

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes*. Wiley, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wellner, J. A. (1977a). A Glivenko-Cantelli theorem and strong laws of large numbers for functions of order statistics. *Ann. Statist.* **5**, 473–480. MR651528

Wellner, J. A. (1977b). Distributions related to linear bounds for the empirical distribution function. *Ann. Statist.* **5**, 1003–1016. MR458673

Wellner, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. verw. Geb.* **45**, 73–88. MR651392

Wellner, J. A. and Koltchinskii, V. (2003). A note on the asymptotic distribution of Berk-Jones type statistics under the null hypothesis. *High Dimensional Probability III*, 321–332. Birkhäuser, Basel (2003). MR2033896

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 332–336
© Institute of Mathematical Statistics, 2004

A theorem on compatibility of systems of sets with applications

A. Goswami [1] **and B. V. Rao** [1]

Indian Statistical Institute

Abstract: A general theorem on compatibility of two systems of subsets of a separable metric space is proved. This theorem is used to deduce results about points of continuity of functions, filtrations and operator semigroups among other things.

In this paper we prove the following result which, in spirit, is very much similar to a classical 1908 result of W. H. Young [2, page 304] on real functions and unifies several well-known results.

Theorem. *Let X be a separable metric space and I be any uncountable subset of the real line. Suppose that $\{A_t : t \in I\}$ and $\{B_t : t \in I\}$ are two families of subsets of X with each A_t closed and satisfying the following condition:*

(*) for each $t \in I$, there is a $\delta_t > 0$ such that $A_t \supset B_s$ whenever $s \in (t, t+\delta_t) \cap I$.

Then for all but countably many $t \in I$, $A_t \supset B_t$. The same conclusion holds if, in condition (*), $(t, t+\delta_t)$ is replaced by $(t-\delta_t, t)$.

Proof. Set $I_\delta = \{t \in I : A_t \supset B_s$ whenever $s \in (t, t+\delta) \cap I\}$ and let ρ denote the metric on the space X. If the conclusion were false, then there is some $\delta > 0$ and an uncountable set $S \subset I_\delta$ such that for all $t \in S$ the assertion fails. Since each A_t is closed, we can get $\epsilon > 0$ such that for uncountably many $t \in S$ there exists $x_t \in B_t$ such that $\rho(x_t, A_t) > \epsilon$. Cutting down S, if necessary, we can and shall assume that this holds for all points t in S. Again no loss to assume that S is contained in an interval of length smaller than δ. Now , if $t < t'$ are two distinct points of S, then noticing that $t' \in (t, t+\delta)$ we see that $\rho(x_t, x_{t'}) > \epsilon$. Thus $\{x_t : t \in S\}$ is an uncountable set of elements of S with any two of them separated by distance larger than ϵ, contradicting separability of X. The other part is similarly proved. □

The following propositions illustrate some applications of the theorem – perhaps there are others. In what follows, the closure of a set A is denoted by $\overline{A}$.

Proposition 1. *Let X and I be as above. Let $\{B_t : t \in I\}$ be any family of subsets of X. Then for all but countably many $t \in I$*

$$B_t \subset \bigcap_{\delta > 0} \overline{\bigcup \{B_s : s \in (t, t+\delta) \cap I\}}$$

[1] Division of Theoretical Statistics & Mathematics, Indian Statistical Institute, 203, B. T. Road, Kolkata 700 108, India. e-mail: alok@isical.ac.in; bvrao@isical.ac.in

Keywords and phrases: separable metric space, filtration, infinitesimal σ-fields, operator semigroups.

AMS 2000 subject classifications: 26A15, 60G07, 47D03.

and

$$B_t \subset \bigcap_{\delta>0} \overline{\bigcup\{B_s : s \in (t-\delta, t) \cap I\}}$$

Proof. Fix $\delta > 0$ and put $A_t = \overline{\bigcup\{B_s : s \in (t, t+\delta) \cap I\}}$ for each $t \in I$. Then the Theorem implies that $B_t \subset A_t$ for all but countably many $t \in I$. The proof is now completed by running δ through a sequence decreasing to zero. The second part follows similarly. □

Proposition 2. *Let f be any function defined on an open interval I into a separable metric space X. For each $t \in I$, define*

$$L_t^d = \bigcap_{\delta>0} \overline{f[(t-\delta, t) \cap I]} \qquad L_t = \bigcap_{\delta>0} \overline{f[(t-\delta, t] \cap I]}$$

and

$$R_t^d = \bigcap_{\delta>0} \overline{f[(t, t+\delta) \cap I]} \qquad R_t = \bigcap_{\delta>0} \overline{f[[t, t+\delta) \cap I]}$$

Then for all but countably many $t \in I$, $L_t^d = L_t = R_t^d = R_t$.

Proof. Since by definition, $L_t^d \subset L_t$ and $R_t^d \subset R_t$ for all $t \in I$, it suffices to show that $L_t \subset R_t^d$ and $R_t \subset L_t^d$ for all but countably many $t \in I$. Fixing $\delta > 0$ and putting for each $t \in I$, $A_t = \overline{f[(t, t+\delta) \cap I]}$, $B_t = L_t$, it follows from the Theorem that $B_t \subset A_t$ for all but countably many $t \in I$. Running δ through a sequence decreasing to zero, one obtains $L_t \subset R_t^d$ for all but countably many $t \in I$. The other inclusion $R_t \subset L_t^d$ is proved similarly. □

Corollary (W.H.Young [2]). *Let f be any real-valued function defined on an open interval I. For every $t \in I$, let*

$$\overline{f}(t-) = \limsup_{\substack{s \longrightarrow t \\ s<t}} f(s) \qquad \overline{f}(t+) = \limsup_{\substack{s \longrightarrow t \\ s>t}} f(s)$$

and

$$\underline{f}(t-) = \liminf_{\substack{s \longrightarrow t \\ s<t}} f(s) \qquad \underline{f}(t+) = \liminf_{\substack{s \longrightarrow t \\ s>t}} f(s)$$

Then for all but countably many $t \in I$, $\overline{f}(t-) = \overline{f}(t+)$ and $\underline{f}(t-) = \underline{f}(t+)$. In particular there is a countable set $D \subset I$ such that for $t \in I - D$, if one of the limits $\lim_{\substack{s \longrightarrow t \\ s<t}} f(s)$ or $\lim_{\substack{s \longrightarrow t \\ s>t}} f(s)$ exists, then so does the other and the two are equal.

Proof. In view of the order preserving homeomorphism $x \to \arctan x$, it suffices to consider bounded f only. To complete the proof now, one has to simply observe that $\overline{f}(t-) = \sup L_t^d$, $\overline{f}(t+) = \sup R_t^d$, $\underline{f}(t-) = \inf L_t^d$ and $\underline{f}(t+) = \inf R_t^d$ in the notation of Proposition 2. □

Remark 1. It is possible to improve the above corollary as follows:

"For any function f on an open interval I into a separable metric space X, there is a countable set $D \subset I$, such that, for $t \in I - D$, if either $\lim_{s \to t, s<t} f(s)$ or $\lim_{s \to t, s>t} f(s)$ exists in X, then f is continuous at t."

To see this, we first note that, by Proposition 2, there is a countable set $D \subset I$, such that, $L_t^d = L_t = R_t^d = R_t$, for $t \in I - D$. For such a t, existence of either of the limits stated in the proposition clearly implies that all these four sets L_t^d, L_t, R_t^d, R_t are equal to one singleton set, i.e., that f is continuous at t.

As an immediate consequence of this, we get

Corollary. *Any function $f : I \to X$ which has , at every point $t \in I$, either a left limit or a right limit, can have at most countably many points of discontinuity.*

Similar technique yields some results on differentiability properties of a real function on an interval. Let f be a real-valued function defined on an open interval I. For $x \in I$, let

$$D_{t+}^d = \bigcap_{\delta>0} \overline{\left\{ \frac{f(u)-f(v)}{u-v} : t < u < v < t+\delta \right\}}$$

and

$$D_{t-}^d = \bigcap_{\delta>0} \overline{\left\{ \frac{f(u)-f(v)}{u-v} : t-\delta < u < v < t \right\}}$$

D_{t+} and D_{t-} are defined analogously with the only exception that $t < u < v < t+\delta$ and $t-\delta < u < v < t$ are replaced by $t \le u < v < t+\delta$ and $t-\delta < u < v \le t$ respectively. It should be pointed out that the closures in the above definitions are closures in the extended real line. Using arguments similar to that of Proposition 2 we get

Proposition 3. *For all but countably many $t \in I$*

$$D_{t-}^d = D_{t-} = D_{t+} = D_{t+}^d.$$

From the definition of D_{t-}, it is clear that in case D_{t-} is a singleton then f must have a left derivative at t. Similar argument applies for D_{t+} as well. This easily yields the following

Corollary. *If $f : I \to \mathbb{R}$ is such that for all but countably many t in I, either D_{t-}^d or D_{t+}^d is a singleton then f is differentiable at all but countably many points.*

A more satisfactory result would have been to replace the hypothesis in the above corollary by the apparently weaker condition that f has a left derivative or a right derivative at all but countably many t. The main problem appears to be that f may have a left (right) derivative at a point t without D_{t-} (D_{t+}) being a singleton. But can this happen at uncountably many points t ? We do not know.

For the next few propositions, which are of interest in the context of stochastic processes, we fix the following set-up and notations. $(\Omega, \mathcal{F}, P)$ denotes a probability space where $\mathcal{F}$ is the P-completion of a countably generated σ-field. It is well known that $\mathcal{F}$ is then a polish space with the metric $\rho(A, B) = P(A \Delta B)$, provided one identifies sets A and B in $\mathcal{F}$ whenever $P(A \Delta B) = 0$. For two sub-σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, we say that $\mathcal{A}$ and $\mathcal{B}$ are equal upto P-null sets, and write $\mathcal{A} \sim \mathcal{B}$ to mean that they generate the same σ-field on augmentation by P-null sets of $\mathcal{F}$. Note that any sub-σ-field of $\mathcal{F}$, on augmentation by P-null sets, becomes a closed subset (modulo the above identification) of the separable metric space $\mathcal{F}$. We will use the

same notation for a sub-σ-field of $\mathcal{F}$ as well as the closed subset it gives rise to. In this language, $\mathcal{A} \sim \mathcal{B}$ simply means that $\mathcal{A}$ and $\mathcal{B}$ are equal as closed sets. Also for any family $\{\mathcal{F}_\alpha, \alpha \in \Lambda\}$ of σ-fields the smallest σ-field containing $\mathcal{F}_\alpha$ for all $\alpha \in \Lambda$ will be denoted by $\vee_{\alpha\in\Lambda}\mathcal{F}_\alpha$.

Proposition 4. *Let $\{\mathcal{F}_t, t \in I\}$ be a monotone non-decreasing family of sub-σ-fields of $\mathcal{F}$ where I is an open interval. For each $t \in I$, let $\mathcal{F}_{t-} = \vee_{s<t}\mathcal{F}_s$ and $\mathcal{F}_{t+} = \cap_{s>t}\mathcal{F}_s$. Then for all but countably many $t \in I$,*

$$\mathcal{F}_{t-} \sim \mathcal{F}_t \sim \mathcal{F}_{t+}$$

Proof. Take $A_t = \mathcal{F}_{t-}$ and $B_t = \mathcal{F}_{t+}$, and note that $B_s \subset A_t$ whenever $s < t$. From the Theorem one gets $B_t \subset A_t$ for all but countably many $t \in I$. The proof is now completed in view of the fact that $\mathcal{F}_{t-} \subset \mathcal{F}_t \subset \mathcal{F}_{t+}$ for all $t \in I$. □

As a consequence we have

Proposition 5. *If $\{X_t, t \geq 0\}$ is a stochastic process on $(\Omega, \mathcal{F}, P)$ and if, for each $t > 0$, $\mathcal{F}_t = \sigma\langle X_u, 0 \leq u \leq t\rangle$, then for all but countably many $t > 0$, $\mathcal{F}_{t-} \sim \mathcal{F}_t \sim \mathcal{F}_{t+}$.*

It is interesting to note that the exceptional set of t's in the above proposition, to be denoted by $D(X)$, depends only on the law of the process $\{X_t\}$; that is, for two processes $\{X_t\}$ and $\{Y_t\}$ on $(\Omega, \mathcal{F}, P)$, having the same finite dimensional distributions, $D(X) = D(Y)$. In particular, if $\{X_t, t \geq 0\}$ is a process with stationary increments, then, for any $s \geq 0$, denoting the process $\{X_{s+t} - X_s, t \geq 0\}$ by $\{Y_t\}$, one has $D(X) = D(Y)$. If moreover, the increments of X are independent, then one can show, using the above, that the complement of $D(X)$ is a right interval and, hence, has to contain $(0, \infty)$. The same argument can be used to show that $D(X)$ is actually empty. Thus we have

Proposition 6. *If $\{X_t, t \geq 0\}$ is a process on $(\Omega, \mathcal{F}, P)$ with stationary independent increments, then for all t, $\mathcal{F}_{t-} \sim \mathcal{F}_t \sim \mathcal{F}_{t+}$.*

This is what is usually known as Blumenthal's $0 - 1$ law (see for example [3]), for which the usual proof is via a right continuous modification of the process $\{X_t\}$ and the strong Markov property.

Proposition 7. *Let I be any open interval and $\{\mathcal{G}_t, t \in I\}$ any family of sub-σ-fields of $\mathcal{F}$. For each $t \in I$, define*

$$\mathcal{G}^d_{t+} = \cap_{\delta>0} \vee \{\mathcal{G}_s, t < s < t + \delta\}, \qquad \mathcal{G}_{t+} = \cap_{\delta>0} \vee \{\mathcal{G}_s, t \leq s < t + \delta\}$$
$$\mathcal{G}^d_{t-} = \cap_{\delta>0} \vee \{\mathcal{G}_s, t - \delta < s < t\}, \qquad \mathcal{G}_{t-} = \cap_{\delta>0} \vee \{\mathcal{G}_s, t - \delta < s \leq t\}$$

Then for all but countably many $t \in I$, $\mathcal{G}^d_{t-} \sim \mathcal{G}_{t-} \sim \mathcal{G}^d_{t+} \sim \mathcal{G}_{t+}$.

Proof. Fix $\delta > 0$ and take $A_t = \vee\{\mathcal{G}_s, t < s < t + \delta\}$, $B_t = \mathcal{G}_{t-}$. The Theorem implies that, for all but countably many $t \in I$, $A_t \supset B_t$. By arguments similar to those used in Proposition 2, one concludes that $\mathcal{G}_{t-} \subset \mathcal{G}^d_{t+}$ for all but countably many $t \in I$. Similarly one shows that $\mathcal{G}_{t+} \subset \mathcal{G}^d_{t-}$ for all but countably many $t \in I$. The proof is now complete in view of the inclusions $\mathcal{G}^d_{t-} \subset \mathcal{G}_{t-}$ and $\mathcal{G}^d_{t+} \subset \mathcal{G}_{t+}$. □

In particular, this gives

Proposition 8 (V. S. Borkar [1]). *For a stochastic process* $\{X_t, t \geq 0\}$ *on* $(\Omega, \mathcal{F}, P)$

$$\cap_{\delta>0}\sigma\langle X_s, t-\delta < s < t\rangle \sim \cap_{\delta>0}\sigma\langle X_s, t-\delta < s \leq t\rangle \sim$$
$$\cap_{\delta>0}\sigma\langle X_s, t < s < t+\delta\rangle \sim \cap_{\delta>0}\sigma\langle X_s, t \leq s < t+\delta\rangle$$

for all but countably many $t > 0$.

Remark 2. In an analogous manner, one gets
For a stochastic process $\{X_t, t \geq 0\}$ on $(\Omega, \mathcal{F}, P)$

$$\begin{aligned}
\cap_{\delta>0} \quad & \sigma\langle X_u - X_s,\ t-\delta < s < u < t\rangle \\
\sim \cap_{\delta>0} \quad & \sigma\langle X_u - X_s,\ t-\delta < s < u \leq t\rangle \\
\sim \cap_{\delta>0} \quad & \sigma\langle X_u - X_s,\ t < s < u < t+\delta\rangle \\
\sim \cap_{\delta>0} \quad & \sigma\langle X_u - X_s,\ t \leq s < u < t+\delta\rangle
\end{aligned}$$

for all but countably many $t > 0$.

We end this note with one more application which may have interesting consequences for Markov processes.

Proposition 9. *Let* $\{T_t, t > 0\}$ *be a semigroup of bounded linear operators on a separable Banach space* B *such that, for every* $x \in B$, $\lim_{t\to 0+} T_t x$ *exists in the strong operator topology. Then* $(T_t, t > 0)$ *is strongly continuous. Moreover, the set* $\{x \in B : T_{0+}x = x\}$ *is precisely the closed span of* $\cup_{t>0} T_t B$.

Proof. By the uniform boundedness principle, T_t are uniformly bounded for t in any bounded interval. For any $x \in B$, the map $t \to T_t x$ has, by the corollary following Remark 1, only countably many discontinuities. Separability of B and the boundedness property noted above permit us to choose one countable set of t's, outside of which the map $t \mapsto T_t x$ is continuous for all $x \in B$. The semigroup property, on the other hand, would assert that the continuity points form a right interval. The proof is complete. □

Remark 3. Without separability of X the Theorem fails. For example, put $I = \mathbb{R}$, $A_t = (t, \infty)$, and $B_t = [t, \infty)$ and let X be the real line with discrete topology. It is clear that the Theorem does not hold. However in the non-separable case the Theorem will remain true if we replace *countably many* by *at most* $\aleph$ *many* where $\aleph$ is the weight of X, that is, the least cardinality of a dense set in X. Interestingly, for finite X, the exceptional set of t's cannot have a right accumulation point of order equalling *card*(X).

References

[1] Borkar, V.S (1993): On infinitesimal σ-fields generated by Random processes. Proc. Ind. Acad. Sci. (Math.Sci) **103** p. 329–332. MR1273358

[2] Hobson, E.W. (1957): *The theory of functions of a real variable (vol 1)* Dover publications Inc., New York.

[3] Karatzas, I. and Shreve, S.E. (1988): *Brownian Motion and Stochastic Calculus* Springer Verlag, New York. MR917065

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 337–341
© Institute of Mathematical Statistics, 2004

A question of geometry and probability

Richard A. Vitale[1]

University of Connecticut

Abstract: We introduce the Aleksandrov–Fenchel inequality, apply it to a tail bound for Gaussian processes, and speculate on a further connection.

1. Introduction

Some time ago I brought a question involving geometry and probability to Herman and that led us in an interesting direction [12]. To celebrate this occasion, I am happy to bring another such question.

Recall the planar isoperimetric inequality, which says that for a convex body K of area $A(K)$ and perimeter $L(K)$

$$4\pi \cdot A(K) \leq L^2(K)\,. \tag{1}$$

Consider now a 2×2 matrix M of independent $N(0,1)$ variables and the image body MK. Inserting into (1) and taking expectations gives

$$4\pi \cdot E\left[A(MK)\right] \leq E\left[L^2(MK)\right]. \tag{2}$$

However, it is the case that the following stronger inequality holds:

$$4\pi \cdot E\left[A(MK)\right] \leq \left[EL(MK)\right]^2. \tag{3}$$

It is possible to verify (3) as a simple exercise in Gaussian determinants, but one cannot say that this approach gives a satisfying explanation of what is really going on, for example, why (2) and (3) differ precisely by $\operatorname{Var} L(MK)$.

In fact, a deep theory is in the background, and the question of the title is to ask how it can be systematically exploited in this and other stochastic contexts. In the next sections, we briefly outline the theory and then turn to a specific question connected with Gaussian processes.

2. The Aleksandrov–Fenchel inequality

The bound (3) can be regarded as the first in an infinite sequence of inequalities, each of which is a stochastic formulation of the Aleksandrov–Fenchel (A–F) inequality in convex geometry. The A–F inequality is well–known to specialists as a powerful tool, having as implications the isoperimetric inequality (in all dimensions) and the Brunn–Minkowski inequality [8]. It has been successfully applied to problems in combinatorics as well as to the resolution of the van der Waerden permanent conjecture [6, 7, 14, 15, 16]. Interestingly, the original plan for the classic compilation [4] was to have a sequel based entirely on the A–F inequality. A closely

[1]Department of Statistics, University of Connecticut, Storrs, CT 06269, USA. e-mail: r.vitale@uconn.edu

Keywords and phrases: Aleksandrov–Fenchel inequality, Brunn–Minkowski inequality, convex body, Gaussian process, intrinsic volume, isoperimetric inequality, mean width, Steiner formula, tail bound.

AMS 2000 subject classifications: primary 52A40; secondary 52A05, 52A39, 60D05, 60G15.

related inequality on mixed discriminants [1, 2, 3] has found applications in stochastic settings. In view of this background, it is surprising that the A–F inequality itself has not found more applications in stochastic settings. One exception is questions in the theory of Gaussian processes, to which we turn in the next section.

A quick introduction to the A–F inequality goes as follows. It is part of *Brunn–Minkowski Theory* [13], which deals with the interaction between volume evaluation and vector addition of convex bodies (i.e., compact, convex subsets). For convex bodies $K_1, K_2, \ldots, K_n$ in $\mathbb{R}^d$ and positive coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\mathrm{vol}\,(\lambda_1 K_1 + \lambda_2 K_2 + \cdots + \lambda_n K_n) = \sum_{i_1, i_2, \cdots, i_d = 1}^{n} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_d} V(K_{i_1}, K_{i_2}, \ldots, K_{i_d}), \tag{4}$$

where, without loss of generality, the coefficients $V(\cdot)$ are taken to be symmetric in their arguments. The A–F inequality then asserts the following:

Theorem 1. *For convex bodies $K_1, K_2, \ldots, K_d$ in $\mathbb{R}^d$,*

$$V^2(K_1, K_2, K_3, \ldots, K_d) \geq V(K_1, K_1, K_3, \ldots, K_d)\; V(K_2, K_2, K_3, \ldots, K_d). \tag{5}$$

For the special case of a parallel body $K + \lambda B$ (B, the unit ball in $\mathbb{R}^d$), (4) simplifies to the Steiner formula [17]

$$\mathrm{vol}(K + \lambda B) = \sum_{i=0}^{d} \lambda^i \omega_i V_{d-i}(K), \tag{6}$$

where ω_i is the volume of the unit ball in $\mathbb{R}^i$ and the $V_i(K)$ are the *intrinsic volumes of K* ($V_0 \equiv 1$). Then (5) translates to the sequence $\{i!\, V_i(K)\}_{i=0}^{\infty}$ being log–concave:

$$(i!\, V_i(K))^2 \geq (i-1)! V_{i-1}(K) \;\cdot\; (i+1)! V_{i+1}(K) \qquad i = 1, 2, \ldots, d-1 \tag{7}$$

(elsewhere this property has been called *ultra–logconcavity of order* ∞ [10, 11]).

3. Intrinsic volumes and Gaussian processes

The theory of Gaussian processes has been heavily influenced by convex geometry [5, 9, 18, 22, 23]. Here we draw especially on [19, 20, 21, 22].

A popular approach to Gaussian processes is *canonical indexing*: suppose that $A \subseteq \mathbb{R}^d$ and that $Z = (Z_1, Z_2, \ldots, Z_d)$ are iid $N(0,1)$ variables. A canonically indexed Gaussian process $X_A = \{X_t, t \in A\}$ has the form $X_t = \sum_1^d t_i Z_i = <t, Z>$ (this process evidently has "rank" no greater than d, but similar definitions can be made in Hilbert space for more general processes). If $A = K$, a convex body, then intrinsic volumes come into play. For $j = 1, 2, \ldots, d$ define the vector process $X_t^{j*} = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(j)}\right)$, where the components are independent copies of X_t. Further, define the (random) convex body $X_K^{j*} = \overline{\mathrm{conv}}\{X_t^{j*},\, t \in K\} \subseteq \mathbb{R}^j$. Then

$$V_j(K) = \frac{(2\pi)^{j/2}}{j!\,\omega_j} E\mathrm{vol}_j\left(X_K^{j*}\right). \qquad j = 1, 2, \ldots, d \tag{8}$$

The *Wills functional* is given by

$$W(K) = E\left[\exp \sup_{t\in K}\left(X_t - \frac{1}{2}EX_t^2\right)\right] \tag{9}$$

and has the generating function expansion

$$W(rK) = \sum_{j=0}^{d}\left(\frac{r}{\sqrt{2\pi}}\right)^j V_j(K). \tag{10}$$

An important consideration for Gaussian processes is "size," which is traditionally interpreted as $\sup_t X_t$. Tail probability bounds are of various types, and we illustrate the application of the preceding ideas to a sharpening of a bound in [22]. Fix K, and recall that the A–F inequality implies that $a_j = j!\,V_j(K)$ is a log-concave sequence:

$$\log a_j \le \log a_i + (\log a_{i+1} - \log a_i)(j-i),$$

which implies

$$V_j(K) \le \frac{i!\,V_i(K)}{j!}\left(\frac{(i+1)V_{i+1}(K)}{V_i(K)}\right)^{j-i}.$$

Substituting into (10) and summing $j = 0, \ldots, \infty$ yields

$$\begin{aligned} W(rK) &\le i!\,V_i(K)\left(\frac{V_i(K)}{(i+1)V_{i+1}(K)}\right)^i \exp\left[\frac{(i+1)V_{i+1}(K)r}{\sqrt{2\pi}V_i(K)}\right] \\ &\le \frac{i!\,V_i(K)}{(2\pi)^{i/2}m_i^i(K)}\exp\left[m_i(K)r\right], \end{aligned}$$

where

$$m_i(K) = \frac{i\,V_i(K)}{\sqrt{2\pi}V_{i-1}(K)}. \tag{11}$$

A straightforward application of Markov's inequality then provides the bound

$$P(\sup_t X_t \ge a) \le \inf_i\left\{\frac{i!\,V_i(K)}{(2\pi)^{i/2}m_i^i(K)}\exp\left[-\frac{(m_i(K)-a)^2}{2\sigma^2}\right]\right\},$$

where $a > 0$ and $\sigma^2 = \sup_{t\in K} EX_t^2 = \sup_{t\in K}\|t\|^2$.

This brings us to the issue mentioned in the introduction: the way in which the values $m_i(K)$ have arisen suggests that they may be natural parameters of the process for other questions as well. It is easy to verify that $m_1(K)$ is at once $E\sup_{t\in K} X_t$ and proportional to the mean width of K and, as such, has linear dimension. The succeeding m_i also have linear units, and evidently provide alternate size measures for both K and $\{X_t, t \in K\}$. Their asymptotic behavior reflects the regularity of the process (see [24] for details), but it seems clear that their specific values must also calibrate successive i-th order properties of some type for the process. What these are remains for investigation.

References

[1] Aleksandrov, A. D. (1938). Zur Theorie der gemischten Volumina von konvexen Köpern IV (Russian, German summary). *Mat. Sbornik* **3**, 227–251.

[2] Bapat, R. B. (1988). Discrete multivariate distributions and generalized log-concavity. *Sankhyā Ser. A* **50**, 98–110.

[3] Bapat, R. B. (1990). Permanents in probability and statistics. *Linear Algebra Appl.* **127**, 3–25.

[4] Beckenbach, E. F., and Bellman, R. (1971). *Inequalities*. Springer-Verlag, New York.

[5] Borell, C. (1975). The Brunn–Minkowski inequality in Gauss space. *Invent. Math.* **30**, 207–216.

[6] Egorychev, G. P. (1981). The solution of van der Waerden's problem for permanents. *Adv. in Math.* **42**, 299–305.

[7] Falikman, D. I. (1981). Proof of the van der Waerden conjecture on the permanent of a doubly stochastic matrix. (Russian) *Mat. Zametki* **29**, 931–938.

[8] Gardner, R. J. (2002). The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.* (N.S.) **39**, no. 3, 355–405.

[9] Landau, H. J., and Shepp, L. A. (1970). On the supremum of a Gaussian process. *Sankhyā Ser. A* **32**, 369–378.

[10] Liggett, T. M. (1997). Ultra logconcave sequences and negative dependence. *J. Combin. Theory Ser. A* **79**, 315–325.

[11] Pemantle, R. (2000). Towards a theory of negative dependence. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. *J. Math. Phys.* **41**, 1371–1390. MR1757964

[12] Rubin, H. and Vitale, R. A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **6**, 165-170.

[13] Schneider, R. (1993). *Convex Bodies: the Brunn-Minkowski Theory*. Camb. Univ. Press, New York.

[14] Stanley, R. P. (1981). Two combinatorial applications of the Aleksandrov-Fenchel inequalities. *J. Combin. Theory Ser. A* **31**, 56–65.

[15] Stanley, R. P. (1986). Two poset polytopes. *Discrete Comput. Geom.* **1**, 9–23.

[16] Stanley, R. P. (1989). Log-concave and unimodal sequences in algebra, combinatorics, and geometry. Graph theory and its applications: East and West (Jinan, 1986), 500–535, *Ann. New York Acad. Sci.*, **576**, New York Acad. Sci.

[17] Steiner, J. (1840). Von dem Krümmunsschwerpunkte ebener Curven. *J. Reine Angew. Math.* **21**, 33–63; *Ges. Werke*, vol. 2, Reimer, Berlin, 1882, 99–159.

[18] Sudakov, V. N., and Tsirel'son, B. S. (1978). Extremal properties of half-spaces for spherically invariant measures. *J. Soviet Math.* **9**, 9–18; translated from *Zap. Nauch. Sem. L.O.M.I.* **41**, 14–24 (1974). MR365680

[19] Tsirel'son, B. S. (1982). A geometric approach to maximum likelihood estimation for infinite–dimensional Gaussian location I. *Theory Prob. Appl.* **27**, 411–418.

[20] Tsirel'son, B. S. (1985). A geometric approach to maximum likelihood estimation for infinite–dimensional Gaussian location II. *Theory Prob. Appl.* **30**, 820–828.

[21] Tsirel'son, B. S. (1986). A geometric approach to maximum likelihood estimation for infinite–dimensional location III. *Theory Prob. Appl.* **31**, 470–483.

[22] Vitale, R. A. (1996). The Wills functional and Gaussian processes. *Ann. Probab.* **24**, 2172–2178.

[23] Vitale, R. A. (1999). A log–concavity proof for a Gaussian exponential bound. In *Contemporary Math.: Advances in Stochastic Inequalities* (T.P. Hill, C. Houdré, eds.) **234**, AMS, 209–212. MR1694774

[24] Vitale, R. A. (2001). Intrinsic volumes and Gaussian processes. *Adv. Appl. Prob.* **33**, 354–364.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 342–347
© Institute of Mathematical Statistics, 2004

Generalized Accept–Reject sampling schemes

George Casella[1,*] Christian P. Robert[2,†], and Martin T. Wells[3,*]

University of Florida, Université Paris 9 – Dauphine and Cornell University

Abstract: This paper extends the Accept–Reject algorithm to allow the proposal distribution to change at each iteration. We first establish a necessary and sufficient condition for this generalized Accept–Reject algorithm to be valid, and then show how the resulting estimator can be improved by Rao-Blackwellization. An application of these results is to the perfect sampling technique of Fill (1998), which is a generalized Accept–Reject algorithm.

1. Preface by GC

This paper is especially appropriate for a volume dedicated to Herman Rubin, as he was the first person who ever mentioned the Accept–Reject algorithm to me although, at the time, I didn't understand a word that he was talking about. I was a graduate student at Purdue in the mid-70s, and Herman was always working on some problem, and if he saw you in the hall he would tell you about it. One day he told me he was working on an algorithm that generated "test exponentials" to get normal random variables. I had no idea why anyone would want to do such a thing (remember the 70s ? – we were proving theorems!). Herman eventually wrote a technical report, but I don't think I ever read it and don't know if it ever was published. And then Herman got interested in other things. But when I think of this story I often wonder how much further along Monte Carlo methods would be today if Herman kept his interest in those "test exponentials"!

2. Introduction

Accept–Reject algorithms are based on the use of a proposal distribution g which serves to simulate from a given target density f, when the ratio f/g is bounded by $1/\epsilon$, say. The standard Accept–Reject Algorithm is

Algorithm $\mathbf{A}_1$—Accept–Reject.
`At iteration` i $(i \geq 1)$

1. `Generate` $X_i \sim g$ `and` $U_i \sim \mathcal{U}([0,1])$, *independently.*
2. `If` $U_i \leq \epsilon f(X_i)/g(X_i)$, `accept` $X_i \sim f$;
3. `otherwise, move to iteration` $i+1$.

[†]Research supported by EU TMR network ERB-FMRX-CT96-0095.
[*]Both Prof. Casella and Prof. Wells are supported by National Science Foundation Grant DMS-9971586.
[1]Department of Statistics, University of Florida, Gainesville, FL 32611-8545, USA. e-mail: casella@stat.ufl.edu
[2]Laboratoire de Statistique, CREST, Insee, Paris and Université Paris 9 – Dauphine, France. e-mail: xian@ceremade.u-dauphine.fr
[3]Cornell University, Ithaca, NY 14851. e-mail: mtw1@cornell.edu
Keywords and phrases: Monte Carlo methods, Accept–Reject, stopping rule, recycling, uniform variable.
AMS 2000 subject classifications: 65C60.

Since the inequality is not always satisfied, the algorithm generates pairs (X_i, U_i) that are rejected. These pairs can be recycled in many ways, including the Rao-Blackwellizing approach by Casella and Robert (1996) which replaces the standard estimator δ based on the accepted pairs with the conditional expectation $\mathbb{E}[\delta|x_1, \ldots, x_n, n]$, which integrates out the uniform variables.

We give in this note a necessary and sufficient condition for a generalized Accept–Reject algorithm to be valid and show that Rao–Blackwellization also applies here, allowing the use of the rejected samples to produce an improved estimator.

This work was partially motivated by that of Fill (1998), who developed an interruptible perfect sampling algorithm as an alternative to Propp and Wilson's (1996) coupling from the past technique. Perfect sampling results in iid outputs from the stationary distribution of the MCMC Markov chain (see Dimakos (2001), Robert and Casella (1999) or Casella, Lavine and Robert (2000) for introductions to perfect sampling). At the core of Fill's algorithm, described in Section 5, is an Accept–Reject algorithm with the feature that the proposal distribution can be modified at each step.

The possibility of changing the proposal distribution at each failure/rejection implies that his method does not fall in the category of a standard Accept–Reject algorithm. It is this more general Accept–Reject algorithm that we are interested in.

3. A generalized Accept–Reject algorithms

We consider the following extension to the standard Accept–Reject algorithm:

Algorithm $\mathbf{A}_2$—Generalized Accept–Reject.
`At iteration` i $(i \geq 1)$

1. `Generate` $X_i \sim g_i$ `and` $U_i \sim \mathcal{U}([0,1])$, *independently.*

2. `If` $U_i \leq \epsilon_i f(X_i)/g_i(X_i)$, `accept` $X_i \sim f$;

3. `otherwise, move to iteration` $i+1$. -

Thus, at each iteration i $(0 < i < \infty)$, the algorithm uses a different pair (g_i, ϵ_i) such that $\epsilon_i f(x)/g_i(x) \leq 1$, uniformly in x. Each of these pairs is thus acceptable for the original Accept–Reject scheme. However, the proposal distribution keeps changing at each reject iteration and may be more adaptive than the single Accept–Reject proposal distribution, or even the adaptive rejection algorithm of Gilks and Wild (1992), which uses an envelope on the target density.

If the proposal distribution is parameterized by a parameter θ, we can select a pre-determined sequence of values of θ to monitor the performance in simulating the distribution of interest f. The value of θ at the time of acceptance can then be exploited in further simulations without jeopardizing the independence properties of the algorithm.

The extension of the Accept–Reject Algorithm does not hold in full generality, in the sense that the distribution of the accepted random variable may not necessarily be the correct one. A minimum requirement must be imposed on the sequence of the ϵ_i's (and hence on the g_i's).

If we denote by Z the (possibly defective) random variable that is output by Algorithm A_2, Z has the cdf (for simplicity, in the univariate continuous case):

$$\begin{aligned}
&P(Z \le z) \\
&\quad = \sum_{i=1}^{\infty} P(Z \le z, Z = X_i) \\
&\quad = \sum_{i=1}^{\infty} P\big(X_i \le Z, U_i \le f(X_i)\epsilon_i / g_i(X_i)\big) \prod_{j=1}^{i-1} P\big(U_j \ge f(X_j)\epsilon_j / g_j(X_j)\big) \\
&\quad = \sum_{i=1}^{\infty} \int_{-\infty}^{z} \frac{f(x)\epsilon_i}{g_i(x)} g_i(x)\,dx \prod_{j=1}^{i-1}(1-\epsilon_j) = \int_{-\infty}^{z} f(x)\,dx \sum_{i=1}^{\infty} \epsilon_i \prod_{j=1}^{i-1}(1-\epsilon_j).
\end{aligned}$$

Therefore, the output is distributed from f if $\sum_{i=1}^{\infty} \epsilon_i \prod_{j=1}^{i-1}(1-\epsilon_j) = 1$. The following theorem ties this condition to the divergence of an associated series.

Theorem 3.1. *The Generalized Accept–Reject Algorithm is valid if, and only if, the series* $\sum_{i=1}^{\infty} \log(1-\epsilon_i)$ *diverges, since*

$$\sum_{i=1}^{\infty} \epsilon_i \prod_{j=1}^{i-1}(1-\epsilon_j) = 1 \quad \textit{if and only if} \quad \sum_{i=1}^{\infty} \log(1-\epsilon_i) \textit{ diverges.} \tag{1}$$

Proof. Note first that $\sum \epsilon_i \prod_{j=1}^{i-1}(1-\epsilon_j)$ necessarily converges to a limit less than, or equal to, 1 since

(a) for every $n \ge 1$,

$$\begin{aligned}
\xi_n &= \sum_{i=1}^{n} \epsilon_i \prod_{j=1}^{i-1}(1-\epsilon_j) \\
&= \epsilon_1 + (1-\epsilon_1)\big\{\epsilon_2 + (1-\epsilon_2)[\ldots(1-\epsilon_{n-1})\epsilon_n)\ldots]\big\} \\
&\le \epsilon_1 + (1-\epsilon_1)\big\{\epsilon_2 + (1-\epsilon_2)[\ldots\epsilon_{n-1} + (1-\epsilon_{n-1}))\ldots]\big\} \\
&= 1.
\end{aligned}$$

(b) the sequence $\{\xi_n\}$ is increasing with n.

Now, $\{\xi_n\}$ converges to 1 if, and only if, for every $0 < \eta < 1$, there exists n_0 such that

$$\xi_n > 1 - \eta \quad \text{for} \quad n > n_0. \tag{2}$$

The condition (2) is equivalent to, for $n > n_0$,

$$\begin{aligned}
&\epsilon_1 + (1-\epsilon_1)\big\{\epsilon_2 + (1-\epsilon_2)[\ldots(1-\epsilon_{n-1})\epsilon_n)\ldots]\big\} > 1-\eta \\
&\quad \Leftrightarrow \epsilon_2 + (1-\epsilon_2)\big\{\epsilon_3 + \ldots(1-\epsilon_{n-1})\epsilon_n]\ldots\big\} > \frac{1-\epsilon_1-\eta}{1-\epsilon_1} = 1 - \frac{\eta}{1-\epsilon_1} \\
&\quad \Leftrightarrow \ldots \\
&\quad \Leftrightarrow \epsilon_n > 1 - \frac{\eta}{\prod_{i=1}^{n-1}(1-\epsilon_i)}.
\end{aligned} \tag{3}$$

The sequence $\omega_n = \prod_{i=1}^{n-1}(1-\epsilon_i)$ with $\omega_1 = 1$ is decreasing and nonnegative. Thus, it either converges to 0 or to $\alpha > 0$. If it converges to 0, that is, if $\sum \log(1-\epsilon_i)$

diverges, the ratio η/ω_n goes to $+\infty$ with n and the right hand side in (3) is negative for n large enough, which ensures that (2) holds.

If $\{\omega_n\}$ converges to $\alpha > 0$, the series $\sum \log(1-\epsilon_i)$ converges and $\log(1-\epsilon_n)$ goes to 0 as n goes to infinity by Cauchy's criterion. Thus, $\{\epsilon_n\}$ converges to 0. Therefore, for δ small enough, there exists n_1 such that $\epsilon_n < \delta$ for $n > n_1$. If one chooses η such that $1 - \frac{\eta}{\alpha} = \delta$ and if (2) holds, one gets $\epsilon_n < \delta < \epsilon_n$ for $n > \max(n_0, n_1)$, which is impossible. □

This result has several implications. First, it shows that continued modifications of the proposal distribution in the Accept–Reject algorithm are legitimate as long as the acceptance rate ϵ_n does not converge to zero too fast. Second, the acceptance rate ϵ_n does not have to go to 1 with n, so some ϵ_n's (even an infinity of them) may be equal to 0, and the algorithm remains valid. Note however, that if one ϵ_n is equal to 1, the sequence terminates.

Theorem 3.1 applies to and validates the generalized Accept–Reject algorithm not only when ϵ_n is constant, but also when the ϵ_n's are periodic in n, and when the sequence $\{\epsilon_n\}$ is uniformly bounded away from 0.

4. Rao–Blackwellization

The output from the generalized Accept–Reject algorithm is as follows: A sequence $Y_1, Y_2, \ldots$ of independent random variables is generated from the g_i's along with a corresponding sequence $U_1, U_2, \ldots$ of uniform random variables. We show how to extend the results of Casella and Robert (1996) to this more general algorithm.

Given a function h, the Accept–Reject estimator of $\mathbb{E}^f\{h(X)\}$, based upon a sample $X_1, \ldots, X_t$, with t fixed, is made of the t accepted values among the Y_j's and is given by

$$\hat{\tau}_1 = \frac{1}{t}\sum_{i=1}^{t} h(X_i) = \frac{1}{t}\sum_{i=1}^{N} \mathbf{I}(U_i \le W_i) h(Y_i), \tag{4}$$

where N, the number of Y_j's generated, is a random integer satisfying

$$\sum_{i=1}^{N} I(U_i \le W_i) = t \qquad \text{and} \qquad \sum_{i=1}^{N-1} I(U_i \le W_i) = t-1,$$

with $W_i = f(Y_i)\epsilon_i / g_i(Y_i)$. By the Rao-Blackwell Theorem, the conditional expectation

$$\hat{\tau}_2 = \frac{1}{t}\mathbb{E}\left\{ \sum_{i=1}^{N} \mathbf{I}(U_i \le W_i) h(Y_i) \Big| N, Y_1, \ldots, Y_N \right\} \tag{5}$$

improves upon (4).

The joint distribution of $(N, Y_1, \ldots, Y_N, U_1, \ldots, U_N)$ is given by

$$\begin{aligned} &P(N = n, Y_1 \le y_1, \ldots, Y_n \le y_n, U_1 \le u_1, \ldots, U_n \le u_n) \\ &\quad = \int_{-\infty}^{y_n} g_n(v_n)(u_n \wedge w_n) dv_n \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_{n-1}} g_1(v_1) \ldots g_{n-1}(v_{n-1}) \\ &\qquad \times \sum_{(i_1,\ldots,i_{t-1})} \prod_{j=1}^{t-1} (w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1} (u_{i_j} - w_{i_j})^+ dv_1 \ldots dv_{n-1}, \end{aligned}$$

where $w = \varepsilon f(v)/g(v)$ (with appropriate subscripts) and the last sum is over all subsets of $\{1, \ldots, n-1\}$ of size $t-1$. Therefore, the conditional density of the U_i's

is given by

$$f(u_1,\dots,u_n|N=n,y_1,\dots,y_n) = \left\{\sum_{(i_1,\dots,i_{t-1})}\prod_{j=1}^{t-1} w_{i_j}\prod_{j=t}^{n-1}(1-w_{i_j})\right\}^{-1} \times\left\{\sum_{(i_1,\dots,i_{t-1})}\prod_{j=1}^{t-1}\mathbf{I}(u_{i_j}\le w_{i_j})\prod_{j=t}^{n-1}\mathbf{I}(u_{i_j}>w_{i_j})\right\}\frac{\mathbf{I}(u_n\le w_n)}{w_n},$$

where, analogously, $w=\varepsilon f(y)/g(y)$. Using this distribution we can calculate, conditional on $(N,y_1,\dots,y_N)$, the probability ρ_i of the events $\{U_i\le w_i\}$ and thus derive the weights of $h(Y_i)$ in the estimator $\hat\tau_2$. The calculations involve averaging over permutations of the realized sample and yield, for $i<n$,

$$\rho_i = w_i\sum_{(i_1,\dots,i_{t-2})}\prod_{j=1}^{t-2} w_{i_j}\prod_{j=t-1}^{n-2}(1-w_{i_j})\Big/\sum_{(i_1,\dots,i_{t-1})}\prod_{j=1}^{t-1} w_{i_j}\prod_{j=t}^{n-1}(1-w_{i_j}), \qquad (6)$$

while $\rho_n=1$. The numerator sum is over all subsets of $\{1,\dots,i-1,i+1,\dots,n-1\}$ of size $t-2$, and the denominator sum is over all subsets of size $t-1$. The following result therefore holds.

Theorem 4.1. *For $N=n$, the Rao–Blackwellized version of* (4) *is given by*

$$\hat\tau_2=\frac{1}{t}\sum_{i=1}^{n}\rho_i h(Y_i)$$

where ρ_i is given by equation (6).

5. Perfect sampling

A *perfect sampling algorithm* for a Markov chain is an algorithm that produces a random variable that is exactly distributed according to the stationary distribution of the Markov chain using variables that are (typically) generated from the conditional distributions of the chain. Perfect sampling in Markov chains originated with the ingenious "coupling from the past" algorithm of Propp and Wilson (1996). In practice, however, this algorithm has some drawbacks, such as – for example – not being interruptible and thus creating biases in the output in cases of interruption for insufficient memory and such.

An alternative, interruptible, perfect sampling algorithm was proposed by Fill (1998). Since it is interruptible, Fill's perfect sampling algorithm seems to be somewhat more practical than coupling from the past, although it requires delicate reversibility and coupling arrangements as shown below.

Fill's algorithm (see also Fill *et al.* 1999) can be described as follows:

(a) Starting at an arbitrary state 0, run a finite state Markov chain (X_i) for t (fixed) steps, and record $X_t=x$.

(b) Starting Markov chains in at all possible states at time t, run them in reversed time, coupled with the original chain.

(c) If all these chains have coalesced, that is, if they all are in state 0 at time 0, then accept $X_t=x$ as an observation from the stationary distribution. If not, reject X_t and start again, possibly with different values of 0 and of t.

We now relate the result of the previous sections to Fill's algorithm.

- The surprising feature of this method is that it is a rejection algorithm with the clever twist that the probability of acceptance is exactly the probability of coalescence. This circumvents the problem of calculating this acceptance probability, which is typically not feasible.
- Fill's (1998) algorithm depends on a parameter t, which is the number of forward steps in the Markov chain and which can be modified at each iteration, by, for instance, doubling the value of t in a typical CFTP manner. Thus, the proposal distribution is changing at every iteration, and the algorithm is covered by Theorem 3.1 (but is not covered by the standard Accept–Reject algorithm).
- For Theorem 3.1 to validate Fill's (1998) algorithm, the series $\sum \log(1-\epsilon_i)$ of acceptance probabilities ϵ_i must diverge. The difficulty then lies in establishing this without the ϵ_i's being available, which is the essence of Fill's technique. However, if the selection is periodic, Fill's algorithm is indeed valid, provided some ϵ_i's are different from 0. In fact, in most practical cases Fill's algorithm will have an increasing acceptance rate, so will be covered by Theorem 3.1.
- The application of Theorem 4.1 to Fill's algorithm requires some further work since, in that case, the weights $w_i = f(x_i)/K^t(0, x_i)$ are not directly available. Note however that in some setups $K^t(0, x)$ may be known, while, in others, it can be estimated, since it is also equal to the probability of acceptance, that is, the probability of coalescence in state 0. Thus, we can implement the Rao–Blackwellized improvement with estimated weights.

References

[1] Casella, G., Lavine, M., and Robert, C.P. (2000). An introduction to perfect sampling. *Amer. Statist.* **55** 299–305. MR1939363

[2] Casella, G., and Robert, C. P. (1996). Rao-Blackwellization of sampling schemes. *Biometrika* **83**(1), 81–94. MR1399157

[3] Dimakos, X. K. (2001). "A Guide to Exact Simulation". *International Statistical Review* **69** 27–48.

[4] Fill, J. A. (1998). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, **8**(1) 131–162. MR1620346

[5] Fill, J. A., Machida, M., Murdoch, D. J. and Rosenthal, J. S. (1999). Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures and Algorithms* **17** 219–316. MR1801136

[6] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–348.

[7] Propp, J.G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, **9**(1& 2), 223–252. MR1611693

[8] Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York. MR1707311

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 348–377
© Institute of Mathematical Statistics, 2004

Scalable mining for classification rules in relational databases

MinWang[1,*], Bala Iyer [2], and Jeffrey Scott Vitter[3,†]

IBM T. J. Watson Research Center, IBM Silicon Valley Lab and Purdue University

Abstract: Data mining is a process of discovering useful patterns (knowledge) hidden in extremely large datasets. Classification is a fundamental data mining function, and some other functions can be reduced to it. In this paper we propose a novel classification algorithm (classifier) called MIND (MINing in Databases). MIND can be phrased in such a way that its implementation is very easy using the extended relational calculus SQL, and this in turn allows the classifier to be built into a relational database system directly. MIND is truly scalable with respect to I/O efficiency, which is important since scalability is a key requirement for any data mining algorithm.

We have built a prototype of MIND in the relational database management system DB2 and have benchmarked its performance. We describe the working prototype and report the measured performance with respect to the previous method of choice. MIND scales not only with the size of datasets but also with the number of processors on an IBM SP2 computer system. Even on uniprocessors, MIND scales well beyond dataset sizes previously published for classifiers. We also give some insights that may have an impact on the evolution of the extended relational calculus SQL.

1. Introduction

Information technology has developed rapidly over the last three decades. To make decisions faster, many companies have combined data from various sources in relational databases [16]. The data contain patterns previously undeciphered that are valuable for business purposes. Data mining is the process of extracting valid, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions. The extracted information can be used to form a prediction or classification model, or to identify relations between database records.

Since extracting data to files before running data mining functions would require extra I/O costs, users of IM as well as previous investigations [20, 19] have pointed to the need for the relational database management systems to have these functions built in. Besides reducing I/O costs, this approach leverages over 20 years of research and development in DBMS technology, among them are:

[1]Data Management Department, IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA. e-mail: min@us.ibm.com

[2]IBM Silicon Valley Lab, 555 Bailey Avenue, San Jose, CA 95141, USA. e-mail: balaiyer@us.ibm.com

[3]Purdue University, 150 North University Street, West Lafayette, IN 47907, USA. e-mail: jsv@purdue.edu

[*]Contact author. Support was provided in part by an IBM Graduate Fellowship.

[†]Support was provided in part by the Army Research Office through research grant DAAD19–03–1–0321, by the National Science Foundation through research grant CCR–9877133, and by an IBM research award.

Keywords and phrases: data mining, classification, classifier, relation database, SQL.

AMS 2000 subject classifications: 68P20.

salary	*age*	*credit rating*
65K	30	Safe
15K	23	Risky
75K	40	Safe
15K	28	Risky
100K	55	Safe
60K	45	Safe
62K	30	Risky

Table 1: Training set

- scalability,
- memory hierarchy management [30, 33],
- parallelism [5],
- optimization of the executions [6],
- platform independence, and
- client server API [27].

The classification problem can be described informally as follows: We are given a *training set* (or *DETAIL* table) consisting of many training examples. Each training example is a row with multiple attributes, one of which is a *class label*. The objective of classification is to process the *DETAIL* table and produce a *classifier*, which contains a description (model) for each class. The models will be used to classify future data for which the class labels are unknown (see [4, 28, 26, 9]).

Several classification models have been proposed in the literature, including neutral network, decision trees, statistical models, and genetic models. Among these models, decision tree model is particularly suited for data mining applications due to the following reasons: (1) ease of construction, (2) simple and easy to understand, and (3) acceptable accuracy [29]. Therefore, we focus on decision tree model in this paper. A simple illustration of of training data is shown in Table 1. The examples reflect the past experience of an organization extending credit. From those examples, we can generate the classifier shown in Figure 1.

Although memory and CPU prices are plunging, the volume of data available for analysis is immense and getting larger. We may not assume that the data are memory-resident. Hence, an important research problem is to develop accurate classification algorithms that are scalable with respect to I/O and parallelism. Accuracy is known to be domain-specific (e.g., insurance fraud, target marketing). However, the problem of scalability for large amounts of data is more amenable to a general solution. A classification algorithm should scale well; that is, the classification algorithm should work well even if the training set is huge and vastly overflows internal memory. In data mining applications, it is common to have training sets with several million examples. It is observed in [24] that previously known classification algorithms do not scale.

Random sampling is often an effective technique in dealing with large data sets. For simple applications whose inherent structures are not very complex, this

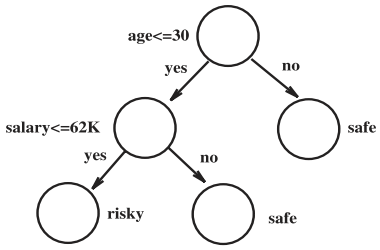


Figure 1: Decision tree for the data in Table 1

approach is efficient and gives good results. However, in our case, we do not favor random sampling for two main reasons:

1. In general, choosing the proper sample size is still an open question. The following factors must be taken into account:

   - The training set size.
   - The convergence of the algorithm. Usually, many iterations are needed to process the sampling data and refine the solution. It's very difficult to estimate how fast the algorithm will give a satisfactory solution.
   - The complexity of the model.

   The best known theoretical upper bounds on sample size suggest that the training set size may need to be immense to assure good accuracy [13, 21].

2. In many real applications, customers insist that *all* data, not just a sample of the data, must be processed. Since the data are usually obtained from valuable resources at considerable expense, they should be used as a whole throughout the analysis.

Therefore, designing a scalable classifier may be necessary or preferable, although we can always use random sampling in places where it is appropriate.

In [24, 29, 18], data access for classification follows "a record at a time" access paradigm. Scalability is addressed individually for each operating system, hardware platform, and architecture. In this paper, we introduce the MIND (MINing in Databases) classifier. MIND rephrases data classification as a classic database problem of summarization and analysis thereof. MIND leverages the extended relational calculus SQL, an industry standard, by reducing the solution to novel manipulations of SQL statements embedded in a small program written in C.

MIND scales, as long as the database primitives it uses scale. We can follow the recommendations in [3, 22] that numerical data be discretized so that each attribute has a reasonable number of distinct values. If so, operations like histogram formation, which have a significant impact on performance, can be done in a linear number of I/Os, usually requiring one, but never more than two passes over the *DETAIL* table [36]. Without the discretization, the I/O performance bound has an extra factor that is logarithmic but fortunately with a very large base M/B, which is the number of disk blocks that can fit in internal memory.

One advantage of our approach is that its implementation is easy. We have implemented MIND as a stored procedure, a common feature in modern DBMSs. In addition, since most modern database servers have very strong parallel query processing capabilities, MIND runs in parallel at no extra cost. A salient feature of MIND and one reason for its efficiency is its ability to do classification without any update to the *DETAIL* table.

We analyze and compare the I/O complexities of MIND and the previous method of choice, the interesting method called SPRINT [29]. Our theoretical analysis and experimental results show that MIND scales well whereas SPRINT can exhibit quadratic I/O times.

We describe our MIND algorithm in the next section; an illustrative example is given in Section 4. A theoretical performance analysis is given in Section 5. We revisit MIND algorithm in Section 6 using a general extension of current SQL standards. In Section 7, we present our experimental results. We make concluding remarks in Section 8.

2. The algorithm

2.1. Overview

A decision tree classifier is built in two phases: a growth phase and a pruning phase. In the growth phase, the tree is built by recursively partitioning the data until each partition is either "pure" (all members belong to the same class) or sufficiently small (according to a parameter set by the user). The form of the split used to partition the data depends upon the type of the attribute used in the split. Splits for a numerical attribute A are of the form $value(A) \leq x$, where x is a value in the domain of A. Splits for a categorical attribute A are of the form $value(A) \in S$, where S is a subset of $domain(A)$. We consider only binary splits as in [24, 29] for purpose of comparisons. After the tree has been fully grown, it is pruned to remove noise in order to obtain the final tree classifier.

The tree growth phase is computationally much more expensive than the subsequent pruning phase. The tree growth phase accesses the training set (or *DETAIL* table) multiple times, whereas the pruning phase only needs to access the fully grown decision tree. We therefore focus on the tree growth phase. The following pseudo-code gives an overview of our algorithm:

```
GrowTree(TrainingSet DETAIL)
  Initialize tree T and put all of records of DETAIL in the root;
  while (some leaf in T is not a STOP node)
    for each attribute i do
      form the dimension table (or histogram) DIM_i;
      evaluate gini index for each non-STOP leaf at each split value
      with respect to attribute i;
    for each non-STOP leaf do
      get the overall best split for it;
    partition the records and grow the tree for one more level according to the
    best splits;
    mark all small or pure leaves as STOP nodes;
  return T;
```

2.2. Leaf node list data structure

A powerful method called SLIQ was proposed in [24] as a semi-scalable classification algorithm. The key data structure used in SLIQ is a *class list* whose size is linear in the number of examples in the training set. The fact that the *class list* must be memory-resident puts a hard limitation on the size of the training set that SLIQ can handle.

In the improved SPRINT classification algorithm [29], new data structures *attribute list* and *histogram* are proposed. Although it is not necessary for the attribute list data structure to be memory-resident, the histogram data structure must be in memory to insure good performance. To perform the split in [29], a *hash table* whose size is linear in the number of examples of the training set is used. When the hash table is too large to fit in memory, splitting is done in multiple steps, and SPRINT does not scale well.

In our MIND method, the information we need to evaluate the split and perform the partition is stored in relations in a database. Thus we can take advantage of DBMS functionalities and memory management. The only thing we need to do is to incorporate a data structure that relates the database relations to the growing classification tree. We assign a unique number to each node in the tree. When loading the training data into the database, imagine the addition of a hypothetical column *leaf_num* to each row. For each training example, *leaf_num* will always indicate which leaf node in the current tree it belongs to. When the tree grows, the *leaf_num* value changes to indicate that the record is moved to a new node by applying a split. A static array called *LNL (*leaf node list) is used to relate the *leaf_num* value in the relation to the corresponding node in the tree. By using a labeling technique, we insure that at each tree growing stage, the nodes always have the identification numbers 0 through $N-1$, where N is the number of nodes in the tree. $LNL[i]$ is a pointer to the node with identification number i. For any record in the relation, we can get the leaf node it belongs to from its *leaf_num* value and *LNL* and hence we can get the information in the node (e.g. split attribute and value, number of examples belonging to this node and their class distribution).

To insure the performance of our algorithm, *LNL* is the only data structure that needs to be memory-resident. The size of *LNL* is equal to the number of nodes in the tree, so *LNL* can always be stored in memory.

2.3. Computing the gini index

A splitting index is used to choose from alternative splits for each node. Several splitting indices have recently been proposed. We use the *gini* index, originally proposed in [4] and used in [24, 29], because it gives acceptable accuracy. The accuracy of our classifier is therefore the same as those in [24, 29].

For a data set S containing N examples from C classes, $gini(S)$ is defined as

$$gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

where p_i is the relative frequency of class i in S. If a split divides S into two subset S_1 and S_2, with sizes N_1 and N_2 respectively, the *gini* index of the divided data $gini_{\text{split}}(S)$ is given by

$$gini_{\text{split}}(S) = \frac{N_1}{N} gini(S_1) + \frac{N_2}{N} gini(S_2) \tag{2}$$

The attribute containing the split point achieving the smallest *gini* index value is then chosen to split the node [4]. Computing the *gini* index is the most expensive part of the algorithm since finding the best split for a node requires evaluating the *gini* index value for each attribute at each possible split point.

The training examples are stored in a relational database system using a table with the following schema: $DETAIL(attr_1, attr_2, \ldots, attr_n, class, leaf_num)$, where $attr_i$ is the ith attribute, for $1 \leq i \leq n$, *class* is the classifying attribute, and *leaf_num* denotes which leaf in the classification tree the record belongs to. In actuality *leaf_num* can be computed from the rest of the attributes in the record and does not need to be stored explicitly. As the tree grows, the *leaf_num* value of each record in the training set keeps changing. Because *leaf_num* is a computed attribute, the *DETAIL* table is never updated, a key reason why MIND is efficient for the DB2 relational database. We denote the cardinality of the class label set by C, the number of the examples in the training set by N, and the number of attributes (not including class label) by n.

3. Database implementation of MIND

To emphasize how easily MIND is embeddable in a conventional database system using SQL and its accompanying optimizations, we describe our MIND components using SQL.

3.1. Numerical attributes

For every level of the tree and for each attribute $attr_i$, we recreate the dimension table (or histogram) called DIM_i with the schema $DIM_i(leaf_num, class, attr_i, count)$ using a simple SQL `SELECT` statement on *DETAIL*:

```
INSERT INTO DIM_i
SELECT leaf_num, class, attr_i, COUNT(*)
FROM DETAIL
WHERE leaf_num <> STOP
GROUP BY leaf_num, class, attr_i
```

Although the number of distinct records in *DETAIL* can be huge, the maximum number of rows in DIM_i is typically much less and is no greater than (#leaves in tree) $\times$ (#distinct values on $attr_i$) $\times$ (#distinct classes), which is very likely to be of the order of several hundreds [25]. By including *leaf_num* in the attribute list for grouping, MIND collects summaries for every leaf in one query. In the case that the number of distinct values of $attr_i$ is very large, preprocessing is often done in practice to further discretize it [3, 22]. Discretization of variable values into a smaller number of classes is sometimes referred to as "encoding" in data mining practice [3]. Roughly speaking, this is done to obtain a measure of aggregate behavior that may be detectable [25]. Alternatively, efficient external memory techniques can be used to form the dimension tables in a small number (typically one or two) linear passes, at the possible cost of some added complexity in the application program to give the proper hints to the DBMS, as suggested in Section 5.

After populating DIM_i, we evaluate the *gini* index value for each leaf node at each possible split value of the attribute i by performing a series of SQL operations that only involve accessing DIM_i.

It is apparent for each attribute i that its DIM_i table may be created in one pass over the *DETAIL* table. It is straightforward to schedule one query per dimension (attribute). Completion time is still linear in the number of dimensions. Commercial DBMSs store data in row-major sequence. I/O efficiencies may be obtained if it is possible to create dimension tables for all attributes in one pass over the *DETAIL* table. Concurrent scheduling of the queries populating the DIM_i tables is the simple approach. Existing buffer management schemes that rely on I/O latency appear to synchronize access to *DETAIL* for the different attributes. The idea is that one query piggy-backs onto another query's I/O data stream. Results from early experiments are encouraging [31].

It is also possible for SQL to be extended to insure that, in addition to optimizing I/O, CPU processing is also optimized. Taking liberty with SQL standards, we write the following query as a proposed SQL operator:

```
SELECT FROM DETAIL
INSERT INTO DIM_1{leaf_num, class, attr_1, COUNT(*)
       WHERE predicate
       GROUP BY leaf_num, class, attr_1}
INSERT INTO DIM_2{leaf_num, class, attr_2, COUNT(*)
       WHERE predicate
       GROUP BY leaf_num, class, attr_2}
   ⋮
INSERT INTO DIM_n{leaf_num, class, attr_n, COUNT(*)
       WHERE predicate
       GROUP BY leaf_num, class, attr_n}
```

The new operator forms multiple groupings concurrently and may allow further RDBMS query optimization.

Since such an operator is not supported, we make use of the object extensions in DB2, the *user-defined function* (udf) [32, 10, 17], which is another reason why MIND is efficient. User-defined functions are used for association in [2]. User-defined function is a new feature provided by DB2 version 2 [10, 17]. In DB2 version 2, the functions available for use in SQL statements extend from the system built-in functions, such as `avg, min, max, sum`, to more general categories, such as user-defined functions (udf). An external udf is a function that is written by a user in a host programming language. The `CREATE FUNCTION` statement for an external function tells the system where to find the code that implements the function. In MIND we use a udf to accumulate the dimension tables for all attributes in one pass over *DETAIL*.

For each leaf in the tree, possible split values for attribute i are all distinct values of $attr_i$ among the records that belong to this leaf. For each possible split value, we need to get the class distribution for the two parts partitioned by this value in order to compute the corresponding *gini* index. We collect such distribution information in two relations, *UP* and *DOWN*.

Relation *UP* with the schema $UP(leaf_num, attr_i, class, count)$ can be generated by performing a self-outer-join on DIM_i:

```
INSERT INTO UP
SELECT d1.leaf_num, d1.attr_i, d1.class, SUM(d2.count)
FROM (FULL OUTER JOIN DIM_i d1, DIM_i d2
      ON d1.leaf_num = d2.leaf_num AND
         d2.attr_i ≤ d1.attr_i AND
         d1.class = d2.class
      GROUP BY d1.leaf_num, d1.attr_i, d1.class)
```

Similarly, relation *DOWN* can be generated by just changing the $\leq$ to $>$ in the `ON` clause. We can also obtain *DOWN* by using the information in the leaf node and the *count* column in *UP* without doing a join on DIM_i again.

DOWN and *UP* contain all the information we need to compute the *gini* index at each possible split value for each current leaf, but we need to rearrange them in some way before the *gini* index is calculated. The following intermediate view can be formed for all possible classes k:

```
CREATE VIEW C_k_UP(leaf_num, attr_i, count) AS
SELECT leaf_num, attr_i, count
FROM UP
WHERE class = k
```

Similarly, we define view C_k*_DOWN* from *DOWN*.

A view *GINI_VALUE* that contains all *gini* index values at each possible split point can now be generated. Taking liberty with SQL syntax, we write

```
CREATE VIEW GINI_VALUE(leaf_num, attr_i, gini) AS
SELECT u1.leaf_num, u1.attr_i, f_gini
FROM C1_UP u1, ...,CC_UP uC, C1_DOWN d1, ...,CC_DOWN dC
WHERE u1.attr_i = ··· = uC.attr_i = d1.attr_i = ··· = dC.attr_i AND
      u1.leaf_num = ··· = uC.leaf_num = d1.leaf_num = ··· = dC.leaf_num
```

where f_{gini} is a function of $u_1.count, \ldots, u_n.count, d_1.count, \ldots, d_n.count$ according to (1) and (2).

We then create a table *MIN_GINI* with the schema *MIN_GINI*(*leaf_num*, *attr_name*, *attr_value*, *gini*):

```
INSERT INTO MIN_GINI
SELECT leaf_num, : i, attr_i, gini
FROM GINI_VALUE a
WHERE a.gini=(SELECT MIN(gini)
              FROM GINI_VALUE b
              WHERE a.leaf_num = b.leaf_num)
```

Table *MIN_GINI* now contains the best split value and the corresponding *gini* index value for each leaf node of the tree with respect to $attr_i$. The table formation query has a nested subquery in it. The performance and optimization of such queries are studied in [6, 26, 15].

We repeat the above procedure for all other attributes. At the end, the best split value for each leaf node with respect to all attributes will be collected in table *MIN_GINI*, and the overall best split for each leaf is obtained from executing the following:

```
CREATE VIEW BEST_SPLIT(leaf_num, attr_name, attr_value) AS
SELECT leaf_num, attr_name, attr_value
FROM MIN_GINI a
WHERE a.gini=(SELECT MIN(gini)
              FROM MIN_GINI b
              WHERE a.leaf_num = b.leaf_num)
```

3.2. Categorical attributes

For categorical attribute i, we form DIM_i in the same way as for numerical attributes. DIM_i contains all the information we need to compute the *gini* index for any subset splitting. In fact, it is an analog of the *count matrix* in [29], but formed with set-oriented operators.

A possible split is any subset of the set that contains all the distinct attribute values. If the cardinality of attribute i is m, we need to evaluate the splits for all the 2^m subsets. Those subsets and their related counts can be generated in a recursive way. The schema of the relation that contains all the k-sets is S_k_IN(*leaf_num*, *class*, $v_1, v_2, ..., v_k$, `count`). Obviously we have $DIM_i = S_1_IN$. S_k_IN is then generated from S_1_IN and S_{k-1}_IN as follows:

```
INSERT INTO S_k_IN
SELECT  p.leaf_num, p.class, p.v_1, ..., p.v_{k-1}, q.v_1, p.count + q.count
FROM (FULL OUTER JOIN S_{k-1}_IN p, S_1_IN q
      ON p.leaf_num = q.leaf_num AND
         p.class = q.class AND
         q.v_1 > p.v_{k-1})
```

We generate relation S_k_OUT from S_k_IN in a manner similar to how we generate *DOWN* from *UP*. Then we treat S_k_IN and S_k_OUT exactly as *DOWN* and *UP* for numerical attributes in order to compute the *gini* index for each k-set split.

A simple observation is that we don't need to evaluate all the subsets. We only need to compute the k-sets for $k = 1, 2, \ldots, \lfloor m/2 \rfloor$ and thus save time. For large m, greedy heuristics are often used to restrict search.

3.3. Partitioning

Once the best split attribute and value have been found for a leaf, the leaf is split into two children. If *leaf_num* is stored explicitly as an attribute in *DETAIL*, then the following `UPDATE` performs the split for each leaf:

```
UPDATE DETAIL
SET leaf_num = Partition(attr_1, ..., attr_n, class, leaf_num)
```

The user-defined function *Partition* defined on a record r of *DETAIL* as follows:

Partition(record r)
  Use the *leaf_num* value of r to locate the tree node n that r belongs to;
  Get the best split from node n;
  Apply the split to r, grow a new child of n if necessary;
  Return a new *leaf_num* according to the result of the split;

$attr_1$	$attr_2$	*class*	*leaf_num*
65K	30	Safe	0
15K	23	Risky	0
75K	40	Safe	0
15K	28	Risky	0
100K	55	Safe	0
60K	45	Safe	0
62K	30	Risky	0

Table 2: Initial relation *DETAIL* with implicit *leaf_num*

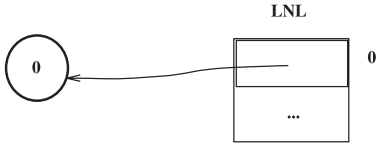


Figure 2: Initial tree

leaf_num	$attr_1$	*class*	*count*
0	15	2	2
0	60	1	1
0	62	2	1
0	65	1	1
0	75	1	1
0	100	1	1

Table 3: Relation DIM_1

However, *leaf_num* is not a stored attribute in *DETAIL* because updating the whole relation *DETAIL* is expensive. We observe that *Partition* is merely applying the current tree to the original training set. We avoid the update by replacing *leaf_num* by function *Partition* in the statement forming DIM_i. If *DETAIL* is stored on non-updatable tapes, this solution is required. It is important to note that once the dimension tables are created, the *gini* index computation for all leaves involves only dimension tables.

4. An example

We illustrate our algorithm by an example. The example training set is the same as the data in Table 1.

Phase 0: Load the training set and initialize the tree and *LNL*. At this stage, relation *DETAIL*, the tree, and *LNL* are shown in Table 2 and Figure 2.

Phase 1: Form the dimension tables for all attributes in one pass over *DETAIL* using user-defined function. The result dimension tables are show in Table 3–4.

Phase 2: Find the best splits for current leaf nodes. A best split is found through a set of operations on relations as described in Section 2.

First we evaluate the *gini* index value for $attr_1$. The procedure is depicted in Table 5–13.

leaf_num	$attr_2$	*class*	*count*
0	23	2	1
0	28	2	1
0	30	1	1
0	30	2	1
0	40	1	1
0	45	1	1
0	55	1	1

Table 4: Relation DIM_2

leaf_num	$attr_1$	*class*	*count*
0	15	1	0
0	15	2	2
0	60	1	1
0	60	2	2
0	62	1	1
0	62	2	3
0	65	1	2
0	65	2	3
0	75	1	3
0	75	2	3
0	100	1	4
0	100	2	3

Table 5: Relation *UP*

We can see that the best splits on the two attributes achieve the same *gini* index value, so relation *BEST_SPLIT* is the same as *MIN_GINI* except that it does not contain the column *gini*. We store the best split in each leaf node of the tree (the root node in this phase). In case of a tie for best split at a node, any one of them ($attr_2$ in our example) can be chosen.

Phase 3: Partitioning. According to the best split found in Phase 2, we grow the tree and partition the training set. The partition is reflected as *leaf_num* updates in relation *DETAIL*. Any new grown node that is pure or "small enough" is marked and reassigned a special *leaf_num* value *STOP* so that it is not processed further. The tree is shown in Figure 3 and the new *DETAIL* is shown in Table 14. Again, note *leaf_num* is never stored in *DETAIL*, so no update to *DETAIL* is necessary.

Phase 4: Repeat Phase 1 through Phase 3 until all the leaves in the tree become *STOP* leaves. The final tree and *DETAIL* are shown in Figure 4 and Table 15.

5. Performance analysis

Building classifiers for large training sets is an I/O bound application. In this section we analyze the I/O complexity of both MIND and SPRINT and compare their performances.

As we described in Section 2.1, the classification algorithm iteratively does two main operations: computing the splitting index (in our case, the *gini* index) and performing the partition. SPRINT [29] forms an *attribute list* (projection of the *DETAIL* table) for each attribute. In order to reduce the cost of computing the

leaf_num	$attr_1$	*class*	*count*
0	15	1	4
0	15	2	1
0	60	1	3
0	60	2	1
0	62	1	3
0	62	2	0
0	65	1	2
0	65	2	0
0	75	1	1
0	75	2	0

Table 6: Relation *DOWN*

leaf_num	$attr_1$	*count*
0	15	0.0
0	60	1.0
0	62	1.0
0	65	2.0
0	75	3.0
0	100	4.0

Table 7: Relation C_1_*UP*

leaf_num	$attr_1$	*count*
0	15	2.0
0	60	2.0
0	62	3.0
0	65	3.0
0	75	3.0
0	100	3.0

Table 8: Relation C_2_*UP*

gini index, SPRINT presorts each attribute list and maintains the sorted order throughout the course of the algorithm. However, the use of attribute lists complicates the partitioning operation. When updating the leaf information for the entries in an attribute list corresponding to some attribute that is *not* the splitting attribute, there is no local information available to determine how the entries should be partitioned. A *hash table* (whose size is linear in the number of training examples that reach the node) is repeatedly queried by random access to determine how the entries should be partitioned. In large data mining applications, the hash table is therefore not memory-resident, and several extra I/O passes may be needed, resulting in highly nonlinear performance.

MIND avoids the external memory thrashing during the partitioning phase by the use of dimension tables DIM_i that are formed while the *DETAIL* table, consisting of all the training examples, is streamed through memory. In practice, the dimension tables will likely fit in memory, as they are much smaller than the *DETAIL* table, and often preprocessing is done by discretizing the examples to make the number of distinct attribute values small. While vertical partitioning of

$leaf_num$	$attr_1$	$count$
0	15	4.0
0	60	3.0
0	62	3.0
0	65	2.0
0	75	1.0

Table 9: Relation C_1_DOWN

$leaf_num$	$attr_1$	$count$
0	15	1.0
0	60	1.0
0	62	0.0
0	65	0.0
0	75	0.0

Table 10: Relation C_2_DOWN

$leaf_num$	$attr_1$	$gini$
0	15	0.22856
0	60	0.40474
0	62	0.21428
0	65	0.34284
0	75	0.42856

Table 11: Relation $GINI_VALUE$

DETAIL may also be used to compute the dimension tables in linear time, we show that it is not a must. Data in and data archived from commercial databases are mostly in row major order. The layout does not appear to hinder performance.

If the dimension tables cannot fit in memory, they can be formed by sorting in linear time, if we make the weak assumption that $(M/B)^c \geq D/B$ for some small positive constant c, where D, M, and B are respectively the dimension table size, the internal memory size, and the block size [7, 36]. This optimization can be obtained automatically if SQL has the multiple grouping operator proposed in Section 3.1 and with appropriate query optimization, or by appropriate restructuring of the SQL operations. The dimension tables themselves are used in a stream fashion when forming the *UP* and *DOWN* relations. The running time of the algorithm thus scales linearly in practice with the training set size.

Now let's turn to the detailed analysis of the I/O complexity of both algorithms. We will use the parameters in Table 16 (all sizes are measured in bytes) in our analysis.

Each record in *DETAIL* has n attribute values of size r_a, plus a class label that we assume takes one (byte). Thus we have $r = nr_a + 1$. For simplicity we regard r_a as some unit size and thus $r = O(n)$. Each entry in a dimension table consists of one node number, one attribute value, one class label and one count. The largest node number is 2^L, and it can therefore be stored in L bits, which for simplicity we assume can fit in one word of memory. (Typically L is on the order of 10–20. If desired, we can rid ourselves of this assumption on L by rearranging *DETAIL* or a

leaf_num	*attr_name*	*attr_value*	*gini*
0	1	62	0.21428

Table 12: Relation *MIN_GINI* after $attr_1$ is evaluated

leaf_num	*attr_name*	*attr_value*	*gini*
0	1	62	0.21428
0	2	30	0.21428

Table 13: Relation *MIN_GINI* after $attr_1$ and $attr_2$ are evaluated

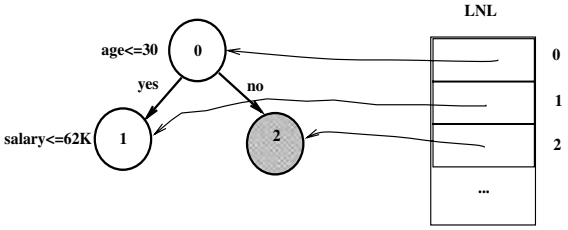


Figure 3: Decision tree at Phase 3

copy of *DETAIL* so that no *leaf_num* field is needed in the dimension tables, but in practice this is not needed.) The largest count is N, so $r_d = O(\log N)$. Counts are used to record multiple instances of a common value in a compressed way, so they always take less space than the original records they represent. We thus have

$$D_k \leq \min\{nN, VC2^k r_d\}. \tag{3}$$

In practice, the second expression in the min term is typically the smaller one, but in our worst-case expressions below we will often bound D_k by nN.

Claim 1. *If all dimension tables fit in memory, that is, $D_k \leq M$ for all k, the I/O complexity of MIND is*

$$O\left(\frac{LnN}{B}\right), \tag{4}$$

which is essentially best possible.

Proof. If all dimension tables fit in memory, then we only need to read *DETAIL* once at each level. Dimension tables for all attributes are accumulated in memory when each *DETAIL* record is read in. When the end of *DETAIL* table is reached, we'll have all the unsorted dimension tables in memory. Then sorting and *gini* index computation are performed for each dimension table, best split will be found for each current leaf node.

The I/O cost to read in *DETAIL* once is $rN/B = O(nN/B)$, and there are L levels in the final classifier, so the total I/O cost is $O(LnN/B)$. □

Claim 2. *In the case when not all dimension tables fit in memory at the same time, but each individual dimension table does, the I/O complexity of MIND is*

$$O\left(\frac{LnN}{B}\log_{M/B} n\right). \tag{5}$$

$attr_1$	$attr_2$	$class$	$leaf_num$
65K	30	Safe	1
15K	23	Risky	1
75K	40	Safe	2⇒$STOP$
15K	28	Risky	1
100K	55	Safe	2⇒$STOP$
60K	45	Safe	2⇒$STOP$
62K	30	Risky	1

Table 14: Relation $DETAIL$ with implicit $leaf_num$ after Phase 3

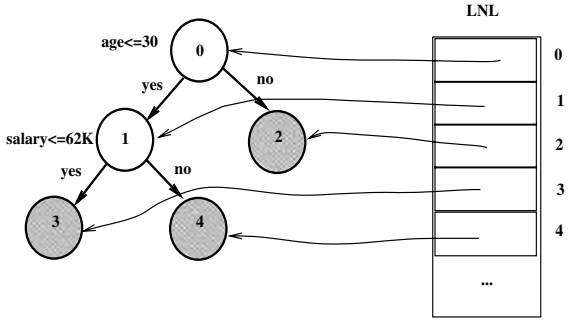


Figure 4: Final decision tree

$attr_1$	$attr_2$	$class$	$leaf_num$
65K	30	Safe	4⇒$STOP$
15K	23	Risky	3⇒$STOP$
75K	40	Safe	$STOP$
15K	28	Risky	3⇒$STOP$
100K	55	Safe	$STOP$
60K	45	Safe	$STOP$
62K	30	Risky	3⇒$STOP$

Table 15: Final relation $DETAIL$ with implicit $leaf_num$

Proof. In the case when not all dimension tables fit in memory at the same time, but each individual dimension table does, we can form, use and discard each dimension table on the fly. This can be done by a single pass through the $DETAIL$ table when $M/n > B$ (which is always true in practice).

MIND keeps a buffer of size $O(M/n)$ for each dimension. In scanning $DETAIL$, for each dimension, its buffer is used to store the accumulated information. Whenever a buffer is full, it is written to disk. When the scanning of $DETAIL$ is finished, many blocks have been obtained for each dimension based on which the final dimension table can be formed easily. For example, there might be two entries $(1, 1, 1, count_1)$, $(1, 1, 1, count_2)$ in two blocks for $attr_1$. They are corresponding to an entry with $leaf_num = 1$, $class = 1$, $attr_1 = 1$ in the final dimension table for $attr_1$ and will become a entry $(1, 1, 1, count_1 + count_2)$ in the final dimension table. All those blocks that corresponds to one dimension are collectively called an *intermediate* dimension table for that dimension.

M	size of internal memory
B	size of disk block
N	# of rows in *DETAIL*
n	# of attributes in *DETAIL* (not including class label)
C	# of distinct class labels
L	depth of the final classifier
D_k	total size of all dimension tables at depth k
V	# of distinct values for all attributes
r	size of each record in *DETAIL*
r_a	size of each attribute value in *DETAIL* (for simplicity, we assume that all attribute values are of similar size.)
r_d	size of each record in a dimension table
r_h	size of each record in a hash table (used in SPRINT)

Table 16: Parameters used in analysis

Now the intermediate dimension table for the first attribute is read into memory, summarized, and sorted into a final dimension table. Then MIND calculates the *gini* index values with respect to this dimension for each leaf node, and keeps the current minimum *gini* index value and the corresponding (*attribute_name*, *attribute_value*) pair in each leaf node. When the calculation for the first attribute is done, the in-memory dimension table is discarded. MIND repeats the same procedure for the second attribute, and so on. Finally, we get the best splits for all leaf nodes and we are ready to grow the tree one more level. The I/O cost at level k is scanning *DETAIL* once, plus writing out and reading in all the intermediate dimension tables once. We denote the total size of all intermediate dimension tables at level k by D'_k. Note that the *intermediate* dimension tables are a compressed version of the original *DETAIL* table, and they take much less space than the original records they represent. So we have

$$D'_k \leq nN.$$

The I/O cost at each level is

$$O\left(\frac{1}{B}\sum_{0\leq k<L} D'_k + \frac{LnN}{B}\right) = O\left(\frac{LnN}{B}\right).$$

In the very unlikely scenario where $M/n < B$, a total of $\log_{M/B} n$ passes over *DETAIL* are needed, resulting in a total I/O complexity in (5). □

Now let's consider the worst case in which some individual dimension tables do not fit in memory. We employ a merge sort process. An interesting point is that the merge sort process here is different from the traditional one: After several passes in the merge sort, the lengths of the runs will not increase anymore; they are upper bounded by the number of rows in the final dimension tables, whose size, although too large to fit in memory, is typically small in comparison with N.

We formally define the special sort problem. We adopt the notations used in [35]:

$$\begin{aligned} N &= \text{problem size (in units of data items)}, \\ M &= \text{internal memory size (in units of data items)}, \end{aligned}$$

$$\begin{aligned} B &= \text{block size (in units of data items)}, \\ m &= \frac{M}{B}, \text{ number of blocks that fits into internal memory}, \end{aligned}$$

where $1 \leq B \leq M < N$.

The special sort problem can be defined as follows:

Definition 1. There are $N'(N' \ll N)$ distinct keys, $\{k_1, k_2, \ldots, k_{N'}\}$, and we assume $k_1 < k_2 < \ldots < k_{N'}$ for simplicity. We have N date items $(k_{x(i)}, count_i)$, for $1 \leq i \leq N$, $1 \leq x(i) \leq N'$.

The goal is to obtain N' data items with the *key* in sorted increasing order and the corresponding *count* summarized; that is, $(k_i, COUNT_i)$, where

$$COUNT_i = \sum_{1 \leq k \leq N, x(k)=i} count_k$$

for $1 \leq i \leq N'$.

Lemma 1. *The I/O complexity of the special sort problem is*

$$O\left(\frac{N}{B} \log_{\frac{M}{B}} \frac{N'}{B}\right) \tag{6}$$

Proof. We perform a modified merge sort procedure for the special sort problem. First N/M sorted "runs" are formed by repeatedly filling up the internal memory, sorting the records according to their key values, combining the records with the same key and summarizing their counts, and writing the results to disk. This requires $O(\frac{N}{B})$ I/Os. Next m runs are continually merged and combined together into a longer sorted run, until we end up with one sorted run containing all the N' records.

In a traditional merge sort procedure, the crucial property is that we can merge m runs together in a linear number of I/Os. To do so we simply load a block from each of the runs and collect and output the B smallest elements. We continue this process until we have processed all elements in all runs, loading a new block from a run every time a block becomes empty. Since there are $O(\log_m \frac{N/B}{m})$ levels in the merge process, and each level requires $O(\frac{N}{B})$ I/O operations, we obtain the $O(\frac{N}{B} \log_m \frac{N}{B})$ complexity for the normal sort problem.

An important difference between the special sort procedure and the traditional one is that in the former, the length of each sorted run will not go beyond N' while in the latter, the length of sorted runs at each level keeps increasing (doubling) until reaching N.

In the special sort procedure, at and after level $k = \lceil \log_{M/B} N'/B \rceil$, the length of any run will be bounded by N' and the number of runs is bounded by $\lceil \frac{N/B}{m^k} \rceil$. (For simplicity, we will ignore all the floors and ceilings in the following discussion.) From level $k+1$ on, the operation we perform at each level is basically combining each m runs (each with a length less than or equal to N') into one run whose length is still bounded by N'. We repeat this operation at each level until we get a single run. At level $k+i$, we combine $\frac{N/B}{m^{k+i-1}}$ runs into $\frac{N/B}{m^{k+i}}$ runs and the I/O at this level is

$$\frac{N/B}{m^{i-1}} \left(1 + \frac{1}{m}\right).$$

We will finish the combining procedure at level $k+p$ where $p = \log_m \frac{N/B}{n'}$, $n' = N'/B$. So the I/O for the whole special sort procedure is:

$$\begin{aligned}
&2\frac{N}{B}k + \frac{N}{B}(1+\frac{1}{m}) + \frac{N/B}{m}\left(1+\frac{1}{m}\right) + \cdots + \frac{N/B}{m^{p-1}}\left(1+\frac{1}{m}\right) \\
&\leq 2\frac{N}{B}\log_m n' + \frac{N}{B}\left(1+\frac{1}{m}\right)\frac{1}{1-1/m} \\
&\approx 2\frac{N}{B}\log_m n' + \frac{N}{B} \\
&= O\left(\frac{N}{B}\log_m n' + \frac{N}{B}\right) \\
&= O\left(\frac{N}{B}\log_{\frac{M}{B}} \frac{N'}{B}\right).
\end{aligned}$$

□

Now we are ready to give the I/O complexity of MIND in the worst case.

Theorem 1. *In the worst case the I/O complexity of MIND is*

$$O\left(\frac{nNL}{B} + \frac{nN}{B}\sum_{0\leq k<L} \log_{M/B} \frac{D_k}{B}\right), \tag{7}$$

which is

$$O\left(\frac{LnN}{B}\frac{\log \frac{nN}{B}}{\log \frac{M}{B}}\right). \tag{8}$$

In most applications, the log *term is negligible, and the I/O complexity of MIND becomes*

$$O\left(\frac{LnN}{B}\right), \tag{9}$$

which matches the optimal time of (4).

Proof. This is similar to the proof in Claim 2. At level k of the tree growth phase, MIND first forms all the intermediate dimension tables with total size $D_k^{'}$ in external memory. This can be done by a single pass through the *DETAIL* table, as follows. MIND keeps a buffer of size $O(M/n)$ for each dimension. In scanning *DETAIL*, MIND accumulates information for each dimension in its corresponding buffer; whenever a buffer is full, it is written to disk. When the scanning of *DETAIL* is finished, MIND performs the special merge sort procedure for the disk blocks corresponding to all (not individual) dimension tables. At the last level of the special sort, the final dimension table for each attribute will be formed one by one. MIND calculates the *gini* index values with respect to each dimension for each leaf node, and keeps the current minimum *gini* index value and the corresponding (*attribute_name*, *attribute_value*) pair in each leaf node. When the calculation for the last attribute is done, we get the best splits for all leaf nodes and we are ready to grow the tree one more level.

The I/O cost at level k is scanning *DETAIL* once, which is $O(nN/B)$, plus the cost of writing out all the *intermediate* dimension tables once, which is bounded by $O(nN/B)$, plus the cost for the special sort, which is $O(\frac{N}{B}\log_{M/B} D_k/B)$.

So the I/O for all levels is

$$\frac{LnN}{B} + \frac{1}{B}\sum_{0\le k<L} D_k^{'} + \frac{nN}{B}\sum_{0\le k<L} \log_{M/B}\frac{D_k}{B}$$

which is

$$O\left(\frac{LnN}{B} + \frac{nN}{B}\sum_{0\le k<L} \log_{M/B}\frac{D_k}{B}\right).$$

□

Now we analyze the I/O complexity of the SPRINT algorithm. There are two major parts in SPRINT: the pre-sorting of all attribute lists and the constructing/searching of the corresponding hash tables during partition. Since we are dealing with a very large *DETAIL* table, it is unrealistic to assume that N is small enough to allow hash tables to be stored in memory. Actually those hash tables need to be stored on disk and brought into memory during the partition phase. It is true that hash tables will become smaller at deeper levels and thus fit in memory, but at the early levels they are very large; for example, the hash table at level 0 has N entries.

Each entry in a hash table contains a *tid*(transaction identifier) which is an integer in the range of 1 to N, and one bit that indicates which child this record should be partitioned to in the next level of the classifier. So we have

$$r_h = \frac{1+\log N}{8}.$$

We can estimate when the hash tables will fit in memory, given the optimistic assumptions that all memory is allocated to hash tables and all hash tables at each node have equal size; that is, a hash table at level k contains $N/2^k$ entries. Thus, a hash table at level k fits in memory if $r_h N/2^k \le M$, or

$$2^k \ge \frac{N}{M}\left(\frac{1+\log N}{8}\right). \tag{10}$$

For sufficiently large k, (10) will be satisfied, that is, hash tables become smaller at deeper nodes and thus fit in memory. But it is clear that even for moderately large detail tables, hash tables at upper levels will not fit in memory.

During the partition phase, each non-splitting attribute list at each node needs to be partitioned into two parts based on the corresponding hash table. One way to do this is to do a random hash table search for each entry in the list, but this is very expensive. Fortunately, there is a better way: First, we bring a large portion of the hash table into memory. The size of this portion is limited only by the availability of the internal memory. Then we scan the non-splitting list once, block by block, and for each entry in the list, we search the in-memory portion of the hash table. In this way, the hash table is swapped into memory only once, and each non-splitting attribute list is scanned N/M times. For even larger N, it is better to do the lookup by batch sorting, but that approach is completely counter to the founding philosophy of the SPRINT algorithm.

A careful analysis gives us the following estimation:

Theorem 2. *The I/O complexity of SPRINT is*

$$O\left(\frac{nN^2\log N}{BM}\right) \tag{11}$$

Proof. To perform the pre-sort of the SPRINT algorithm, we need to read *DETAIL* once, write out the unsorted attribute lists, and sort all the attribute lists. So we have

$$IO_{presort} = O\left(\frac{nN}{B}\log_{\frac{M}{B}}\frac{N}{B}\right).$$

From level 0 through level $k-1$, hash tables will not fit in memory. At level i $(0 \le i \le k-1)$, SPRINT will perform the following operations:

1. Scan the attribute lists one by one to find the best split for each leaf node.
2. According to the best split found for each leaf node, form the hash tables and write them to disk.
3. Partition the attribute list of the splitting attribute for each leaf node.
4. Partition the attribute lists for the $n-1$ non-splitting attributes for each leaf node.

Among these operations, the last one incurs the most I/O cost and we perform it by bringing a portion of a hash table into memory first. The size of this portion is limited only by the availability of the main memory. Then we scan each non-splitting list once, block by block, and for each entry in the list, we search the in-memory portion of the hash table and decide which child this entry should go in the next level. In this way, the hash table is swapped into memory only once, and the non-splitting list is scanned multiple times. The I/O cost of this operation is

$$O\left(\frac{nNh_i}{B}\right)$$

where h_i is the number of portions we need to partition a hash table into due to the limitation of the memory size.

From level k to level L the hash table will fit in memory, and the I/O costs for those levels is $O((L-k)nN/B)$, which is significantly smaller than those for the previous levels.

So the I/O cost of SPRINT becomes

$$O\left(\frac{nN}{B}\log_{\frac{M}{B}}\frac{N}{B} + \sum_{0\le i\le k-1}\frac{nNh_i}{B} + \frac{(L-k)nN}{B}\right) \tag{12}$$

Note that we have

$$h_i = \frac{r_h N}{2^i M} = \frac{N}{2^i M}\left(\frac{1+\log N}{8}\right)$$

So

$$\frac{N}{M}\left(\frac{1+\log N}{8}\right) \le \sum_{0\le i\le k-1} h_i \le \frac{2N}{M}\left(\frac{1+\log N}{8}\right) \tag{13}$$

Applying (13) to (12), we get the I/O complexity of SPRINT in (11). □

Examination of (8) and (11) reveals that MIND is clearly better in terms of I/O performance. For large N, SPRINT does a quadratic number of I/Os, whereas MIND scales well.

6. Algorithm revisited using schema SQL

In Section 3.1, we described the MIND algorithm using SQL-like statements. Due to the limitation of current SQL standards, most of those SQL-like statements are not supported directly in today's DBMS products. Therefore, we need to convert them to currently supported SQL statements, augmented with new facilities like user defined functions. Putting logic within a user-defined function hides the operator from query optimization. If classification was a subquery or part of a large query, it would not be possible to obtain all join reorderings, thereby risking suboptimal execution.

Current SQL standards are mainly designed for efficient OLTP (On-Line Transactional Processing) queries. For non-OLTP applications, it is true that we can usually reformulate the problem and express the solution using standard SQL. However, this approach often results in inefficiency. Extending current SQL with ad-hoc constructs and new optimization considerations might solve this problem in some particular domain, but it is not a satisfactory solution. Since supporting OLAP (On-Line Analytical Processing) applications efficiently is such an important goal for today's RDBMSs, the problem deserves a more general solution.

In [23] an extension of SQL, called SchemaSQL, is proposed. SchemaSQL offers the capability of uniform manipulation of data and meta-data in relational multidatabase systems. By examining the SQL-like queries in Section 3.1, we can see that this capability is what we need in the MIND algorithm. To show the power of extended SQL and the flexibility and general flavor of MIND, in this section, we rewrite all the queries in Section 3.1 using SchemaSQL.

First we give an overview of the syntax of SchemaSQL. For more details see [23].

In a standard SQL query, the *tuple* variables are declared in the `FROM` clause. A variable declaration has the form $\langle range \rangle \langle var \rangle$. For example, in the query below, the expression *student* T declares T as a variable that ranges over the (tuples of the) relation $student(student_id, department, GPA)$:

`SELECT` $student_id$
`FROM` $student\ T$
`WHERE` $T.department = CS$ `AND` $T.GPA = A$

The SchemaSQL syntax extends SQL syntax in several directions:

1. The federation consists of databases, with each database consisting of relations.
2. To permit meta-data queries and reconstruction views, SchemaSQL permits the declaration of other types of variables in addition to the tuple variables permitted in SQL.
3. Aggregate operations are generalized in SchemaSQL to make horizontal and block aggregations possible, in addition to the usual vertical aggregation in SQL.

SchemaSQL permits the declaration of variables that can range over any of the following five sets:

1. names of databases in a federation,
2. names of the relations in a database,

3. names of the columns in the scheme of a relation,

4. tuples in a given relation in database, and

5. values appearing in a column corresponding to a given column in a relation.

Variable declarations follow the same syntax as $\langle range \rangle \langle var \rangle$ as in SQL, where *var* is any identifier. However, there are two major differences:

1. The only kind of *range* permitted in SQL is a set of tuples in some relation in the database, where in SchemaSQL any of the five kinds of *range* can be used to declare variables.

2. The *range* specification in SQL is made using constant, i.e., an identifier referring to a specific relation in a database. By contrast, the diversity of *range*s possible in SchemaSQL permits *range* specifications to be *nested*, in the sense that it is possible to say, for example, that R is a variable ranging over the relation names in database D, and that T is a tuple in the relation denoted by R.

Range specifications are one of the following five types of expressions, where *db*, *rel*, *col* are any *constant* or *variable identifiers*.

1. The expression $\rightarrow$ denotes a *range* corresponding to the set of database names in the federation.

2. The expression $db \rightarrow$ denotes the set of relation names in the database *db*.

3. The expression $db :: rel \rightarrow$ denotes the set of names of column in the schema of the relation *rel* in the database *db*.

4. $db :: rel$ denotes the set of tuples in the relation *rel* in the database *db*.

5. $db :: rel.col$ denotes the set of values appearing in the column named *col* in the relation *rel* in the database *db*.

For example, consider the clause `FROM` $db1 \rightarrow R$, $db1 :: R\ T$. It declares R as a variable ranging over the set of relation names in the database $db1$ and T as a variable ranging over the tuples in each relation R in the database $db1$

Now we are ready to rewrite all the SQL-like queries in Section 3.1 using SchemaSQL. Assume that our training set is stored in relation *DETAIL* in a database named $FACT$. We first generate all the dimension tables with the schema $(leaf_num, class, attr_val, count)$ in a database named *DIMENSION*, using a simple SchemaSQL statement:

```
CREATE VIEW DIMENSION :: R(leaf_num, class, attr_val, count) AS
SELECT T.leaf_num, T.class, T.R, COUNT(*)
FROM FACT :: DETAIL → R,
     FACT :: DETAIL T
WHERE R <>' class' AND
      R <>' leaf_num' AND
      T.leaf_num <> STOP
GROUP BY T.leaf_num, T.class, T.R
```

The variable R is declared as a column name variable ranging over the column names of relation *DETAIL* in the database *FACT*, and the variable T is declared as a tuple variable on the same relation. The conditions on R in the `WHERE` clause make the variable R range over all columns except the columns named *class* and *leaf_num*. If there are n columns in *DETAIL* (excluding columns *class* and *leaf_num*), this query generates n `VIEW`s in database *DIMENSION*, and the name of each `VIEW` is the same as the corresponding column name in *DETAIL*. Note that the attribute name to relation name transformation is done in a very natural way, and the formation of multiple `GROUP BY`s is done by involving *DETAIL* only once.

Those views will be materialized, so that in the later operations we do not need to access *DETAIL* any more.

Relations corresponding to *UP* with the schema (*leaf_num*, *attr_val*, *class*, *count*) can be generated in a database named *UP* by performing a self-outer-join on dimension tables in database *DIMENSION*:

```
CREATE VIEW UP :: R(leaf_num, attr_val, class, count) AS
SELECT d1.leaf_num, d1.attr_val, d1.class, SUM(d2.count)
FROM (FULL OUTER JOIN DIMENSION :: R d1,
                      DIMENSION :: R d2,
                      DIMENSION → R
      ON d1.leaf_num = d2.leaf_num AND
         d1.attr_val ≤ d2.attr_val AND
         d1.class = d2.class
      GROUP BY d1.leaf_num, d1.attr_val, d1.class)
```

The variable R is declared as a relation name variable ranging over all the relations in database *DIMENSION*. Variables d_1 and d_2 are both tuple variables over the tuples in each relation R in database *DIMENSION*. For each relation in database *DIMENSION*, a self-outer-join is performed according to the conditions specified in the query, and the result is put into a `VIEW` with the same name in database *UP*.

Similarly, relations corresponding to *DOWN* can be generated in a database named *DOWN* by just changing the $\leq$ to $>$ in the `ON` clause.

Database *DOWN* and database *UP* contain all the information we need to compute all the *gini* index values. Since standard SQL only allows vertical aggregations, we need to rearrange them before the *gini* index is actually calculated as in Section 3.1. In SchemaSQL, aggregation operations are generalized to make horizontal and block aggregations possible. Thus, we can generate views that contain all *gini* index values at each possible split point for each attribute in a database named *GINI_VALUE* directly from relations in *UP* and *DOWN*:

```
CREATE VIEW GINI_VALUE :: R(leaf_num, attr_val, gini) AS
SELECT u.leaf_num, u.attr_val, f_gini
FROM UP :: R u,
     DOWN :: R d,
     UP → R
WHERE u.leaf_num = d.leaf_num AND
      u.attr_val = d.attr_val
GROUP BY u.leaf_num, u.attr_val
```

where f_{gini} is a function of *u.class*, *d.class*, *u.count*, *d.count* according to (1) and (2).

R is declared as a variable ranging over the set of relation names in database UP, u is a variable ranging over the tuples in each relation in database UP, and d is a variable ranging over the tuples in the relation with the same name as R in database $DOWN$. Note that the set of relation names in databases UP and $DOWN$ are the same. For each of the relation pairs with the same name in UP and $DOWN$, this statement will create a view with the same name in database $GINI_VALUE$ according to the conditions specified. It is interesting to note that f_{gini} is a block aggregation function instead of the usual vertical aggregation function in SQL. Each view named R in database $GINI_VALUE$ contains the *gini* index value at each possible split point with respect to attribute named R.

Next, we create a *single* view MIN_GINI with the schema $MIN_GINI(leaf_num, attr_name, attr_val, gini)$ in a database named $SPLIT$ form the *multiple* views in database $GINI_VALUE$:

```
CREATE VIEW SPLIT :: MIN_GINI(leaf_num, attr_name, attr_val, gini) AS
SELECT T1.leaf_num, R1, T1.attr_val, gini
FROM GINI_VALUE → R1,
     GINI_VALUE :: R1 T1
WHERE T1.gini =(SELECT MIN(T2.gini)
                FROM GINI_VALUE → R2,
                     GINI_VALUE :: D2 T2
                WHERE R1 = R2 AND
                      T1.leaf_num = T2.leaf_num)
```

R_1 and R_2 are variables ranging over the set of relation names in database $GINI_VALUE$. T_1 and T_2 are tuple variables ranging over the tuples in relations specified by R_1 and R_2, respectively. The clause $R_1 = R_2$ enforces R_1 and R_2 to be the same relation. Note that relation name R_1 in database $GINI_VALUE$ becomes the column value for the column named $attr_name$ in relation MIN_GINI in database $SPLIT$. Relation MIN_GINI now contains the best split value and the corresponding *gini* index value for each leaf node of the tree with respect to all attributes.

The overall best split for each leaf is obtained from executing the following:

```
CREATE VIEW SPLIT :: BEST_SPLIT( leaf_num, attr_name, attr_val) AS
SELECT T1.leaf_num, T1.attr_name, T1.attr_val
FROM SPLIT :: MIN_GINI T1
WHERE T1.gini =(SELECT MIN(gini)
                FROM SPLIT :: MIN_GINI T2
                WHERE T1.leaf_num = T2.leaf_num)
```

This statement is similar to the statement generating relation $BEST_SPLIT$ in Section 3.1. T_1 is declared as a tuple variable ranging over the tuples of relation MIN_GINI in database $SPLIT$. For each $leaf_num$, ($attr_name$, $attr_val$) pair that achieving the minimum *gini* index value is inserted into relation $BEST_SPLIT$.

We have shown how to rewrite all the SQL-like queries in MIND algorithm using SchemaSQL. In our current prototype of MIND, the first step, generating all the dimension tables from $DETAIL$, is most costly and all the later steps only need to access small dimension tables. We use udf to reduce the cost of the first step. All the SQL-like queries in Section 3.1 in the later steps are translated into equivalent SQL queries. Those translations usually lead to poor performance. But since those queries only access small relations in MIND, the performance loss is negligible.

While udf provides a solution to our classification algorithm, we a believe general extension of SQL is needed for efficient support of OLAP applications.

An alternative way to generate all the dimension tables from *DETAIL* would be using the newly proposed *data cube* operator [14] since dimension tables are different subcubes. But it usually takes a long time to generate the data cube without precomputation and the fact that the *leaf_num* column in *DETAIL* keeps changing from level to level when we grow the tree makes precomputation infeasible.

7. Experimental results

There are two important metrics to evaluate the quality of a classifier: *classification accuracy* and *classification time*. We compare our results with those of SLIQ [24] and SPRINT [29]. (For brevity, we include only SPRINT in this paper; comparisons showing the improvement of SPRINT over SLIQ are given in [29].) Unlike SLIQ and SPRINT, we use the classical database methodology of summarization. Like SLIQ and SPRINT, we use the same metric (*gini* index) to choose the best split for each node, we grow our tree in a breadth-first fashion, and we prune it using the same pruning algorithm. Our classifier therefore generates a decision tree identical to the one produced by [24, 29] for the same training set, which facilitates meaningful comparisons of run time. The accuracy of SPRINT and SLIQ is discussed in [24, 29], where it is argued that the accuracy is sufficient.

For our scaling experiments, we ran our prototype on large data sets. The main cost of our algorithm is that we need to access *DETAIL* n times (n is the number of attributes) for each level of the tree growth due to the absence of the multiple `GROUP BY` operator in the current SQL standard. We recommend that future DBMSs support the multiple `GROUP BY` operator so that *DETAIL* will be accessed only once regardless of the number of attributes. In our current working prototype, this is done by using user-defined function as we described in Section 3.1.

Owing to the lack of a classification benchmark, we used the synthetic database proposed in [1]. In this synthetic database, each record consists of nine attributes as shown in Table 17. Ten classifier functions are proposed in [1] to produce databases with different complexities. We run our prototype using function 2. It generates a database with two classes: Group A and Group B. The description of the class predicate for Group A is shown below.

Function 2, Group A

$((\text{age} < 40) \wedge (50\text{K} \leq \text{salary} \leq 100\text{K})) \vee$
$((40 \leq \text{age} < 60) \wedge (75\text{K} \leq \text{salary} \leq 125\text{K})) \vee$
$((\text{age} \geq 60) \wedge (25\text{K} \leq \text{salary} \leq 75\text{K}))$

Our experiments were conducted on an IBM RS/6000 workstation running AIX level 4.1.3. and DB2 version 2.1.1. We used training sets with sizes ranging from 0.5 million to 5 million records. The relative response time and response time per example are shown in Figure 5 and Figure 6 respectively. Figure 5 hints that our algorithm achieves linear scalability with respect to the training set size. Figure 6 shows that the time per example curve stays flat when the training set size increases. The corresponding curve for [29] appears to be growing slightly on the largest cases. Figure 7 is the performance comparison between MIND and SPRINT. MIND ran on a processor with a slightly slower clock rate. We can see that MIND performs better than SPRINT does even in the range where SPRINT scales well, and MIND continues to scale well as the data sets get larger.

We also ran MIND on an IBM multiprocessor SP2 computer system. Figure 8 shows the parallel speedup of MIND.

attribute	*value*
salary	uniformly distributed from 20K to 150K
commission	$salary \geq 75K \Rightarrow commission = 0$ else uniformly distributed from 10K to 75K
age	uniformly distributed from 20 to 80
loan	uniformly distributed from 0 to 500K
elevel	uniformly chosen from 0 to 4
car	uniformly chosen form 1 to 20
zipcode	uniformly chosen from 10 available zipcodes
hvalue	uniformly distributed from $0.5k100000$ to $1.5k100000$, where $k \in \{0, \ldots, 9\}$ is zipcode
hyear	uniformly distributed from 1 to 30

Table 17: Description of the synthetic data

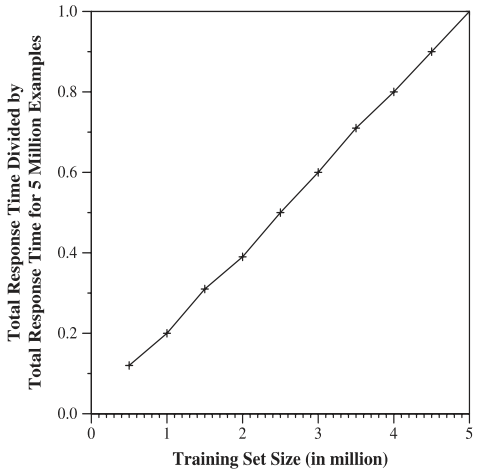


Figure 5: Relative total response time. The y-value denotes the total response time for the indicated training set size, divided by the total response time for 5 million examples.

Another interesting measurement we obtained from uniprocessor execution is that accessing *DETAIL* to form the dimension tables for all attributes takes 93%–96% of the total execution time. To achieve linear speedup on multiprocessors, it is critical that this step is parallelized. In the current working prototype of MIND, it is done by user-defined function with a scratch-pad accessible from multiple processors.

8. Conclusions

The MIND algorithm solves the problem of classification within the relational database management systems. Our performance measurements show that MIND demonstrates scalability with respect to the number of examples in training sets and the number of parallel processors. We believe MIND is the first classifier to successfully run on datasets of $N = 5$ million examples on a uniprocessor and

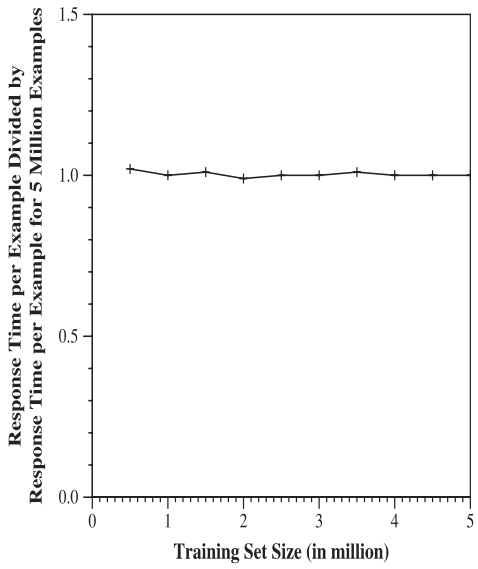


Figure 6: Relative response time per example. The y-value denotes the response time per example for the indicated training set size, divided by response time per example when processing 5 million examples.

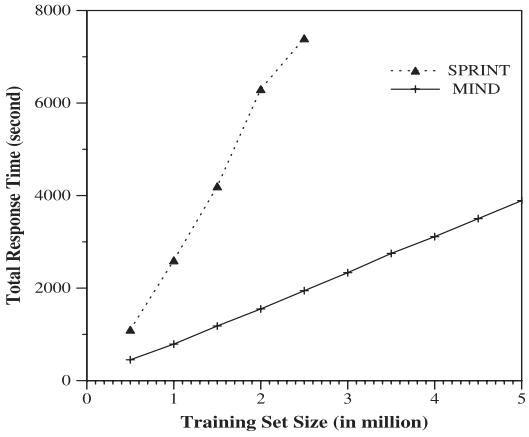


Figure 7: Performance comparison of MIND and SPRINT

yet demonstrate effectively non-increasing response time per example as a function of N. It also runs faster than previous algorithms on file systems.

There are four reasons why MIND is fast, exhibits excellent scalability, and is able to handle data sets larger than those tackled before:

1. MIND rephrases the data mining function classification as a classic DBMS problem of summarization and analysis thereof.

2. MIND avoids any update to the *DETAIL* table of examples. This is of significant practical interest; for example, imagine *DETAIL* having billions of rows.

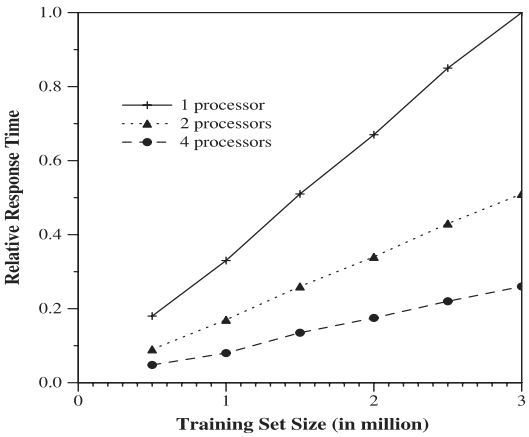


Figure 8: Speedup of MIND for multiprocessors. The y-value denotes the total response time for the indicated training set size, divided by the total response time for 3 million examples.

3. In the absence of a multiple concurrent grouping SQL operator, MIND takes advantage of the user-defined function capability of DB2 to achieve the equivalent functionality and the resultant performance gain.

4. Parallelism of MIND is obtained at little or no extra cost because the RDBMS parallelizes SQL queries.

We recommend that extensions be made to SQL to do multiple groupings and the streaming of each group to different relations. Most DBMS operators currently take two streams of data (tables) and combine them into one. We believe that we have shown the value of an operator that takes a single stream input and produces multiple streams of outputs.

References

[1] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami. An interval classifier for database mining applications. In *Proceedings of the 1992 International Conference on Very Large Databases*, pages 560–573, Vancouver, Canada, August 1992.

[2] R. Agrawal and K. Shim. Developing tightly-coupled data mining applications on a relational database system. In *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining*, August 1996.

[3] P. Adrians and D. Zantinge. *Data Mining.* Addison-Wesley, 1996.

[4] L. Breiman et al. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

[5] C. K. Baru et al. DB2 parallel edition. *IBM Systems Journal*, 34(2), 1995.

[6] G. Bhargava, P. Goel, and B. Iyer. Hypergraph based reordering of outer join queries with complex predicates. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, 1995.

[7] R. D. Barve, E. F. Grove, and J. S. Vitter. Simple randomized mergesort on parallel disks. *Parallel Computing*, 23(4), 1997. MR1453082

[8] D. Chamberlin et al. Seqel: A structured english query language. In *Proc. of ACM SIGMOD Workshop on Data Description, Access, and Control*, May 1974.

[9] J. Catlett. *Megainduction: Machine Learning on Very Large Databases*. PhD thesis, University of Sydney, 1991.

[10] D. Chamberlin. *Using the New DB2: IBM's Object-Relational Database System*. Morgan Kaufmann, 1996.

[11] D. Chamberlin. Personal communication, 1997.

[12] E. F. Codd. A relational model of data for large shared data banks. *CACM*, 13(6), June 1970.

[13] T. Dietterich, M. Kearns, and Y. Mansour. Applying the weak learning framework to understand and improve C4.5. In *Proceedings of the 13th International Conference on Machine Learning*, pages 96–104, 1996.

[14] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tabs and subtotals. In *Proceedings of the 12th Annual IEEE Conference on Data Engineering (ICDE '96)*, pages 131–139, 1996.

[15] R. A. Ganski and H. K. T. Wong. Optimization of nested sql queried revisited. In *Proceeding of the 1987 ACM SIGMOD International Conference on Management of Data*, 1987.

[16] S. Hasty. Mining databases. *Apparel Industry Magazine*, 57(5), 1996.

[17] IBM. *IBM DATABASE 2 Application Programming Guide-for common servers*, version 2 edition.

[18] IBM Germany. *IBM Intelligence Miner User's Guide*, version 1 edition, July 1996.

[19] T. Imielinsk and H. Mannila. A database perspective on knowledge discovery. *Communication of the ACM*, 39(11), November 1996.

[20] T. Imielinski. From file mining to database mining. In *Proceedings of the 1996 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, May 1996.

[21] M. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. In *Proceedings of the 28th ACM Symposium on the Theory of Computing*, pages 459–468, 1996. MR1427544

[22] H. Lu et al. On preprocessing data for efficient classification. In *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, May 1996.

[23] L. Lakshmanan, F. Sadri, and I. Subramanian. SchemaSQL–a language for interoperability in relational multi-database systems. In *Proceedings of the 1996 International Conference on Very Large Databases*, 1996.

[24] M. Mehta, R. Agrawal, and J. Rissanen. SLIQ: A fast scalable classifier for data mining. In *Proceedings of the 5th International Conference on Extending Database Technology*, Avignon, France, March 1996.

[25] H. Messatfa. Personal communications, 1997.

[26] S. K. Murthy. *On Growing Better Decision Trees from Data*. PhD thesis, Johns Hopkins University, 1995.

[27] T. Nguyen and V. Srinivasan. Accessing relational databases from the world web web. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, 1996.

[28] J. Ross Quilan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.

[29] J. C. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. In *Proceedings of the 1996 International Conference on Very Large Databases*, Mumbai (Bombay), India, September 1996.

[30] S. Sarawagi. Query processing in tertiary memory databases. In *Proceedings of the 1995 International Conference on Very Large Databases*, 1995.

[31] J. B. Sinclair. Personal communication, 1997.

[32] M. Stonebraker and L. A. Rowe. The design of postgres. In *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data*, 1986.

[33] S. Sarawagi and M. Stonebraker. Benefits of reordering execution in tertiary memory databases. In *Proceedings of the 1996 International Conference on Very Large Databases*, 1996.

[34] J. Ullman. *Principles of Database Systems*. Computer Science Press, second edition, 1982. MR669881

[35] J. S. Vitter. External memory algorithms and data structures: Dealing with MASSIVE DATA. *ACM Computing Surveys*, 33(2):209–271, June 2001. MR1965532

[36] D. E. Vengroff and J. S. Vitter. I/O-efficient scientific computation using TPIE. In *Proceedings of the Goddard Conference on Mass Storage Systems and Technologies*, NASA Conference Publication 3340, Volume II, pages 553–570, College Park, MD, September 1996.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 378–384
© Institute of Mathematical Statistics, 2004

A simple proof of a condition for cointegration

T. W. Anderson[1,*]

Stanford University

Abstract: A simple proof is given for a theorem concerning the first difference and some linear functions of a cointegrated autoregressive process being stationary.

1. Introduction

Many macroeconometric models are formulated in terms of autoregressive processes or autoregressive processes with moving average innovations. The most appropriate process in a given situation may not be stationary, but some linear relations of the components may be stationary; such a process is called cointegrated. Johansen (1995) has given alternative conditions for the cointegrated components and first differences of other components to be stationary. Here we give a proof of one condition that is more straightforward and transparent than what is in the literature.

A p-dimensional m-order autoregressive process $\{\mathbf{Y}_t\}$ is defined by

$$\mathbf{Y}_t = \mathbf{B}_1\mathbf{Y}_{t-1} + \mathbf{B}_2\mathbf{Y}_{t-2} + \ldots + \mathbf{B}_m\mathbf{Y}_{t-m} + \mathbf{Z}_t, \tag{1.1}$$

where the $\mathbf{Z}_t$'s are independent unobservable innovations with $\mathcal{E}\mathbf{Z}_t = \mathbf{0}$, $\mathcal{E}\mathbf{Z}_t\mathbf{Z}_t' = \mathbf{\Sigma}$, and $\mathcal{E}\mathbf{Z}_t\mathbf{Y}_{t-s}' = \mathbf{0}$, $0 < s$. Let

$$\mathbf{B}(\lambda) = \lambda^m\mathbf{I}_p - \lambda^{m-1}\mathbf{B}_1 - \ldots - \mathbf{B}_m, \tag{1.2}$$

and let the roots of $|\mathbf{B}(\lambda)| = 0$ be λ_i, $i = 1, \ldots, mp$. If $|\lambda_i| < 1$, $i = 1, \ldots, mp$, the process $\{\mathbf{Y}_t\}$ may be stationary. If one or more of the roots are 1, the process is nonstationary, but some order of differencing may yield a stationary process. When some linear functions of a nonstationary process are stationary, the model is called *cointegrated*. We call a process defined by the Equation (1.1) *stationary* if it is possible to assign a distribution to $(\mathbf{Y}_{-m+1}, \ldots, \mathbf{Y}_{-1}, \mathbf{Y}_0)$ such that (1.1) generates a process $\mathbf{Y}_{-m+1}, \mathbf{Y}_{-m+2}, \ldots$ that is stationary. Throughout this paper it is assumed that n of the roots are 1 and the other roots satisfy $|\lambda_i| < 1$, $i = n+1, \ldots, mp$.

An "error-correction form" of the autoregressive process is

$$\Delta\mathbf{Y}_t = \mathbf{\Pi}\mathbf{Y}_{t-1} + \mathbf{\Pi}_1\Delta\mathbf{Y}_{t-1} + \ldots + \mathbf{\Pi}_{m-1}\Delta\mathbf{Y}_{t-m+1} + \mathbf{Z}_t, \tag{1.3}$$

where $\Delta\mathbf{Y}_t = \mathbf{Y}_t - \mathbf{Y}_{t-1}$,

$$\mathbf{\Pi}_j = -(\mathbf{B}_{j+1} + \ldots + \mathbf{B}_m), \qquad j = 1, \ldots, m-1, \tag{1.4}$$

$$\mathbf{\Pi} = \mathbf{B}_1 + \mathbf{B}_2 + \ldots + \mathbf{B}_m - \mathbf{I}_p. \tag{1.5}$$

Note that $\mathbf{\Pi}_j = \mathbf{\Pi}_{j+1} - \mathbf{B}_{j+1}$ and $\mathbf{\Pi} = -\mathbf{B}(1)$.

[*] This paper is dedicated to my friend and co-author Herman Rubin, who stimulated and educated me as well as collaborated with me.

[1] Departments of Economics and Statistics, Stanford University, Stanford, CA 94305, USA. e-mail: twa@stat.stanford.edu

Keywords and phrases: autoregressive process, error correction form, stationarity.
AMS 2000 subject classifications: primary 62P20; secondary 60G12.

Suppose the rank of $\mathbf{\Pi}$ is k. Then $\mathbf{\Pi}$ can be written $\mathbf{\Pi} = \mathbf{A}\mathbf{\Gamma}'$, where $\mathbf{A}$ and $\mathbf{\Gamma}$ are $p \times k$ matrices of rank k. Let $\mathbf{A}_\perp$ and $\mathbf{\Gamma}_\perp$ be $p \times (p-k)$ matrices of rank $p-k$ such that $\mathbf{A}'_\perp \mathbf{A} = \mathbf{0}$ and $\mathbf{\Gamma}'_\perp \mathbf{\Gamma} = \mathbf{0}$. Then a necessary and sufficient condition that $\Delta \mathbf{Y}_t$ and $\mathbf{\Gamma}' \mathbf{Y}_t$ are stationary is that

$$\operatorname{rank} \left[\mathbf{A}'_\perp \left(\mathbf{I} - \sum_{i=1}^{m-1} \mathbf{\Pi}_i \right) \mathbf{\Gamma}_\perp \right] = p - k \tag{1.6}$$

[Theorem 4.2, Johansen (1995)]. The proof of this statement involves an expansion of $\mathbf{B}(\lambda)$ around $\lambda = 1$.

If $\{\mathbf{Y}_t\}$ is stationary, it is said to be $I(0)$. If $\{\mathbf{Y}_t\}$ is not $I(0)$, but $\{\Delta \mathbf{Y}_t\}$ is stationary, the process $\{\mathbf{Y}_t\}$ is said to be $I(1)$.

Corollary 4.3 of Johansen asserts that if k is the rank of $\mathbf{\Pi}$ and $k < p$, then the multiplicity of $\lambda = 1$ as a zero of $|\mathbf{B}(\lambda)|$ is equal to $p - k$ if and only if $\{\mathbf{Y}_t\}$ is $I(1)$. The proof of this statement depends on his Theorem 4.2 and its proof.

In this paper the condition is formulated as

Rank Condition. *There are n linearly independent solutions to*

$$\boldsymbol{\omega}' \mathbf{\Pi} = \mathbf{0}, \tag{1.7}$$

where n is the multiplicity of $\lambda = 1$ as a root of the characteristic equation $|\mathbf{B}(\lambda)| = 0$.

Let n independent solutions of (1.7) be assembled into the matrix $\mathbf{\Omega}_1 = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$; then $\mathbf{\Omega}'_1 \mathbf{\Pi} = \mathbf{0}$ and the rank of $\mathbf{\Omega}_1$ is n.

2. First-order case

First we treat the special case of $m = 1$. Then (1.1) is

$$\mathbf{Y}_t = \mathbf{B}_1 \mathbf{Y}_{t-1} + \mathbf{Z}_t; \tag{2.1}$$

the error-correction form is

$$\Delta \mathbf{Y}_t = \mathbf{\Pi} \mathbf{Y}_{t-1} + \mathbf{Z}_t, \tag{2.2}$$

where $\mathbf{\Pi} = \mathbf{B}_1 - \mathbf{I}_p$; and $\mathbf{B}(\lambda) = \lambda \mathbf{I}_p - \mathbf{B}_1$.

Theorem 1 ($m = 1$)**.** *Suppose the Rank Condition holds. Then the rank of $\mathbf{\Pi}$ is $k = p - n$, and there exists a $p \times k$ matrix $\mathbf{\Omega}_2$ such that*

$$\mathbf{\Omega}'_2 \mathbf{\Pi} = \mathbf{\Upsilon}_{22} \mathbf{\Omega}'_2, \tag{2.3}$$

$\mathbf{\Upsilon}_{22}$ $(k \times k)$ is nonsingular, and $\mathbf{\Omega} = (\mathbf{\Omega}_1, \mathbf{\Omega}_2)$ is nonsingular. Define

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{X}_{1t} \\ \mathbf{X}_{2t} \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}'_1 \mathbf{Y}_t \\ \mathbf{\Omega}'_2 \mathbf{Y}_t \end{bmatrix}, \qquad \mathbf{W}_t = \begin{bmatrix} \mathbf{W}_{1t} \\ \mathbf{W}_{2t} \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}'_1 \mathbf{Z}_t \\ \mathbf{\Omega}'_2 \mathbf{Z}_t \end{bmatrix}. \tag{2.4}$$

Then $\Delta \mathbf{X}_{1t}, \mathbf{X}_{2t}$ defines a stationary process.

Proof. Let $\mathbf{\Omega}'_1 = (\mathbf{I}_n, \mathbf{\Omega}'_{21})$ and $\mathbf{\Pi}' = (\mathbf{\Pi}'_1, \mathbf{\Pi}'_2)$, where $\mathbf{\Pi}_2$ is $k \times p$. (The rows of $\mathbf{\Omega}_1$ and the columns of $\mathbf{\Pi}$ can be ordered so that $\mathbf{\Omega}_{11}$ is nonsingular and can be set as $\mathbf{I}_n$.) Then the Rank Condition is

$$\mathbf{0} = \mathbf{\Omega}'_1 \mathbf{\Pi} = \left(\mathbf{I}_n, \mathbf{\Omega}'_{21} \right) \begin{bmatrix} \mathbf{\Pi}_1 \\ \mathbf{\Pi}_2 \end{bmatrix} = \mathbf{\Pi}_1 + \mathbf{\Omega}'_{21} \mathbf{\Pi}_2, \tag{2.5}$$

which implies $\mathbf{\Pi}_1 = -\mathbf{\Omega}_{21}'\mathbf{\Pi}_2$ and

$$\mathbf{\Pi} = \begin{bmatrix} -\mathbf{\Omega}_{21}' \\ \mathbf{I}_k \end{bmatrix} \mathbf{\Pi}_2. \tag{2.6}$$

Define $\mathbf{\Omega}_2 = \mathbf{\Pi}_2'$ $(p \times k)$ and

$$\mathbf{\Upsilon}_{22} = \mathbf{\Pi}_2 \begin{bmatrix} -\mathbf{\Omega}_{21}' \\ \mathbf{I}_k \end{bmatrix} = \mathbf{\Omega}_2' \begin{bmatrix} -\mathbf{\Omega}_{21}' \\ \mathbf{I}_k \end{bmatrix}. \tag{2.7}$$

Then (2.3) is satisfied. Note that $\mathbf{\Upsilon}_{22}$ $(k \times k)$ is nonsingular, that is, of rank k, because if $\mathbf{\Upsilon}_{22}$ were singular there would exist a k-vector $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}'\mathbf{\Upsilon}_{22} = \mathbf{0}$ and then $\boldsymbol{\gamma}'\mathbf{\Pi}_2$ would be another left-sided eigenvector of $\mathbf{\Pi}$ associated with the root 0, but that would imply more than n linearly independent vectors satisfying $\boldsymbol{\omega}'\mathbf{\Pi} = \mathbf{0}$ and hence more than n zeros of $|\mathbf{B}(\lambda)|$ at $\lambda = 1$, which is contrary to assumption. Note that (2.6) is a factorization $\mathbf{\Pi} = \mathbf{A}\mathbf{\Gamma}'$ with $\mathbf{\Gamma}' = \mathbf{\Pi}_2$.

The matrix $\mathbf{\Omega}$ satisfies

$$\mathbf{\Omega}'\mathbf{\Pi} = \begin{bmatrix} \mathbf{\Omega}_1' \\ \mathbf{\Omega}_2' \end{bmatrix} \begin{bmatrix} -\mathbf{\Omega}_{21}' \\ \mathbf{I}_k \end{bmatrix} \mathbf{\Pi}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Upsilon}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_1' \\ \mathbf{\Omega}_2' \end{bmatrix} = \mathbf{\Upsilon}\mathbf{\Omega}', \tag{2.8}$$

$$\mathbf{\Omega}'\mathbf{B} = \mathbf{\Omega}'(\mathbf{\Pi} + \mathbf{I}) = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_{22} \end{bmatrix} \mathbf{\Omega}' = \mathbf{\Psi}\mathbf{\Omega}'. \tag{2.9}$$

where $\mathbf{\Psi}_{22} = \mathbf{\Upsilon}_{22} + \mathbf{I}_k$. Let $\mathbf{\Pi}_2 = (\mathbf{\Pi}_{21}, \mathbf{\Pi}_{22})$. Then $\mathbf{\Omega}$ is nonsingular because

$$|\mathbf{\Omega}| = \begin{vmatrix} \mathbf{I}_n & \mathbf{0} \\ -\mathbf{\Omega}_{21} & \mathbf{I}_k \end{vmatrix} \begin{vmatrix} \mathbf{I}_n & \mathbf{\Pi}_{21}' \\ \mathbf{\Omega}_{21} & \mathbf{\Pi}_{22}' \end{vmatrix} = \begin{vmatrix} \mathbf{I}_n & \mathbf{\Pi}_{21}' \\ \mathbf{0} & \mathbf{\Upsilon}_{22}' \end{vmatrix} = |\mathbf{\Upsilon}_{22}'| \neq 0. \tag{2.10}$$

Hence (2.4) is a nonsingular linear transformation.

The transformed process $\mathbf{X}_t$ satisfies the autoregressive model

$$\mathbf{X}_t = \mathbf{\Psi}\mathbf{X}_{t-1} + \mathbf{W}_t, \tag{2.11}$$

$$\Delta\mathbf{X}_t = \mathbf{\Upsilon}\mathbf{X}_{t-1} + \mathbf{W}_t, \tag{2.12}$$

where

$$\mathbf{\Psi} = \mathbf{\Omega}'\mathbf{B}_1(\mathbf{\Omega}')^{-1} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_{22} \end{bmatrix} \tag{2.13}$$

has eigenvalues λ_i, $i = 1, \ldots, p$, and $\mathbf{\Psi}_{22}$ has eigenvalues λ_i, $i = n+1, \ldots, p$, and $\mathbf{\Upsilon} = \mathbf{\Psi} - \mathbf{I}_p$. From (2.11) to (2.13) we obtain

$$\begin{aligned} \begin{bmatrix} \Delta\mathbf{X}_{1t} \\ \mathbf{X}_{2t} \end{bmatrix} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_{22} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{X}_{1,t-1} \\ \mathbf{X}_{2,t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{W}_{1t} \\ \mathbf{W}_{2t} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{W}_{1t} \\ \mathbf{\Psi}_{22}\mathbf{X}_{2,t-1} + \mathbf{W}_{2t} \end{bmatrix} \end{aligned} \tag{2.14}$$

as generating the process $(\Delta\mathbf{X}_{1t}', \mathbf{X}_{2t}')'$. Since the eigenvalues of the coefficient matrix in (2.14) are 0 of multiplicity n and λ_i, $i = n+1, \ldots, p$, the process $(\Delta\mathbf{X}_{1t}', \mathbf{X}_{2t}')'$ is a stationary process. □

The transformation $\mathbf{X}_t = \mathbf{\Omega}'\mathbf{Y}_t$ is a change of coordinates such that the first n coordinates of $\mathbf{X}_t$ define a random walk, which is an $I(1)$ process. The other k coordinates define a stationary process. Thus $\{\mathbf{X}_t\}$ is an $I(1)$ process; that is, $\Delta\mathbf{X}_t$ is an $I(0)$ process. The process $\mathbf{Y}_t = (\mathbf{\Omega}')^{-1}\mathbf{X}_t$ is a mixture of an $I(1)$ and an $I(0)$ process.

3. General case

Theorem 2. *When the Rank Condition holds,*

$$\begin{bmatrix} \Delta \mathbf{Y}_t \\ \mathbf{\Pi}_2 \mathbf{Y}_t \end{bmatrix} \tag{3.1}$$

defines a stationary process.

Proof. For arbitrary m the models (1.1) and (1.3) can be written in "stacked" form as

$$\widetilde{\mathbf{Y}}_t = \widetilde{\mathbf{B}}_1 \widetilde{\mathbf{Y}}_{t-1} + \widetilde{\mathbf{Z}}_t \tag{3.2}$$

and

$$\Delta \widetilde{\mathbf{Y}}_t = \widetilde{\mathbf{\Pi}} \widetilde{\mathbf{Y}}_{t-1} + \widetilde{\mathbf{Z}}_t, \tag{3.3}$$

where

$$\widetilde{\mathbf{Y}}_t = \begin{bmatrix} \mathbf{Y}_t \\ \mathbf{Y}_{t-1} \\ \mathbf{Y}_{t-2} \\ \vdots \\ \mathbf{Y}_{t-m+1} \end{bmatrix}, \widetilde{\mathbf{Z}}_t = \begin{bmatrix} \mathbf{Z}_t \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \widetilde{\mathbf{B}}_1 = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \dots & \mathbf{B}_{m-1} & \mathbf{B}_m \\ \mathbf{I}_p & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_p & \mathbf{0} \end{bmatrix}, \tag{3.4}$$

and $\widetilde{\mathbf{\Pi}} = \widetilde{\mathbf{B}}_1 - \mathbf{I}_{mp}$. [See Anderson (1971), Section 5.3, for example.] Let $\widetilde{\mathbf{B}}(\lambda) = \lambda \mathbf{I}_{mp} - \widetilde{\mathbf{B}}_1$. Then $|\mathbf{B}(\lambda)| = |\widetilde{\mathbf{B}}(\lambda)|$. We shall prove Theorem 2 by using Theorem 1 with $\mathbf{Y}_t$ replaced by $\widetilde{\mathbf{Y}}_t$.

Suppose that there are n linearly independent solutions to $\widetilde{\boldsymbol{\omega}}' \widetilde{\mathbf{\Pi}} = \mathbf{0}$. Let these solutions be assembled into the $n \times mp$ matrix $\widetilde{\mathbf{\Omega}}_1' = (\widetilde{\mathbf{\Omega}}_{11}', \dots, \widetilde{\mathbf{\Omega}}_{m1}')$. Then

$$\begin{aligned} \mathbf{0} &= \widetilde{\mathbf{\Omega}}_1' \widetilde{\mathbf{\Pi}} \\ &= \Big[\widetilde{\mathbf{\Omega}}_{11}'(\mathbf{B}_1 - \mathbf{I}_p) + \widetilde{\mathbf{\Omega}}_{21}',\ \widetilde{\mathbf{\Omega}}_{11}' \mathbf{B}_2 - \widetilde{\mathbf{\Omega}}_{21}' + \widetilde{\mathbf{\Omega}}_{31}', \dots, \\ &\qquad \widetilde{\mathbf{\Omega}}_{11}' \mathbf{B}_{m-1} - \widetilde{\mathbf{\Omega}}_{m-1,1}' + \widetilde{\mathbf{\Omega}}_{m1}',\ \widetilde{\mathbf{\Omega}}_{11}' \mathbf{B}_m - \widetilde{\mathbf{\Omega}}_{m1}' \Big]. \end{aligned} \tag{3.5}$$

This equation implies

$$\widetilde{\mathbf{\Omega}}_{m1}' = \widetilde{\mathbf{\Omega}}_{11}' \mathbf{B}_m = -\widetilde{\mathbf{\Omega}}_{11}' \mathbf{\Pi}_{m-1}, \tag{3.6}$$

$$\widetilde{\mathbf{\Omega}}_{m-j,1}' = \widetilde{\mathbf{\Omega}}_{11}' \mathbf{B}_{m-j} + \widetilde{\mathbf{\Omega}}_{m-j+1}' = -\widetilde{\mathbf{\Omega}}_{11}' \mathbf{\Pi}_{m-j-1}, \qquad j = 1, \dots, m-1, \tag{3.7}$$

$$\mathbf{0} = \widetilde{\mathbf{\Omega}}_{11}'(\mathbf{B}_1 - \mathbf{I}_p) + \widetilde{\mathbf{\Omega}}_{21}' = \widetilde{\mathbf{\Omega}}_{11}' \mathbf{\Pi}. \tag{3.8}$$

It follows that

$$\widetilde{\mathbf{\Omega}}_1' = \widetilde{\mathbf{\Omega}}_{11}' [\mathbf{I}_p, -\mathbf{\Pi}_1, \dots, -\mathbf{\Pi}_{m-1}]. \tag{3.9}$$

Lemma. *There is a $pm \times n$ matrix $\widetilde{\mathbf{\Omega}}_1$ of rank n such that $\widetilde{\mathbf{\Omega}}_1' \widetilde{\mathbf{\Pi}} = \mathbf{0}$ if and only if there is a $p \times n$ matrix $\widetilde{\mathbf{\Omega}}_{11}$ of rank n such that $\widetilde{\mathbf{\Omega}}_{11}' \mathbf{\Pi} = \mathbf{0}$.*

Thus the Rank Condition on the mp-dimensional $\widetilde{\mathbf{Y}}_t$ in terms of $\widetilde{\mathbf{\Pi}}$ is equivalent to the Rank Condition on $\mathbf{Y}_t$, where $\mathbf{\Pi}$ is defined by (1.5).

It follows from Theorem 1 applied to (3.2) that the rank of $\widetilde{\mathbf{\Pi}}$ is $\tilde{k} = mp - n$. Let

$$\widetilde{\mathbf{\Pi}} = \begin{bmatrix} \widetilde{\mathbf{\Pi}}_{\cdot n} \\ \widetilde{\mathbf{\Pi}}_{\cdot \tilde{k}} \end{bmatrix} = \begin{bmatrix} (\mathbf{B}_1 - \mathbf{I}_p)_{\cdot n} & \mathbf{B}_{2\cdot n} & \cdots & \mathbf{B}_{m-1\cdot n} & \mathbf{B}_{m\cdot n} \\ (\mathbf{B}_1 - \mathbf{I}_p)_{\cdot k} & \mathbf{B}_{2\cdot k} & \cdots & \mathbf{B}_{m-1\cdot k} & \mathbf{B}_{m\cdot k} \\ \mathbf{I}_p & -\mathbf{I}_p & & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & & \mathbf{I}_p & -\mathbf{I}_p \end{bmatrix}, \tag{3.10}$$

where $\widetilde{\mathbf{\Pi}}_{\cdot n}$ has n rows and $(\ )_{\cdot n}$ denotes the first n rows of $(\ )$ and $(\ )_{\cdot k}$ denotes the last k rows of $(\ )$. The $pm \times \tilde{k}$ matrix $\widetilde{\mathbf{\Omega}}_2 = \widetilde{\mathbf{\Pi}}'_{\cdot \tilde{k}}$ satisfies

$$\widetilde{\mathbf{\Omega}}_2' \widetilde{\mathbf{\Pi}} = \widetilde{\mathbf{\Upsilon}}_{22} \widetilde{\mathbf{\Omega}}_2', \tag{3.11}$$

$\widetilde{\mathbf{\Upsilon}}_{22}$ is nonsingular, and $\widetilde{\mathbf{\Omega}} = (\widetilde{\mathbf{\Omega}}_1, \widetilde{\mathbf{\Omega}}_2)$ is nonsingular. Define $\widetilde{\mathbf{X}}_t = \widetilde{\mathbf{\Omega}}' \widetilde{\mathbf{Y}}_t$ and $\widetilde{\mathbf{W}}_t = \widetilde{\mathbf{\Omega}}' \widetilde{\mathbf{Z}}_t$. Then $\widetilde{\mathbf{X}}_t = (\widetilde{\mathbf{X}}_{1t}', \widetilde{\mathbf{X}}_{2t}')'$ satisfies

$$\widetilde{\mathbf{X}}_{1t} = \widetilde{\mathbf{X}}_{1,t-1} + \widetilde{\mathbf{W}}_{1t}, \tag{3.12}$$

$$\widetilde{\mathbf{X}}_{2t} = \widetilde{\mathbf{\Psi}}_{22} \widetilde{\mathbf{X}}_{2,t-1} + \widetilde{\mathbf{W}}_{2t}, \tag{3.13}$$

where the eigenvalues of $\widetilde{\mathbf{\Psi}}_{22}$ are λ_i, $i = n+1, \ldots, mp$,

$$\widetilde{\mathbf{X}}_{2t} = \widetilde{\mathbf{\Omega}}_2' \widetilde{\mathbf{Y}}_t = \begin{bmatrix} (\mathbf{B}_1 - \mathbf{I}_p)_{\cdot k} \mathbf{Y}_t + \mathbf{B}_{2\cdot k} \mathbf{Y}_{t-1} + \ldots + \mathbf{B}_{m\cdot k} \mathbf{Y}_{t-m+1} \\ \mathbf{Y}_t - \mathbf{Y}_{t-1} \\ \vdots \\ \mathbf{Y}_{t-m+2} - \mathbf{Y}_{t-m+1} \end{bmatrix}, \tag{3.14}$$

and

$$\widetilde{\mathbf{W}}_{2t} = \widetilde{\mathbf{\Omega}}_2' \widetilde{\mathbf{Z}}_t = \begin{bmatrix} (\mathbf{B}_1 - \mathbf{I}_p)_{\cdot k} \mathbf{W}_t \\ \mathbf{W}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{3.15}$$

Thus $\{\widetilde{\mathbf{X}}_{1t}\}$ is an $I(1)$ process of dimension n and $\{\widetilde{\mathbf{X}}_{2t}\}$ is an $I(0)$ process of dimension $\tilde{k}$.

Now we want to transform $\{\widetilde{\mathbf{X}}_t\}$ so that $k = p - n$ coordinates constitute the cointegrated part of $\{\mathbf{Y}_t\}$ and the other coordinates are components of $\Delta\mathbf{Y}_t, \ldots, \Delta\mathbf{Y}_{t-m+1}$. In terms of $\mathbf{Y}_t$ (3.12) can be written

$$\sum_{j=1}^{m} \widetilde{\mathbf{\Omega}}_{j1}' \Delta\mathbf{Y}_{t-j+1} = \widetilde{\mathbf{\Omega}}_{11}' \left(\Delta\mathbf{Y}_t - \sum_{j=2}^{m} \mathbf{\Pi}_{j-1} \Delta\mathbf{Y}_{t-j+1} \right) = \widetilde{\mathbf{\Omega}}_{11}' \mathbf{Z}_t = \widetilde{\mathbf{W}}_{1t}. \tag{3.16}$$

Let

$$\widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{I}_k & -\mathbf{\Pi}_{1\cdot k} & \cdots & -\mathbf{\Pi}_{m-1\cdot k} \\ \mathbf{0} & \mathbf{I}_p & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_p \end{bmatrix}, \tag{3.17}$$

$$\widetilde{\mathbf{V}}_{2t} = \widetilde{\mathbf{M}} \widetilde{\mathbf{X}}_{2t} = \begin{bmatrix} \mathbf{\Pi}_{\cdot k} \mathbf{Y}_t \\ \Delta\mathbf{Y}_t \\ \vdots \\ \Delta\mathbf{Y}_{t-m+2} \end{bmatrix}, \quad \widetilde{\mathbf{U}}_{2t} = \widetilde{\mathbf{M}} \mathbf{W}_{2t} = \begin{bmatrix} \mathbf{\Pi}_{\cdot k} \mathbf{W}_t \\ \mathbf{W}_t \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{3.18}$$

Here $\mathbf{\Pi}_{\cdot k}$ denotes the last k rows of $\mathbf{\Pi}$ defined by (1.5); that is, $\mathbf{\Pi}_{\cdot k} = \mathbf{\Pi}_2$ in (2.6). Let $\widetilde{\mathbf{\Theta}} = \widetilde{\mathbf{M}}\widetilde{\mathbf{\Psi}}_{22}\widetilde{\mathbf{M}}^{-1}$. Then $\widetilde{\mathbf{V}}_{2t}$ satisfies

$$\widetilde{\mathbf{V}}_{2t} = \widetilde{\mathbf{\Theta}}\widetilde{\mathbf{V}}_{2,t-1} + \widetilde{\mathbf{U}}_{2t}. \tag{3.19}$$

The eigenvalues of $\widetilde{\mathbf{\Theta}}$ are λ_i, $i = n+1, \ldots, mp$. Hence $\widetilde{\mathbf{V}}_{2t}$ defines a stationary process. In fact

$$\widetilde{\mathbf{V}}_{2t} = \sum_{s=0}^{\infty} \widetilde{\mathbf{\Theta}}^s \widetilde{\mathbf{U}}_{2,t-s}. \tag{3.20}$$

Since the last $m-2$ blocks of $\widetilde{\mathbf{U}}_{2t}$ are $\mathbf{0}$'s, the last $m-2$ blocks of (3.19) are identities. The first $k+p$ rows of (3.19) define a stationary process for $\mathbf{\Pi}_{\cdot k}\mathbf{Y}_t$ and $\Delta\mathbf{Y}_t$. □ □

Discussion. *The process* $\{\mathbf{Y}_t\}$ *is* cointegrated *of rank* k*, and* $\mathbf{\Pi}_{\cdot k}$ *is the* cointegrating matrix.

The orthogonality conditions of $\mathbf{A}_\perp$ and $\mathbf{\Gamma}_\perp$ are equivalent to $\mathbf{A}_\perp\mathbf{\Pi} = \mathbf{0}$ and $\mathbf{\Pi}\mathbf{\Gamma}_\perp = \mathbf{0}$. Hence, $\mathbf{A}_\perp$ consists of $p-k$ left-sided characteristic vectors of $\mathbf{\Pi}$ corresponding to the characteristic root of 0 and $\mathbf{\Gamma}_\perp$ consists of $p-k$ right-sided characteristic vectors corresponding to the root of 0. The matrix $\mathbf{\Gamma}$ corresponds to $\mathbf{\Omega}_2 = \mathbf{\Pi}_2'$.

4. Inference

The model (1.3) has the form of regression

$$\mathbf{Y}_t = \mathbf{A}_1\mathbf{X}_{1t} + \mathbf{A}_2\mathbf{X}_{2t} + \mathbf{Z}_t, \tag{4.1}$$

where $\mathbf{A}_1$ is of rank k. The maximum likelihood estimator of $\mathbf{A}_1$ under normality of $\mathbf{Z}_t$ is the reduced rank regression estimator introduced by Anderson (1951). Johansen (1988), (1995) also derived the estimator for (1.3) and gives some asymptotic theory suitable for the cointegrated model. Anderson (2000), (2001), (2002) has given more details of the asymptotic theory.

Acknowledgement

The author thanks Cheng Hsiao for helpful suggestions.

References

[1] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics*, **22**, 327–351. [Correction, *Annals of Statistics*, **8**, (1980), p. 1400.] MR42664

[2] Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, John Wiley and Sons, Inc., New York. MR283939

[3] Anderson, T. W. (2000). The asymptotic distribution of canonical correlations in cointegrated models. *Proceedings of the National Academy of Sciences*, **97**, 7068–7073. MR1769813

[4] Anderson, T. W. (2001). The asymptotic distribution of canonical correlations and variates in higher-order cointegrated models, *Proceedings of the National Academy of Sciences*, **98**, 4860–4865. MR1828055

[5] Anderson, T. W. (2002). Reduced rank regression in cointegrated models, *Journal of Econometrics*, **106**, 203–216. MR1884248

[6] Johansen, Soren (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254. MR986516

[7] Johansen, Soren (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford. MR1487375

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 385–395
© Institute of Mathematical Statistics, 2004

Forecasting NBA basketball playoff outcomes using the weighted likelihood*

Feifang Hu[1] **and James V. Zidek**[2]

University of Virginia and University of British Columbia

Abstract: Predicting the outcome of a future game between two sports teams poses a challenging problem of interest to statistical scientists as well as the general public. To be effective such prediction must exploit special contextual features of the game. In this paper, we confront three such features and address the need to: (i) use all relevant sample information; (ii) reflect the home court advantage. To do so we use the relevance weighted likelihood of Hu and Zidek (2002). Finally we demonstrate the value of the method by showing how it could have been used to predict the 1996–1997 NBA Final series results. Our relevance likelihood-based method proves to be quite accurate.

1. Introduction

This paper demonstrates the use of weighted likelihood (WL) to predict the winner of 1996–1997 National Basketball Association (NBA) Finals between the Chicago Bulls and the Utah Jazz. However, as we try to indicate, the WL has much wider applicability inside as well outside the domain of sports.

Statistical methods have been extensively used in sports (Bennett 1998). Harville (1977) uses regression analysis to rate high school and college football teams based on observed score differences. In a later paper (Harville 1980), he develops a method for forecasting the point spread of NFL games by using similar techniques. In related papers, Schwertman et al (1996) and Carlin (1996) tackle NCAA basketball. Both papers (like this one) estimate the probability that team i beats j. They (unlike us) are based on pre-game information. The first uses a logistic regression analysis of win - loss records and various functions of seed numbers (that is ranks assigned to the teams going into a tournament), as a way of incorporating prior knowledge and expert opinion. The second extends earlier unpublished work of Schwertman et al (1993) by using other external information such as "...the RPI index, Sagarin ratings, and so on..." in addition to seed numbers. Like Harville (1980) and Stern (1992), Carlin uses published point spreads to capture pregame information and does a linear regression analysis of observed point spreads on pregame information. Models derived from that analysis can be used to predict game winners.

Our approach, unlike those described above, does not attempt to take pre-game information into consideration although it may be possible to do that through the weights in the WL. That issue remains to be explored. Instead, our goal is to introduce the WL method and show how it can be used. No doubt improvements that

*The work reported in this article was partially supported by a grant from the Natural Science and Engineering Research Council of Canada.

[1]Department of Statistics, University of Virginia, 104 Halsey Hall, P.O. Box 400135, Charlottesville, Virginia 22904-4135, USA. e-mail: fh6e@virginia.edu

[2]Department of Statistics, The University of British Columbia, 6356 Agricultural Road, Vancouver BC V6T 1Z2, Canada. e-mail: jim@stat.ubc.ca

Keywords and phrases: likelihood, relevance weighted likelihood, predicting basketball scores, sports statistics, NBA, basketball.

AMS 2000 subject classifications: 62M20, 62P99.

build on earlier work could enhance the method. However, we do assess our approach against a logistical method that embraces the celebrated method of Bradley and Terry (1952) that also underlies the work of Schwertman et al. (1996).

The genesis of our work lies in two statistical problems encountered in sports: (i) the prediction of the outcome of a future game between two specified sports teams; (ii) the assessment of the accuracy of this prediction. Since typically these two teams will not have met more than just a few times in the given season, little direct information will be available to the forecaster. The consequent small sample size will make naive predictions inaccurate and the associated prediction intervals excessively large.

Turning to the NBA Finals, we note that the winner is the team that wins a best of 7 series (that is, the first team to win four games). To predict that outcome, one might sequentially determine the prediction probability of a Bulls' win in each of a series of successive games. To find that probability, the 1996–1997 season data would be used. However, the Bulls met the Jazz just twice, providing the only "direct" information available, in the terminology of Hu and Zidek (1993) and Hu (1994). However, that small sample cannot generate accurate predictions.

To overcome this data deficiency, observe that the Bulls (like the Jazz) played 82 games in the season (2 with the Jazz and 80 with other teams). The 160 games these two played against other teams provide "relevant" information, in the Hu-Zidek terminology.

To use both the "direct" and "relevant" information in some simple yet flexible way, Hu (1994) proposes the "relevance weighted likelihood". Hu and Zidek (2002) extend that likelihood and Wang (2001) further extended it to get the "weighted likelihood (WL)", the terminology we use in this paper.

The method of weighted likelihood has been applied to a neurophysiology experiment (Hu and Rosenberger, 2000). In that paper, they find that both bias and mean square error are significantly reduced by using the weighted likelihood method. Hu and Zidek (2001) use the WL to predict the number of goals (with prediction intervals) for each of the Vancouver Canucks and Calgary Flames in their NHL games against each other during the 1996–1997 season. They (Hu and Zidek 2002) show how the WL can be used to construct generalizations of the classical Shewhart control charts. Their generalization includes the moving average and exponentially moving average charts and allows for a variety of failure modes when processes go out of control. This application introduces the weighted likelihood ratio test. In that same paper, they show how the James Stein estimator, including generalizations, can be found with the WL.

A particularly important class of applications arise in estimating parameters that are interrelated, leading to natural relationships among the associated populations and inducing transfers of information from their associated samples. Van Eeden and Zidek (2002) show how such interrelations may be exploited through the WL when the means of two normal populations with known variances are ordered. The analogous problem when the mean difference is bounded is treated in Van Eeden and Zidek (2000). Finally, we would mention an application to disease mapping in Wang (2001).

In Section 2, we apply the WL in the NBA forecasting application above by taking advantage of special features of sports data. The maximum WL estimator (MWLE) is developed for predicting the result of a future game. The mean square error of this MWLE is given. Moreover, we construct approximate confidence intervals using the asymptotic theory for the MWLE given by Hu (1997).

In Section 3, we apply the method developed in Section 2 to predict the 1996–1997 NBA playoff results, specifically for games involving the Chicago Bulls and the Utah Jazz. Our predictions agree quite well with the actual outcomes.

To validate that positive performance assessment, in Section 3 we consider the playoff games played by the Bulls against each of three other teams, the Miami Heat, the Atlanta Hawks, and the New York Knicks. Similarly, playoff games between the Heat and Knicks are considered. These additional predictions are also in good agreement with the actual game outcomes.

Many other approaches can be taken in our application. In Section 4, our method is shown to compare favorably with a "purpose built" competitor, an extension of the Bradley Terry model (Bradley and Terry 1952). Moreover, it proves to have all the flexibility and much of the simplicity of its classical predecessor proposed by Fisher. Thus, we are able to recommend it as a practical alternative to its competitors for the application considered.

2. Sports data and the WL

2.1. Contextual features

Usually in sports, the outcome of any one game derives from the combined efforts of two teams that have seldom played each other before. Yet these games yield the only direct sample information available about the relative strength of these two teams. At the same time, each of these teams will have played many games against other teams thereby generating relevant (although not direct) sample information. The predictive probability of a win in the next game between these two teams, should combine both kinds of information.

In some sports, the home team has a great advantage (see Section 3) that must be accounted for when the data are analyzed (although in their application, Hu and Zidek (2002) ignored that advantage). Finally, the outcome of any one game will depend on both the offensive and defensive capabilities of the teams involved. Satisfactory prediction of future games requires that we combine information about the offense and defense of the two teams involved in any specific game.

2.2. The weighted likelihood

To develop a statistical model for the analysis of sports data, one should recognize the distinctive contextual features described in the last subsection. Let $Y_{AB}(h)$ be a Bernoulli random variable that is 1 if team, A, wins against team B when B is at home. Similarly, let $Y_{AB}(r)$ be a random variable that is 1 or 0 according as team A wins against team B when team B is at home. Note that $Y_{AB}(h) = 1 - Y_{BA}(r)$. As an approximation, assume the time series of Y's for different games and team pairs are independent in this paper. Clearly, a more sophisticated approach like that of Hu, Rosenberger, and Zidek (2000) would allow dependent game outcomes.

Suppose the $\{Y_{AB}(h)\}$ and $\{Y_{AB}(r)\}$ have probability density functions $f(y, p_{AB}(h))$ and $f(y, p_{AB}(r))$ respectively. To predict the game result, $(Y_{AB}(h), Y_{BA}(r))$ or $(Y_{AB}(r), Y_{BA}(h))$, we have to estimate the parameters $p_{AB}(h)$ and $p_{AB}(r)$.

To create the weights required in implementing the WL, we choose the same weight in the likelihood factor corresponding to each of the games A played against teams other than B, irrespective of the opponent. From Hu and Zidek (2002), we may use the weighted likelihood method to estimate the parameters $p_{AB}(h)$ and

$p_{AB}(r)$. The log weighted likelihood of $p_{AB}(h)$ thus becomes

$$\sum_{i=1}^{k_{AB}} \log f\big(y_{AB}(h), p_{AB}(h)\big) + \alpha_{AB}(h) \sum_{A(B)} \log f\big(y_{A(B)}(h), p_{AB}(h)\big) + \beta_{AB}(h) \sum_{(A)B} \log f\big(y_{(A)B}(h), p_{AB}(h)\big), \tag{1}$$

where k_{AB} is the number of games that A against B at home; $\sum_{A(B)}$ denotes the sum over all games that A played against teams other than B in the league with A at home and $y_{A(B)}(h)$ the corresponding binary game outcomes; $\sum_{(A)B}$ is the sum over all games that B played against teams other than A when B is away and $y_{(A)B}(h)$ the corresponding outcomes. Let $\hat{p}_{AB}^{MWLE}(h)$ be the corresponding maximum weighted likelihood estimate (MWLE) of $p_{AB}(h)$. The MWLE of $p_{AB}(r)$ can be defined in a similar way.

We adopt the approximate Akaike criterion (Akaike, 1977, Akaike, 1985, and Hu and Zidek, 2002) to select the weights $\alpha_{AB}(h)$ and $\beta_{AB}(h)$ by minimizing with respect to both,

$$E\big(\hat{p}_{AB}(h) - p_{AB}(h)\big)^2. \tag{2}$$

The resulting optima will, however, depend on the unknown p's being estimated. To address this problem we can use 'plug - in' estimators obtained in any reasonable way, for these p's, to obtain $\hat{\alpha}_{AB}(h)$ and $\hat{\beta}_{AB}(h)$ from Equation (2). One possible way of doing this is demonstrated in Section 3.

In most applications, we need confidence intervals (or the equivalent) for the parameters. The impossibility of finding exact confidence intervals based on the MWLE leads us to use approximate ones based on the asymptotic normality of the MWLE (see Theorem 5 of Hu, 1997). We obtain such a 95% confidence interval for $p_{AB}(h)$ as

$$\Big[\hat{p}_{AB}^{MWLE}(h) - \hat{bias}_{AB} - 1.96\sqrt{\hat{var}_{AB}}, \hat{p}_{AB}^{MWLE}(h) + \hat{bias}_{AB} + 1.96\sqrt{\hat{var}_{AB}}\Big]. \tag{3}$$

Here $\hat{bias}_{AB}$ and $\hat{var}_{AB}$ are the estimators of the bias and variance given in Theorem 5 of Hu (1997). With those estimates $\hat{p}_{AB}(h)$ and $\hat{p}_{BA}(r)$, we can find the predictive probabilities of winning, losing and drawing the game (along with their approximate confidence intervals) when a game is played at the home of Team A.

3. Predicting the NBA playoff results

In this section, we turn to the problem of predicting the outcomes of NBA playoff games. Our analysis concerns the 1996-1997 season.

The home team advantage is significant in the NBA. We tested the null-hypothesis of no home team advantage against the alternative of a home team advantage and found a p-value of about 10^{-7} suggesting the need to separate home and away games.

To describe our application, let $Y_{AB}(h) \sim Bernoulli(p_{AB}(h))$ be independently distributed random variables representing a "win" or "loss" by team A in any one game played against team B while A is at home. We first estimate the predictive probabilities $p_{AB}(h)$ and $p_{BA}(h)$ where 'A' and 'B' denote respectively the Chicago Bulls and the Utah Jazz, two top NBA teams.

The use of the weighted likelihood seems especially appealing here given the paucity of "direct" information about the relative strengths of A and B. In fact,

the Jazz played only one game in Chicago. The classical likelihood leaves no chance of finding reasonable parameter estimates. In contrast, the MWLE brings in information from games each of these teams played against others in the NBA. That is, the MWLE uses the information in the "relevant sample" in addition to that in the "direct sample".

We find the MWLE of $p_{AB}(h)$ (from the weighted likelihood (1)) to be

$$\begin{aligned}\hat{p}_{AB}^{MWLE}(h) &= \bar{y}_{AB}(h) + \alpha_{AB}(h)\big(\bar{y}_{A(B)}(h) - \bar{y}_{AB}(h)\big) \\ &\quad + \beta_{AB}(h)\big(\bar{y}_{(A)B}(h) - \bar{y}_{AB}(h)\big),\end{aligned} \tag{4}$$

where $\bar{y}_{AB}(h)$ denotes the fraction of wins for A in the $k_{AB}(h)$ games played against B during the season with A at home. The $\bar{y}_{A(B)}(h)$ represents the corresponding fraction of wins for A in the $k_{A(B)}(h)$ games played against teams other than B with A at home.

By using the approximate Akaike criterion with a reasonable estimate $\hat{p}_{AB}(h)$ (described below), an optimal weight may be estimated by

$$\hat{\alpha}_{AB}(h) = \frac{V_{AB}(h)[V_{(A)B}(h) + (\bar{y}_{(A)B}(h) - \hat{p}_{AB}(h))(\bar{y}_{(A)B}(h) - \bar{y}_{A(B)}(h))]}{C + D} \tag{5}$$

and

$$\hat{\beta}_{AB}(h) = \frac{V_{AB}(h)[V_{A(B)}(h) + (\bar{y}_{A(B)}(h) - \hat{p}_{AB}(h))(\bar{y}_{A(B)}(h) - \bar{y}_{(A)B}(h))]}{C + D} \tag{6}$$

where

$$\begin{aligned}V_{AB}(h) &= \frac{\hat{p}_{AB}(h)(1 - \hat{p}_{AB}(h))}{k_{AB}(h)}, \\ V_{A(B)}(h) &= \frac{\bar{y}_{A(B)}(h)(1 - \bar{y}_{A(B)}(h))}{k_{A(B)}(h)}, \\ V_{(A)B}(h) &= \frac{\bar{y}_{(A)B}(h)(1 - \bar{y}_{(A)B}(h))}{k_{(A)B}(h)}, \\ C &= V_{AB}(h)\big[V_{(A)B}(h) + V_{A(B)}(h) + \big(\bar{y}_{(A)B}(h) - \bar{y}_{A(B)}(h)\big)^2\big]\end{aligned}$$

and

$$\begin{aligned}D &= V_{A(B)}(h)\big(\bar{y}_{(A)B}(h) - \hat{p}_{AB}(h)\big)^2 + V_{(A)B}(h)\big(\bar{y}_{A(B)}(h) - \hat{p}_{AB}(h)\big)^2 \\ &\quad + V_{A(B)}(h)V_{(A)B}(h).\end{aligned}$$

The corresponding mean square error of the MWLE may be estimated by

$$\begin{aligned}\hat{MSE}_{MWLE} &= \big[\hat{\alpha}_{AB}(h)\big(\bar{y}_{A(B)}(h) - \hat{p}_{AB}^{MWLE}(h)\big) \\ &\quad + \hat{\beta}_{AB}(h)\big(\bar{y}_{(A)B}(h) - \hat{p}_{AB}^{MWLE}(h)\big)\big]^2 \\ &\quad + \hat{\alpha}_{AB}^2(h)\frac{\bar{y}_{A(B)}(h)(1 - \bar{y}_{A(B)}(h))}{k_{A(B)}(h)} \\ &\quad + \hat{\beta}_{AB}^2(h)\frac{\bar{y}_{(A)B}(h)(1 - \bar{y}_{(A)B}(h))}{k_{(A)B}(h)} \\ &\quad + (1 - \hat{\alpha}_{AB}(h) - \hat{\beta}_{AB}(h))^2\frac{\hat{p}_{AB}^{MWLE}(h)(1 - \hat{p}_{AB}^{MWLE}(h))}{k_{AB}(h)}.\end{aligned}$$

The 95% confidence interval of $p_{AB}(h)$ based on the MWLE would be: $[\hat{p}_{AB}^{MWLE}(h) - \hat{bias}_{AB}(h) - 1.96\sqrt{\hat{var}_{AB}(h)}, \hat{p}_{AB}^{MWLE}(h) + \hat{bias}_{AB}(h) + 1.96\sqrt{\hat{var}_{AB}(h)}]$, where

$$\hat{bias}_{AB}(h) = |\hat{\alpha}_{AB}(h)(\bar{y}_{A(B)}(h) - \hat{p}_{AB}^{MWLE}(h)) + \hat{\beta}_{AB}(h)(\bar{y}_{(A)B}(h) - \hat{p}_{AB}^{MWLE}(h))|$$

and

$$\begin{aligned} \hat{var}_{AB}(h) &= \hat{\alpha}_{AB}^2(h)\frac{\bar{y}_{A(B)}(h)(1-\bar{y}_{A(B)}(h))}{k_{A(B)}(h)} \\ &\quad + \hat{\beta}_{AB}^2(h)\frac{\bar{y}_{(A)B}(h)(1-\bar{y}_{(A)B}(h))}{k_{(A)B}(h)} \\ &\quad + (1-\hat{\alpha}_{AB}(h)-\hat{\beta}_{AB}(h))^2\frac{\hat{p}_{AB}^{MWLE}(h)(1-\hat{p}_{AB}^{MWLE}(h))}{k_{AB}(h)}. \end{aligned}$$

We now describe how we found the plug-in estimates, the optimal weights, the win probabilities and the corresponding confidence intervals by considering the Bulls against the Jazz while the Bulls are at home.

During the regular season, the Bulls played 41 games at home. One game was against the Jazz and the Bulls won this game. So $k_{AB} = 1$ and $\bar{Y}_{AB} = 1$. The Bulls played 40 games against teams other than the Jazz and won 38 of these games. Thus, $k_{A(B)} = 40$ and $\bar{Y}_{A(B)} = 0.95$. The Jazz played 40 ($k_{(A)B} = 40$) games against teams other than the Bulls on road and won 26 of these games. $\bar{Y}_{(A)B} = 1 - 26/40 = 0.35$. For this case, the plug-in estimate,

$$\hat{p}_{AB}(h) = \frac{k_{AB}\bar{Y}_{AB} + k_{A(B)}\bar{Y}_{A(B)} + k_{(A)B}\bar{Y}_{(A)B}}{k_{AB} + k_{A(B)} + k_{(A)B}} = \frac{1+38+14}{1+40+40} = \frac{53}{81} = 0.6543.$$

The corresponding values in equation (5) and (6) can be calculated by using above results. And the values are: $V_{AB}(h) = 0.2262$, $V_{A(B)}(h) = 0.0011875$, $V_{(A)B}(h) = 0.0056875$, $C = 0.082987$ and $D = 0.000637$. Subsistute these values into equation (5) and (6), we get the optimal weights:

$$\hat{\alpha}_{AB}(h) = 0.50925, \quad \text{and} \quad \hat{\beta}_{AB}(h) = 0.4831.$$

The MWLE in (1) is then

$$\hat{p}_{AB}^{MWLE}(h) = 0.66.$$

The corresponding mean square error, bias and variance of this MWLE are

$$MSE_{MWLE} = 0.001653, \ \hat{bias}_{AB}(h) = 0.002 \quad \text{and} \quad \hat{var}_{AB}(h) = 0.001648.$$

The 95% confidence interval of $p_{AB}(h)$ based on this MWLE is then $[0.58, 0.74]$.

The above MWLE is based on the games with all teams that the Bulls played at home or the Jazz played on the road. Each game has the same weight in the weighted likelihood. This seems unreasonable because some of the teams are significantly weaker than others. Now we only use the teams (10 teams in 1996/97 season) which won at least 50 games in the season. By using the games with these 10 teams, we calculate the win probabilities as well as the confidence intervals, which is denoted by MWLE1.

Before the 1996–1997 finals between the Bulls and the Jazz, both teams had played the first and second round as well as the conference finals. This additional information is used in constructing MWLE2.

Table 1: The Bulls predictive win probabilities (with mean square error) and confidence intervals based on MWLE, MWLE1 and MWLE2 for a future game between the Bulls and the Jazz during the 1996–1997 season.

	MWLE	MWLE1	MWLE2
At Chicago	0.66 (0.002)	0.77 (0.007)	0.75 (0.004)
95% C.I.	$[0.58, 0.74]$	$[0.60, 0.94]$	$[0.62, 0.89]$
At Utah	0.40 (0.002)	0.36 (0.008)	0.34 (0.004)
95% C.I.	$[0.32, 0.48]$	$[0.16, 0.55]$	$[0.21, 0.47]$

Table 2: The predictive probabilities of a Bulls' win against the Jazz together with confidence intervals for MWLE, MWLE1 and MWLE2 in the 1996–1997 Final.

	Game #	Game 4	Game 5	Game 6	Game 7	Total 90+%[a]	C.I.
MWLE	Bulls' Win	0.07	0.11	0.21	0.21	0.61	[0.43,0.77]
	Jazz Win	0.04	0.13	0.11	0.11	0.39	[0.23,0.56]
MWLE1	Bulls' Win	0.07	0.11	0.27	0.26	0.71	[0.30,0.95]
	Jazz Win	0.02	0.11	0.08	0.08	0.29	[0.05,0.70]
MWLE2	Bulls' Win	0.07	0.10	0.26	0.26	0.69	[0.37,0.92]
	Jazz Win	0.02	0.12	0.09	0.08	0.31	[0.08,0.63]

We now use MWLE, MWLE1 and MWLE2 to predict the 1996–1997 Finals between the Bulls and Jazz. We report the point estimates of the probabilities, the mean square errors and the confidence intervals of $p_{AB}(h)$ in Table 1

Based on the probabilities and the confidence intervals of Table 1, we can find the probabilities with which the Bulls (and the Jazz) will win the Finals in four, five, six and seven games. Also we can calculate the total win probabilities for the Bulls against the Jazz based on their home and away win probabilities given by each of the three estimation methods. Confidence intervals for these win probabilities may be obtained as well. In Table 2 where the results are reported, and in the tables that follow, that interval is obtained for any pair of teams say A and B from the 95% asymptotic intervals for A's home- and A's away-win-against-B probabilities. Since those intervals are stochastically dependent, we use a Bonferonni argument and obtain an asymptotic interval of confidence at least 90%. In obtaining that interval, we rely on the heuristically obvious fact that the overall win probability must be a monotonically increasing function of the home and away win probabilities.

Table 1 indicates general agreement between MWLE1 and MWLE2. But MWLE gives a much smaller estimator of a Chicago win at home. Both MWLE1 and MWLE2 predict that the Bulls would win the Finals with high probability. Also MWLE1 and MWLE2 predict the Bulls will win at Game 6. These predictions agree with the actual result: the Bulls won the Finals in six games.

To explore the performance of our method further, we have also calculated prediction probabilities for other pairs of teams, the Bulls vs. the Miami Heat, the Atlanta Hawks, the New York Knicks as well as the Miami Heat against the Knicks. The detailed results are not reported in this paper.

For the Bulls against the Miami Heat, both MWLE1 and MWLE2 also predict that most probably the Bulls will win at Game 5. That prediction proved to be correct in the playoff. When the Bulls play the Atlanta Hawks, MWLE1 and MWLE2 also predict a Bulls' win at game 5 with the highest probabilities (0.43 and 0.40). [In the playoffs the Bulls did win at game 5.]

Our analysis shows that a Heat - Knicks game will be close. MWLE and MWLE1 predict that the Heat have a slight advantage in the playoffs, while MWLE2 favors the Knicks slightly. In fact, the Heat won at game 7. However, an accident occurred in that series leading to a suspension of several New York players in games 6 and 7. Undoubtedly this influenced the outcome.

Overall, MWLE is more conservative in that its predictions are closer to 0.5 than the other methods. This is because MWLE uses some not-so-relevant information from games involving weak teams. When the Bulls and the Jazz play weak teams each wins. Thus, these data will tend to increase both of their success rates. However, since they both enjoy that benefit, the relevant difference in their estimated strengths will diminish, making the MWLE tend toward 0.5. MWLE1 and MWLE2 agree with each other, the latter giving slightly more precise predictions (as measured by the length of the associated predictive intervals in Table 1) because it incorporates the playoff games.

The Bulls and the Knicks did not meet in the playoffs. However, MWLE1 and MWLE2 predict a hypothetical Bulls' win with probabilities 0.75 and 0.78 had they met. Both predict a hypothetical Bulls' win for the series at game 5.

4. Concluding remarks

The method in this paper provides guidelines for the development of a prediction strategy. Its implementation, more specifically the construction of weights entails the incorporation of any special features that may obtain when the game is played. For example, one might need to incorporate the knowledge that certain key players cannot play in that game. [This last consideration did arise in the playoff between the Miami Heat and the New York Knicks.]

The need for the incorporation of such features was reaffirmed by an unpublished analysis carried out in the summer of 1998 by Farouk Nathoo. In that analysis, he twice simulated the entire 1997/1998 season based on the previous year's results. In his report he compared the simulation results with the actual results. Among other things he found the fraction of wins for each of the 29 NBA teams and for example we include the results for the Atlantic Division and give these results in Table 3.

We see in this example that the simulated winning percentages are in reasonable agreement with the actual results except in the case of the Nets, the Knicks and the Celtics. Given the severity of the challenge of predicting the outcomes of all games over an entire year, we find our results encouraging.

The WL method can be applied in other sports such as baseball, hockey (see Hu and Zidek 2002), soccer. In this paper, we chose the same weight for all teams. This seems unreasonable in some cases and there we may be able to use the rank of the teams to get better weights. This is another topic for the future.

Finally, we would note the abundance of alternative approaches, Bayesian (Berger, 1985) and non-Bayesian that could be used in this context. Some specific methods were discribed in the Introduction, We intend to compare our approach with some of these in future work. Here, we restricted our comparisons to an extension of one of the non - Bayesian approaches based on that of Bradley and

Table 3: The percentage of wins in the actual and two simulated 1997/1998 season for the NBA's Atlantic Division based on the WL win probability estimators obtained at the end of the previous season.

Team	Win % : Actual	Win % Simulation 1	Win % : Simulation 2
Heat	67	66	59
Nets	52	35	38
Knicks	52	65	66
Wizards	51	52	54
Magic	50	54	48
Celtics	44	20	26
Sixers	38	27	35

Terry (1952) to estimate the probabilities of a Bulls' win for both home and away games against the Jazz. (We found the corresponding probabilities for the remaining teams as well but do not report them here.) With these probabilities we could then compute the termination probabilities analogous to those in Table 2.

To be more precise, we fitted a logistic model using the software R with the response variable being 1 or 0 according as the outcome of any game during the season was a visitor or home victory. We used dummy variables to represent visitor and home teams in each game throughout the season. Thus for example, Bulls = 1 and Supersonics = 1, all other dummies being 0, would mean those two teams were playing for that particular game, the visitors being the Bulls. For each of the factors, "visitor" and "home" we represented by the dummies in this way, we arbitrarily chose the 76ers' as the baseline team. Thus, in effect, the fitted intercept, suitably transformed, provides an estimate of the likelihood of a "1" in the purely hypothetical situation where the 76ers' played themselves at home as the visitors. The coefficients for the remaining dummies represent the deviations from the 76ers' performance for each of the other teams depending on whether they were playing at home or away.

The results differed somewhat from those obtained by the MWLE2 WL method. To be specific we found the probability of a Bulls' win at home to be 0.76 as compared with the 0.75 seen in Table 1 while the corresponding probabilities for the Jazz were 0.71 and 0.66 respectively. These differences became more pronounced when we computed the probabilities corresponding to Table 2. We see a comparison of the results in Table 4.

In Table 4, we see that the Bradley–Terry extension points to a Bulls' victory on Game 7 while the MWLE2 is ambivalent between games 6 and 7. Obviously a more extensive comparison would be needed to assess the relative performance of the methods. But considering the large number of parameters needed by the logistic model, these very preliminary results make the weighted likelihood model more desirable for forecasting the outcomes of NBA playoff games.

However, we would not expect our method to do as well as it did above, when competing in particular contexts with purpose built methods. Instead, we see its value deriving from its relative ease of use and its broad domain of applicability, features it shares with the classical likelihood itself. That is, we see it as a valuable

Table 4: The predictive probabilities of a Bulls' win against the Jazz for both the MWLE2 and Bradley–Terry (logistic) based methods in the 1996–1997 Final.

	Game #	Game 4	Game 5	Game 6	Game 7
MWLE2	Bulls' Win	0.07	0.10	0.26	0.26
	Jazz Win	0.02	0.12	0.09	0.08
Bradley–Terry	Bulls' Win	0.05	0.08	0.25	0.28
	Jazz Win	0.03	0.14	0.09	0.09

tool in the statistical toolbox. In this paper, we have tried to demonstrate its value from that perspective.

In particular, although in this manuscript we have used only binary outcome information about team wins or losses, the theory can be extended to incorporate more complex outcome information such as the scores, for example. In that case, we could have defined the $Y_{AB}(h)$ to be the score of team A against team B when team A is at home and so on.

The referee pointed to another direction for future work when he or she noticed that "the weighted likelihood method to estimate the probability of A beating B (at home) uses information about B at A, C at A, and B at C. It seems logical to use information concerning A at B." We agree. However, we have not been able to do that yet since we do not know how to relate $p_{AB}(h)$ and $p_{AB}(r)$ through the WLE.

Acknowledgements

We are indebted to Farouk Nathoo and an anonymous referee for comments that improved the clarity of our presentation.

References

[1] Akaiki, H. (1977). On entropy maximization principle. In P. R. Krishnaiah, (ed.), *Applications of Statistics.* Amsterdam: North-Holland, 27–41. MR501456

[2] Akaiki, H. (1985). Prediction and entropy. *A Celebration of statistics, The ISI Centenary Volume.* Berlin, Spring-Verlag. MR816142

[3] Bennett, J (Ed) (1998). *Statistics in sport.*, London, Arnold Applications of Statistics.

[4] Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis, Second edition.* New York: Springer-Verlag. MR804611

[5] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs - I. The method of paired comparisons. *Biometrika*, 39, 324–345. MR70925

[6] Carlin, BP (1996). Improved NCAA basketball tournament modelling via point spread and team strength information. *The Amer Statistician*, 50, 39–43.

Edwards, A. W. F. (1984). *Likelihood.* First

[7] Harville, DA (1977). The use of linear-model methodology to rate high school or college football teams. *JASA*, 72, 278–289.

[8] Harville, DA (1980). Predictions for national football league games via linear-model methodology. *JASA*, 75, 516–524.

[9] Hu, F. (1994). Relevance weighted smoothing and a new bootstrap method. Ph.D. Thesis, Department of Statistics, University of British Columbia.

[10] Hu, F. (1997). Asymptotic properties of relevance weighted likelihood estimations. *Canadian Journal of Statistics*, 25, 45–60. MR1451672

[11] Hu, F. and Rosenberger, W. F. (2000). Anaylysis of time trends in adaptive designs with application to a neurophysiology experiment. *Statistics in Medicine*, 19, 2067–2075.

[12] Hu, F. and Zidek J. V. (1993b). Relevant samples and their information. *Unpublished manuscript.*

[13] Hu, F. and Zidek, J. V. (2001). The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference.* S.E.Ahmed and N.M.Reid, eds.) New York: Springer, 211–234. MR1855554

[14] Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, 30, 347–371. MR1944367

[15] Hu, F. Rosenberger, W. F. and Zidek, J. V. (2000). The relevance weighted likelihood for dependent data. *Metrika.* 51, 223–243. MR1795371

[16] Schwertman, N. C., Schenk, L., and Holbrook, B. C. (1993). More Probability models for the NCAA regional baketball tournaments. Technical Report, Department of Mathematics and Statistics, California State University, Chico.

[17] Schwertman, N. C., Schenk, L., and Holbrook, B. C. (1996). More probability models for the NCAA regional baketball tournaments. *The Amer Statistician*, 50, 34–38. MR1422072

[18] Stern, H. (1992). Who's number one? - rating football teams. In *Proceedings of the Section on Sports Statistics (Vol. 1).* Alexandria, VA: American Statistical Association, pp. 1–6.

[19] Van Eeden, C., and Zidek, J. V. (2000). Combining the data from two normal populations to estimate the mean of one when their means diffence is bounded. Available from http://hajek.stat.ubc.ca/~jim/pubs. Submitted.

[20] Van Eeden, C., and Zidek, J. V. (2002). Combining sample information in estimating ordered normal means. *Sankyha, Series A*, 64, 588–610. MR1985401

[21] Wang, S. (2001). Maximum relevance weighted likelihood estimation. Ph.D. Thesis, Department of Statistics, University of British Columbia.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 396–407
© Institute of Mathematical Statistics, 2004

Distributions of failure times associated with non-homogeneous compound Poisson damage processes

S. Zacks[1]

SUNY, Binghamton

Abstract: Failure time distributions are derived for non-homogeneous compound Poisson cumulative damage processes. We focus attention on Weibull type processes with exponential damage size. The hazard functions are illustrated and their asymptotic behavior investigated. Moment equations and maximum likelihood estimates are studied for the homogeneous case.

1. Introduction

Bogdanoff and Kozin, in their book (1985) define cumulative damage (CD) as the "irreversible accumulation of damage throughout life, that ultimately leads to failure". Such damage can be manifested by corrosion, cracks, physical wear in bearing, piston rings, locks, etc. We focus attention on damage processes that occur at random times, according to some non-homogeneous Poisson process. The amount of damage that accumulates follows a specified distribution. Thus, the amount of damage at time t, is a realization of a random process $\{Y(t), t \geq 0\}$, where $Y(t) \geq 0$ is a non-decreasing process with $Y(t) \to \infty$ a.s. as $t \to \infty$.

A system subjected to such a damage process fails at the first instant at which $Y(t) \geq \beta$, where $0 < \beta < \infty$ is a threshold specific to the system. Thus, the distribution of the failure times is a stopping time distribution. We present in the present paper the methodology of deriving these distributions. We are interested in particular in a family of non-homogeneous Poisson processes having an intensity function of the Weibull type, namely $\lambda(t) = (\lambda t)^\nu$, $0 < \lambda, \nu < \infty$. In Section 2 we specify compound non-homogeneous Poisson damage processes, and the distribution of the cumulative damage $Y(t)$, at time t. In Section 3 we derive the density and the reliability function of failure times driven by such processes. In particular we focus attention on cumulative Weibull processes with exponentially distributed damage amount in each occurrence. We investigate and illustrate the behavior of the distribution of failure times and the hazard function. In Section 4 we develop estimators of the parameters of the failure distribution in the homogeneous case ($\nu = 1$).

An extensive list of publications on damage processes is given in Bogdanoff and Kozin (1985). They provide empirical examples, and mention (p. 28) the non-homogeneous Poisson process with a Weibull intensity function. The theory for a discrete Markov chain model, having b states of damage is developed in this book. A recent paper on the subject is that of W. Kahle and H. Wendt (2000). They have modeled damage by a marked point process, and focus attention on

[1]Department of Mathematical Sciences, Binghamton University, SUNY, Binghamton, NY 13902-6000, USA. e-mail: shelly@math.binghamton.edu

Keywords and phrases: cumulative damage processes, non-homogeneous compound Poisson processes, distributions of stopping times, reliability functions, hazard functions, moment equation estimators, maximum likelihood estimators.

AMS 2000 subject classifications: 60J55, 60J75, 62N05.

doubly stochastic compound Poisson processes. Their formulation is close to ours, but they do not provide an explicit formula for the distribution of failure times. Other related papers are those concerned with shock models, like Esary, Marshall and Proshan (1973), Feng, Adachi and Kowada (1994), Shaked (1983), Soczyk (1987).

2. Compound cumulative damage processes

We consider cumulative damage processes (CDP) modeled by *non-homogeneous compound Poisson processes*. In this model, the system is subjected to shocks at random times, $0 < \tau_1 < \tau_2 < \cdots$, following a non-homogeneous Poisson process, with an *intensity function* $\lambda(t)$ (see Kao, 1997, pp. 56). The amount of damage to the system at the n-th shock is a random variable X_n, $n \geq 1$. We assume that $X_0 \equiv 0$, $X_1, X_2, \ldots$ are i.i.d., and that the sequence $\{X_n, n \geq 1\}$ is *independent* of $\{\tau_n, n \geq 1\}$.

Let $\{N(t), t \geq 0\}$ be a non-homogeneous Poisson counting process, with $N(0) = 0$ where

$$N(t) = \max\{n : \tau_n \leq t\}. \tag{1}$$

$\{N(t), t \geq 0\}$ is a process of independent increments such that, for any $0 \leq s < t < \infty$,

$$P\{N(t) - N(s) = n\} = e^{-(m(t)-m(s))} \frac{(m(t) - m(s))^n}{n!}, \tag{2}$$

$n = 0, 1, \ldots$, where $m(t) = \int_0^t \lambda(s)ds$, $0 \leq t < \infty$. The *compound damage process* (CDP) $\{Y(t), t \geq 0\}$ is defined as

$$Y(t) = \sum_{n=0}^{N(t)} X_n. \tag{3}$$

It is a compound non-homogeneous Poisson process. The compound Poisson Process (CPP) is the special case of a constant intensity function, $\lambda(t) = \lambda$, for all $0 < t < \infty$, $0 < \lambda < \infty$. We restrict attention in the present paper to the family of compound Weibull processes (CWP), in which $\lambda(t) = \lambda\nu(\lambda t)^{\nu-1}$, $0 < t < \infty$ for $0 < \lambda$, $\nu < \infty$. Furthermore, we assume that X_n, $n \geq 1$, are absolutely continuous random variables, having a common distribution function, F, and density f.

The cdf of $Y(t)$, at $t > 0$, has a discontinuity at $y = 0$, and is absolutely continuous on $0 < y < \infty$. It is given by

$$D(y; t) = \sum_{n=0}^{\infty} e^{-m(t)} \frac{(m(t))^n}{n!} F^{(n)}(y). \tag{4}$$

with $D(0; t) = \exp(-m(t))$, and $F^{(n)}$ is the n-fold convolution of F, i.e.,

$$F^{(n)}(y) = \begin{cases} F(y), & \text{if } n = 1 \\ \int_0^y f(x)F^{(n-1)}(y - x)dx, & \text{if } n \geq 2. \end{cases} \tag{5}$$

The defective density of $Y(t)$ on $(0, \infty)$ is

$$d(y; t) = \sum_{n=1}^{\infty} e^{-m(t)} \frac{(m(t))^n}{n!} f^{(n)}(y). \tag{6}$$

where $f^{(n)}$ is the n-fold convolution of the density f. We will use the notation $p(n;\mu)$ and $P(n;\mu)$ for the probability function and cdf, respectively, of the Poisson distribution with mean μ. Accordingly, the density of the CWP, at $0 < y < \infty$ and $0 < t < \infty$ is

$$d(y;t,\lambda,\nu) = \sum_{n=1}^{\infty} p(n;(\lambda t)^{\nu}) f^{(n)}(y), \tag{7}$$

and its cdf is

$$D(y;t,\lambda,\nu) = \sum_{n=0}^{\infty} p(n;(\lambda t)^{\nu}) F^{(n)}(y). \tag{8}$$

We consider a special case of these functions, when the amount of damage X_n is exponentially distributed, with parameter μ, i.e., $E\{X_n\} = \frac{1}{\mu}$. In this special case $f^{(n)}(y) = \mu p(n-1;\mu y)$ and $F^{(n)}(y) = 1 - P(n-1;\mu y)$. The results of this paper can be generalized to damage processes driven by compound renewal processes with any distribution F.

3. Cumulative damage failure distributions

A cumulative damage failure time is the stopping time

$$T(\beta) = \inf\{t > 0 : Y(t) \geq \beta\}, \tag{9}$$

where $0 < \beta < \infty$. Since $Y(t)$ is non-decreasing a.s., we immediately obtain that, in the continuous case,

$$P\{T(\beta) > t\} = D(\beta;t), \quad 0 < t < \infty. \tag{10}$$

This is the reliability (survival) function of the system. Thus, for the CWP, with general damage distribution,

$$P\{T(\beta) > t\} = \sum_{n=0}^{\infty} p(n;(\lambda t)^{\nu}) F^{(n)}(\beta). \tag{11}$$

In the special case of exponential damage distribution,

$$P\{T(\beta) > t\} = 1 - \sum_{n=1}^{\infty} p(n;(\lambda t)^{\nu}) P(n-1;\mu\beta). \tag{12}$$

We see in (3.4) that, in the exponential case, the distribution of $T(\beta)$ depends on μ and β only through $\zeta = \mu\beta = \beta/E\{X_1\}$. Accordingly, let $R(t;\lambda,\nu,\zeta)$ denote the reliability function of a system under CWP with exponential damage distribution (CWP/E).

Theorem 1. *Under* CWP/E *the reliability function is*

$$R(t;\lambda,\nu,\zeta) = \sum_{j=0}^{\infty} p(j;\zeta) P(j;(\lambda t)^{\nu}). \tag{13}$$

Proof. According to (3.4),

$$R(t;\lambda,\nu,\zeta) \quad = \quad 1 - \sum_{n=1}^{\infty} p(n;(\lambda t)^{\nu}) \sum_{j=0}^{n-1} p(j;\zeta)$$

$$
\begin{aligned}
&= 1 - \sum_{j=0}^{\infty} p(j;\zeta) \sum_{n=j+1}^{\infty} p(n;(\lambda t)^{\nu}) \\
&= 1 - \sum_{j=0}^{\infty} p(j;\zeta)(1 - P(j;(\lambda t)^{\nu})).
\end{aligned}
$$

This implies (3.5). □

It is obvious from (3.1) that $P\{T(\beta) < \infty\} = 1$ for any $0 < \beta < \infty$. This follows also from the following theorem.

Theorem 2. *Under* CWP/E, $R(0;\lambda,\nu,\zeta) = 1$, $R(t;\lambda,\nu,\zeta)$ *is strictly decreasing in* t, *for* (λ,ν,ζ) *fixed, and* $\lim_{t\to\infty} R(t;\lambda,\nu,\zeta) = 0$, *for any* (λ,ν,ζ) *in* $\mathbb{R}^{3+}$.

Proof. According to (3.5), since $\lim_{t\to 0} P(j;(\lambda t)^{\nu}) = 1$ for all $j = 0, 1, \ldots$ and any $0 < \lambda, \nu < \infty$, the bounded convergence theorem implies that

$$
\lim_{t\to 0} R(t;\lambda,\nu,\zeta) = \sum_{j=0}^{\infty} p(j;\zeta) \lim_{t\to 0} P(j;(\lambda t)^{\nu}) = 1.
$$

Furthermore, the Poisson family is an MLR family and $P(j;(\lambda t)^{\nu}) \downarrow t$. Hence, $R(t;\lambda,\nu,\zeta) \downarrow t$, i.e., $\frac{\partial}{\partial t} R(t;\lambda,\nu,\zeta) < 0$, for any fixed (λ,ν,ζ), $0 < \lambda,\nu,\zeta < \infty$. Finally, since $\lim_{t\to\infty} P(j;(\lambda t)^{\nu}) = 0$ for any fixed $j \geq 0$, $0 < \lambda, \nu < \infty$, the dominated convergence theorem implies that $\lim_{t\to\infty} R(t;\lambda,\nu,\zeta) = 0$, for any $0 < \lambda,\nu,\zeta < \infty$. □

Theorem 3. *Under* CWP/E, *the density of* $T(\zeta)$, $0 < \zeta < \infty$, *is*

$$
f(t;\lambda,\nu,\zeta) = \lambda\nu(\lambda t)^{\nu-1} \sum_{j=0}^{\infty} p(j;\zeta) p(j;(\lambda t)^{\nu}), \tag{14}
$$

and its m*-th moment,* $m \geq 1$, *is*

$$
E\{(T(\zeta))^m\} = \frac{1}{\lambda^m} \sum_{j=0}^{\infty} p(j;\zeta) \frac{\Gamma(j+1+\frac{m}{\nu})}{\Gamma(j+1)}. \tag{15}
$$

Proof. It is easy to verify that

$$
\frac{\partial}{\partial \omega} P(j;\omega) = -p(j;\omega), \quad 0 < \omega < \infty.
$$

Moreover,

$$
\begin{aligned}
f(t;\lambda,\nu,\zeta) &= -\frac{\partial}{\partial t} P\{T(\beta) > t\} \\
&= -\frac{\partial}{\partial t} \sum_{j=0}^{\infty} p(j;\zeta) P(j;(\lambda t)^{\nu}).
\end{aligned}
$$

This implies (3.6), since $R(t;\lambda,\nu,\zeta)$ is an analytic function of t, or by bounded convergence. To prove (3.7) we write

$$
E\{(T(\zeta))^m\} = \int_0^{\infty} t^m f(t;\lambda,\nu,\zeta) dt
$$

$$
\begin{aligned}
&= \nu\lambda^{\nu}\sum_{j=0}^{\infty} p(j;\zeta)\frac{\lambda^{\nu j}}{j!}\int_0^{\infty} t^{m+\nu(j+1)-1}e^{-(\lambda t)^{\nu}}dt \\
&= \sum_{j=0}^{\infty} p(j;\zeta)\frac{\lambda^{\nu(j+1)}}{j!}\int_0^{\infty} u^{\frac{m}{\nu}+j}e^{-\lambda^{\nu}u}du \\
&= \frac{1}{\lambda^m}\sum_{j=0}^{\infty} p(j;\zeta)\frac{\Gamma(j+1+\frac{m}{\nu})}{\Gamma(j+1)}.
\end{aligned}
$$

□

Corollary. *In the homogeneous case* ($\nu = 1$) *with exponential damage, the expected value, variance and coefficient of skewness of* $T(\zeta)$ *are, correspondingly,*

$$E\{T(\beta) \mid \lambda, \nu = 1, \zeta\} = \frac{1+\zeta}{\lambda}, \tag{16}$$

$$V\{T(\zeta) \mid \lambda, \nu = 1, \zeta\} = \frac{1+2\zeta}{\lambda^2} \tag{17}$$

and

$$\gamma_1(T(\zeta)) = \frac{2(1+3\zeta)}{(1+2\zeta)^{3/2}}. \tag{18}$$

Notice also that equation (3.7) shows that moments of $T(\zeta)$ of all orders exist, since moments of all orders of the Poisson distribution exist. In Figure 1 we present several densities of $T(\zeta)$, for $\lambda = 1$, $\zeta = 5$ and $\nu = 1.1, 1, .9$. According to eq. (3.6),

$$\lim_{t\to 0} f(t;\lambda,\nu,\zeta) = \begin{cases} \infty, & \text{if } \nu < 1 \\ \lambda e^{-\zeta}, & \text{if } \nu = 1 \\ 0, & \text{if } \nu > 1. \end{cases} \tag{19}$$

Indeed, $\lim_{t\to 0} p(j;(\lambda t)^{\nu}) = I\{j = 0\}$, i.e., 1 if $j = 0$ and 0 otherwise. Thus, $\lim_{t\to 0}\sum_{j=0}^{\infty} p(j;\zeta)p(j;(\lambda t)^{\nu}) = p(0;\zeta) = e^{-\zeta}$. The densities $f(t;\lambda,\nu,\zeta)$ are uni-modal whenever $\nu \geq 1$, and bi-modal when $\nu < 1$. Figure 1 does not show the behavior of these densities in the interval $(0, 1)$. We see that the density becomes more symmetric as ζ grows. Indeed, $\frac{\partial}{\partial\zeta}\gamma_1(T(\zeta)) = -\frac{6\zeta}{(1+2\zeta)^{5/2}} < 0$ for all $0 < \zeta < \infty$.

From eq. (3.5) we obtain immediately that the reliability function $R(t;\lambda,\nu,\zeta)$, is a strictly increasing function of ζ, for each fixed (t,λ,ν). This result is obvious from (3.1) if $\mu = 1$. Generally, for fixed t, λ, ν $P(j;(\lambda t)^{\nu})$ is an increasing function of j. Hence, since the Poisson family $\{p(\cdot;\zeta), 0 < \zeta < \infty\}$ is a monotone likelihood ratio family (MLR), $E_{\zeta}\{P(J;(\lambda t)^{\nu})\}$ is an increasing function of ζ.

The hazard function under CWP/E damage processes is

$$h(t;\lambda,\nu,\zeta) = \frac{\lambda\nu(\lambda t)^{\nu-1}\sum_{j=0}^{\infty} p(j;\zeta)p(j;(\lambda t)^{\nu})}{\sum_{j=0}^{\infty} p(j;\zeta)P(j;(\lambda t)^{\nu})}. \tag{20}$$

We obtain from (3.11) since $\lim_{t\to 0} P(j;(\lambda t)^{\nu}) = 1$ for all $j \geq 0$, that,

$$\lim_{t\to 0} h(t;\lambda,\nu,\zeta\}) = \begin{cases} \infty, & \text{if } 0 < \nu < 1 \\ \lambda e^{-\zeta}, & \text{if } \nu = 1 \\ 0, & \text{if } \nu > 1. \end{cases} \tag{21}$$

In Figure 2 we illustrate the hazard function (3.12) for $\lambda = 1$, $\zeta = 5$ and $\nu = .53, .55, .57$.

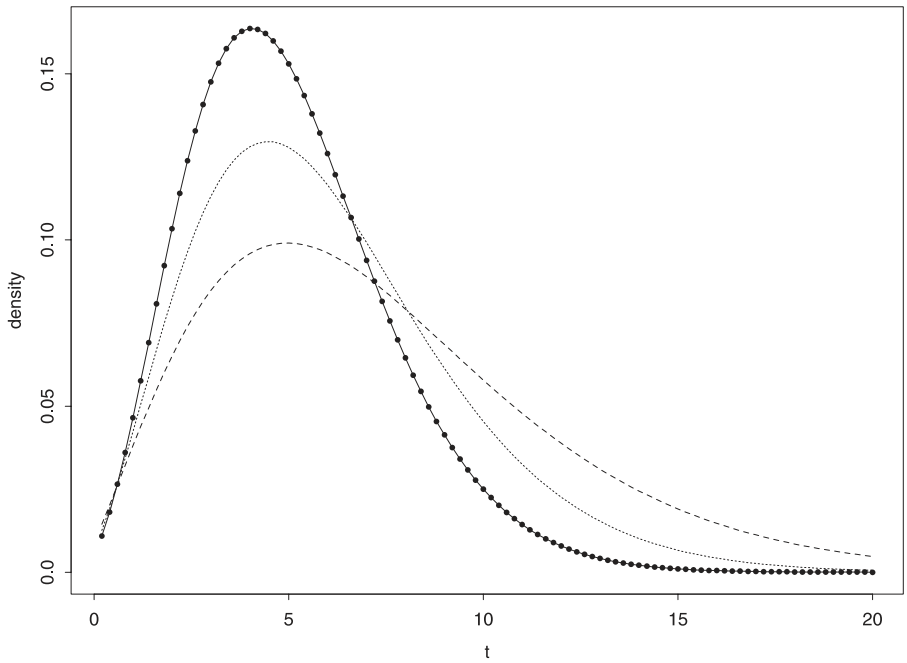


Figure 1: Densities of $T(\zeta)$, $\lambda = 1$, $\zeta = 5$, $\nu = 1.1$ —●—, $\nu = 1.0$ $\cdots$, $\nu = 0.9$ - - -

Similar types of hazard functions were discussed by Aalen and Gjesing (2003).

We examine now the asymptotic behavior of the hazard function (3.12), as $t \to \infty$. Make first the transformation $u = (\lambda t)^{\nu}$. In terms of u, the hazard function is

$$h^*(u; \lambda, \nu, \zeta) = \lambda \nu u^{1-1/\nu} \cdot \frac{E_\zeta\{p(J;u)\}}{E_\zeta\{P(J;u)\}}, \tag{22}$$

where $J \sim \text{Pois}(\zeta)$.

Theorem 4. *For a fixed λ, ν, ζ, the asymptotic behavior of the hazard function is*

$$\lim_{u\to\infty} h^*(u; \lambda, \nu, \zeta) = \begin{cases} \infty, & \text{if } \nu > 1 \\ \lambda, & \text{if } \nu = 1 \\ 0, & \text{if } \nu < 1. \end{cases} \tag{23}$$

Proof. Since $p(j;u) \le P(j;u)$ for $j = 0, 1, \ldots$ and each u, $0 < u < \infty$,

$$\overline{\lim_{u\to\infty}} \frac{E_\zeta\{p(J;u)\}}{E_\zeta\{P(J;u)\}} \le 1. \tag{24}$$

We now prove that

$$\lim_{u\to\infty} \frac{E_\zeta\{p(J;u)\}}{E_\zeta\{P(J;u)\}} = 1. \tag{25}$$

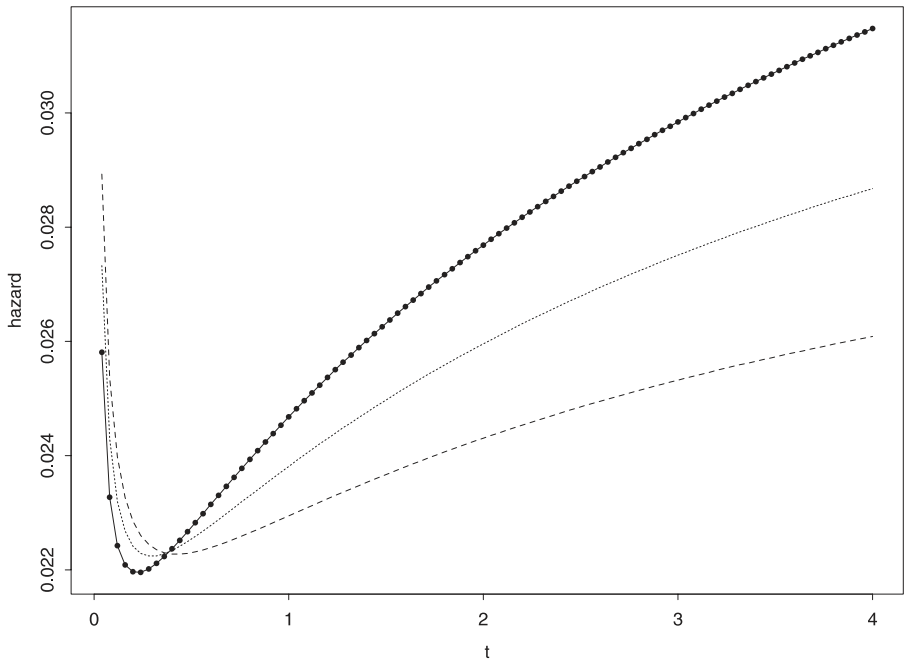


Figure 2. Hazard Functions, $\lambda = 1$, $\zeta = 5$, $\nu = .57$ —●—, $\nu = .55$ $\cdots$, $\nu = .53$ - - -

First, by dominated convergence, $\lim_{u\to\infty} E_\zeta(p(J;u)\} = E_\zeta\{\lim_{u\to\infty} p(J;u)\} = 0$. Similarly, $\lim_{u\to\infty} E_\zeta\{P(J;u)\} = 0$. By L'Hospital rule,

$$\begin{aligned}
\lim_{u\to\infty} \frac{E_\zeta\{p(J;u)\}}{E_\zeta\{P(J;u)\}} &= \lim_{u\to\infty} \frac{\frac{d}{du}E_\zeta\{p(J;u)\}}{\frac{d}{du}E_\zeta\{P(J;u)\}} \\
&= \lim_{u\to\infty} \frac{E_\zeta\{p(J;u) - p(J-1;u)\}}{E_\zeta\{p(J;u)\}} \\
&= 1 - \lim_{u\to\infty} \frac{\sum_{j=0}^{\infty} p(n+1;\zeta)p(n;u)}{\sum_{n=0}^{\infty} p(n;\zeta)p(n;u)}.
\end{aligned}$$

Furthermore,

$$\frac{\sum_{n=0}^{\infty} p(n+1;\zeta)p(n;u)}{\sum_{n=0}^{\infty} p(n;\zeta)p(n;u)} = \zeta \frac{\sum_{n=0}^{\infty} \frac{1}{n+1} p(n;\zeta)p(n;u)}{\sum_{n=0}^{\infty} p(n;\zeta)p(n;u)}$$

Fix a positive integer K (arbitrary). Then,

$$\begin{aligned} R(\zeta,u) &= \frac{\sum_{n=0}^{\infty}\frac{1}{n+1}p(n;\zeta)p(n;u)}{\sum_{n=0}^{\infty}p(n;\zeta)p(n;u)} \\ &\le \frac{\sum_{n=0}^{K}\frac{1}{n+1}p(n;\zeta)p(n;u)+\frac{1}{K+2}\sum_{n=K+1}^{\infty}p(n;\zeta)p(n;u)}{\sum_{n=0}^{K}p(n;\zeta)p(n;u)+\sum_{n=K+1}^{\infty}p(n;\zeta)p(n;u)} \end{aligned} \tag{26}$$

Finally, since $p(n;u)\to 0$ as $u\to\infty$ for each $n=0,1,\ldots,$

$$\lim_{u\to\infty}\sum_{j=0}^{K}\frac{1}{j+1}p(j;\zeta)p(j;u)=\lim_{u\to\infty}\sum_{j=0}^{K}p(j;\zeta)p(j;u)=0.$$

Thus,

$$\begin{aligned} \overline{\lim_{u\to\infty}}R(\zeta;u) &\le \frac{1}{K+2}\lim_{u\to\infty}\frac{\sum_{j=K+1}^{\infty}p(j;\zeta)p(j;u)}{\sum_{j=K+1}^{\infty}p(j;\zeta)p(j;u)} \\ &= \frac{1}{K+2}, \quad \text{for all fixed } \zeta. \end{aligned}$$

□

In Figure 3 we illustrate a hazard function for $\lambda=1$, $\zeta=5$, $\nu=0.5$.

4. Estimation of parameters

Let $T_1,T_2,\ldots,T_n$ be i.i.d. random failure times following CWP/E. The likelihood function of the parameters (λ,ν,ζ) is

$$L(\lambda,\nu,\zeta;T_1,\ldots,T_n)=(\lambda)^{n\nu}\nu^n\left(\prod_{i=1}^{n}T_i^{\nu-1}\right)\cdot\prod_{i=1}^{n}\sum_{j=0}^{\infty}p(j;\zeta)p(j;(\lambda T_i)^\nu). \tag{27}$$

Accordingly, the minimal sufficient statistic is the trivial one $(T_{(1)},\ldots,T_{(n)})$, where $0<T_{(1)}\le T_{(2)}\le\cdots\le T_{(n)}$.

4.1. Moment equations estimators of λ, ζ in the homogeneous case, $\nu=1$.

Let $M_1=\frac{1}{n}\sum_{i=1}^{n}T_i$ and $M_2=\frac{1}{n}\sum_{i=1}^{n}T_i^2$ be the first two sample moments. The moment equations estimators (MEE) of λ and ζ are obtained by solving the equations,

$$\frac{1+\hat{\zeta}}{\hat{\lambda}}=M_1 \tag{28}$$

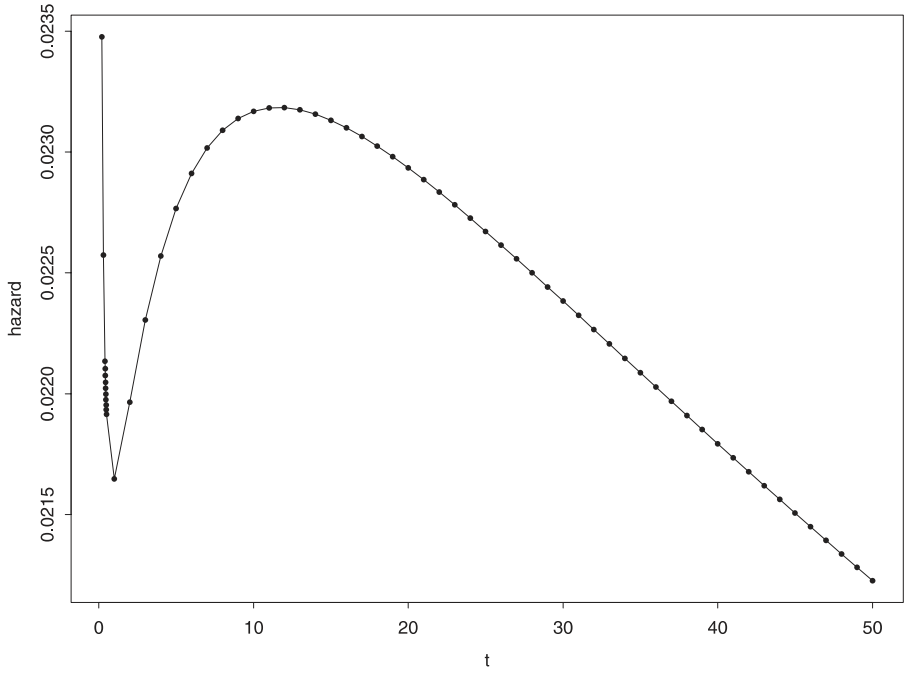


Figure 3. Hazard Function for $\lambda = 1$, $\nu = .5$, $\zeta = 5$

and

$$\frac{2 + 4\hat{\zeta} + \hat{\zeta}^2}{\hat{\lambda}^2} = M_2. \tag{29}$$

Or, equivalently,

$$\hat{\lambda} = \frac{1 + \hat{\zeta}}{M_1}, \tag{30}$$

and $\hat{\zeta}$ is the positive root of the quadratic equation

$$\hat{\zeta}^2 \left(1 - \frac{M_1^2}{M_2}\right) - 2\hat{\zeta}\left(\frac{2M_1^2}{M_2} - 1\right) - \left(\frac{2M_1^2}{M_2} - 1\right) = 0. \tag{31}$$

A real root exists provided $M_2 < 2M_1^2$. Since $2M_1^2 - M_2 \xrightarrow[n\to\infty]{a.s.} (\frac{\zeta}{\lambda})^2 > 0$, an MEE exists for n sufficiently large. It is given by

$$\hat{\zeta} = \frac{(2M_1^2 - M_2)^{1/2}(M_1 + (2M_1^2 - M_2)^{1/2})}{M_2 - M_1^2}. \tag{32}$$

Both $\hat{\lambda}$ and $\hat{\zeta}$ are strongly consistent estimators of λ and ζ, respectively. The mean squared errors of these estimators can be approximated by the delta method. We obtain

$$\text{MSE}\{\hat{\lambda}\} = \frac{\lambda^2}{n} \cdot \frac{1 + 12\zeta + 58\zeta^2 + 144\zeta^3 + 192\zeta^4 + 128\zeta^5 + 32\zeta^6}{\zeta^2(1 + 2\zeta)^4} + O\left(\frac{1}{n^2}\right), \tag{33}$$

and

$$\text{MSE}\{\hat{\zeta}\} = \frac{1}{n\zeta^2}(2(1 + \zeta)^4 - (1 + \zeta)^2 - \zeta^2) + O\left(\frac{1}{n^2}\right). \tag{34}$$

In the following table we compare the values of the MSE, as approximated by eq.'s (4.7) and (4.8), to those obtained by simulations. When $\nu = 1$ the distribution of T is like that of $\chi^2[2;\zeta]/(2\lambda)$, where $\chi^2[2;\zeta]$ is a non-central chi-square with 2 degrees of freedom, and parameter of non-centrality ζ. Thus

$$T \sim (N_1^2(\sqrt{\zeta},1) + N_2^2(\sqrt{\zeta},1))/(2\lambda),$$

where $N_i(\sqrt{\zeta},1)$ $(i = 1,2)$ are i.i.d. normal random variables with mean $\sqrt{\zeta}$ and variance 1. 10,000 simulation runs yield the following results

Table 1. *MSE Values of the MEE By Delta Method and By Simulations*

λ	ζ	n	Delta Method		Simulation	
			$\hat{\lambda}$	$\hat{\zeta}$	$\hat{\lambda}$	$\hat{\zeta}$
1	5	50	0.0568	2.0248	0.0744	2.6286
		100	0.0284	1.0124	0.0322	1.1475
2	5	50	0.2272	2.0248	0.3058	2.6746
		100	0.1136	1.0124	0.1323	1.1828

We notice that the delta method for samples of size 50 or 100 is not sufficiently accurate. It yields values which are significantly smaller than those of the simulation. Also, since the MEE $\hat{\lambda}$ and $\hat{\zeta}$ are continuously differentiable functions of the sample moments M_1 and M_2, the asymptotic distributions of $\hat{\lambda}$ and $\hat{\zeta}$ are normal, with means λ and ζ and variances given by (4.7) and (4.8).

4.2. Maximum likelihood estimators, $\nu = 1$

The log-likelihood function of (λ,ζ), given $\mathbf{T}^{(n)}$ is

$$l(\lambda,\zeta;\mathbf{T}^{(n)}) = n\log\lambda + \sum_{i=1}^{n}\log E_\zeta\{p(J;\lambda T_i)\}, \tag{35}$$

where $J \sim \text{Pois}(\zeta)$. Accordingly, the score functions are

$$\frac{\partial}{\partial\lambda}l(\lambda,\zeta;\mathbf{T}^{(n)}) = \frac{n}{\lambda} - \sum_{i=1}^{n}T_i + \zeta\sum_{i=1}^{n}T_iW(\lambda,\zeta,T_i), \tag{36}$$

and

$$\frac{\partial}{\partial\zeta}l(\lambda,\zeta,\mathbf{T}^{(n)}) = -n + \lambda\sum_{i=1}^{n}T_iW(\lambda,\zeta,T_i), \tag{37}$$

where

$$W(\lambda,\zeta,T) = \frac{E_\zeta\{\frac{1}{1+J}p(J;\lambda T)\}}{E_\zeta\{p(J;\lambda T)\}}. \tag{38}$$

Let $\hat{\lambda}$ and $\hat{\zeta}$ be the maximum likelihood estimators (MLE) of λ and ζ, respectively.

From (4.10) and (4.11) we obtain that, as in (4.4),

$$\hat{\lambda} = \frac{1+\hat{\zeta}}{M_1}. \tag{39}$$

Substituting $\hat{\lambda}$ in (4.11) we obtain the function

$$\overset{\bullet}{l}(\zeta) = (1+\zeta)\sum_{i=1}^{n} U_i W\left(\frac{1+\zeta}{M_1}, \zeta, M_1 U_i\right) - n, \tag{40}$$

where $U_i = T_i/M_1$. More specifically,

$$\overset{\bullet}{l}(\zeta) = (1+\zeta)\sum_{i=1}^{n} U_i \frac{E_\zeta\{\frac{1}{1+J} p(J;(1+\zeta)U_i)\}}{E_\zeta\{p(J;(1+\zeta)U_i)\}} - n. \tag{41}$$

Notice that $\overset{\bullet}{l}(0) = 0$. The MLE of ζ, $\hat{\zeta}$, is the positive root of $\overset{\bullet}{l}(\zeta) \equiv 0$. $N = 1,000$ simulation runs gave the following estimates of the MSE of $\hat{\lambda}$ and $\hat{\zeta}$, when $\lambda = 1$, $\zeta = 5$ and $n = 50$, namely:

$$\widehat{\mathrm{MSE}}(\hat{\lambda}) = 0.06015 \quad \text{and} \quad \widehat{\mathrm{MSE}}(\hat{\zeta}) = 2.13027.$$

As expected, these estimators of the MSE of $\hat{\lambda}$ and $\hat{\zeta}$ are smaller than those of the MEE estimates, given in Table 1. The asymptotic distribution of the MLE vector $(\hat{\lambda}, \hat{\zeta})$ is bivariate normal with mean (λ, ζ) and covariance matrix AV, which is the inverse of the Fisher information matrix. The asymptotic variance-covariance matrix of the MLE can be estimated by simulation. $N = 10,000$ simulation runs gave, for the case of $\lambda = 1$, $\zeta = 5$ the asymptotic variance-covariance matrix

$$AV = \frac{1}{n}\begin{bmatrix} 2.33917 & 13.04000 \\ 13.04000 & 83.30706 \end{bmatrix}.$$

Thus, the asymptotic variance of $\hat{\zeta}$ for $n = 50$ is $AV(\hat{\zeta}) = \frac{83.30706}{50} = 1.66614$. We see that the estimated variance of $\hat{\zeta}$ is, as in the case of the MEE, considerably larger than its asymptotic variance. The convergence is apparently very slow.

References

[1] Aalen, O. O. and Gjessing, H. K. (2003). A look behind survival data: underlying processes and quasi-stationarity, *Mathematical and Statistical Methods in Reliability*, Bo H. Lindquist & K. A. Doksum, Editors, pp. 221, World Scientific, New Jersey.

[2] Bogdanoff, J. L. and Kozim, F., *Probabilistic Models of Cumulative Damage*, John Wiley, New York, 1985.

[3] Esary, J. D., Marshall, A. W. and Proshan, F. (1973). Shock models and wear processes, *The Annals of Probability*, **1**:627-649.

[4] Feng, W., Adachi, K., and Kowada, M. (1994). Optimal replacement under additive damage in a Poisson random environment, *Commun. Statist.-Stochastic Models*, **10**:679-700.

[5] Kao, Edward, P. C., *An Introduction to Stochastic Processes*, Duxbury Press, New York, 1997.

[6] Kahle, W. and Wendt, H. (2000). Statistical analysis of damage processes, in *Recent Advances in Reliability Theory, Methodology, Practice and Inference*, N. Limuios and M. Nikulin (Editors), Birkhäuser, Boston, pp. 199.

[7] Karlin, S. (1957). Polya type distributions, II, *Ann. Math. Statist.*, **28**:281-308.

[8] Shaked, M. (1983). Wear and damage processes from shock models, *Reliability Theory and Models*, Proc. Symp./N.C. 43-64.

[9] Sobczyk, K. (1987). Stochastic models for fatigue damage of materials, *Adv. Appl. Prob.*, **19**:652-673.

A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 408–417
© Institute of Mathematical Statistics, 2004

Conversations with Herman Rubin

Mary Ellen Bock[1]

Purdue University

Abstract: Herman Rubin was born October 27, 1926 in Chicago, Illinois. He obtained his Ph.D. in Mathematics from the University of Chicago in 1948 at the age of 21. He has been on the faculty of Stanford University, the University of Oregon, Michigan State University and Purdue University, where he is currently Professor of Statistics and Professor of Mathematics. He is a Fellow of the Institute of Mathematical Statistics and of the American Association for the Advancement of Science as well as a member of Sigma Xi.

He is well known for his broad ranging mathematical research interests and for fundamental contributions in Bayesian decision theory, in set theory, in estimations for simultaneous equations, in probability and in asymptotic statistics.

These conversations took place during the 2003–2004 academic year at Purdue University.

Herman, it is great that the IMS is bringing out this Festschrift for you. I am delighted to be able to prepare this interview with you. I guess we always want to know about childhood. So Herman, where did you grow up?

I was born in Chicago, Illinois, and grew up there, the oldest of three children. Both of my parents were immigrants, my father from Russia and my mother from Russian-occupied Poland. My mother's sister was also an immigrant and she taught me to read at the age of three.

What was your educational background? Did you receive special training in mathematics?

I went to the Chicago public schools for grammar school and was a voracious reader in the public library. But the material was organized by grade level and I did not find much on mathematics beyond arithmetic. But the summer before I went to high school I discovered algebra when I came upon a book about it while visiting New York City. After reading the book, I tested out of algebra in the first month of the first year of high school. In high school I found many more advanced books about mathematics in the public library; I taught myself material through calculus while taking plane geometry in high school. After two years at the public high schools, I was given a scholarship to a combined high school/college program at the University of Chicago. I could have graduated high school after a total of three years but delayed the official high school graduation by one year because I could take more college courses and not pay the college tuition. I received the high school diploma in June of 1943, the bachelor degree SB (Mathematics major with Physics minor) in December of 1944 and the master degree SM (Mathematics) in March of 1945. At the University of Chicago, almost all of my courses beyond the bachelor's were in abstract mathematics but my Ph.D. dissertation was in statistics.

How did you get interested in the field of statistics if most of your courses were in abstract mathematics?

My interest developed during my stint at the Cowles Commission for Research in Economics (CCRE) which was housed at the University of Chicago. In 1944

[1]Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 42907-20689, USA. e-mail: mbock@purdue.edu

CCRE needed a mathematics research assistant because their current assistant was being drafted into the U.S. military. At the time I was a student in the undergraduate/graduate program at the University of Chicago and, aside from my mathematical abilities, one of my qualifications was that I was too young to be drafted. So in July of 1944 at the age of seventeen I became a research assistant for CCRE.

I became interested in statistics because the leader of CCRE, Jacob Marschak, who took over in 1943, had decided to concentrate the work of the group on the problems of stochastic simultaneous equations found in economics.

Who worked with you when you joined CCRE?

My initial work was with Tjalling Koopmans who had joined CCRE at the same time as I. He was brought in to concentrate on the mathematical aspects. My first paper was a solution to a problem of Koopmans for the approximate distribution of the circular serial correlation coefficients under the null hypothesis and it appeared in the *Annals of Mathematical Statistics* in 1945. The main problem I worked on with Koopmans was to estimate the parameters of a system of stochastic equations including lags and to derive their properties. (Individual equations might have more than one dependent variable and least squares was already known to be inconsistent when applied individually to each equation.) I developed some Maximum Likelihood techniques and their properties for the time series lags to attack the problem.

I understand that the work at CCRE was interrupted.

The work was interrupted because I was drafted into the U.S. Army in March, 1945, at the age of 18. The bulk of the work I mentioned was published as a joint paper by myself with Koopmans and Roy Leipnik. (Roy was a research assistant in CCRE from February, 1945, to July, 1946, and took over the work with Koopmans after I was drafted.)

I was discharged from the Army in December, 1945, and returned to the University of Chicago as a graduate student and CCRE as a research assistant in January, 1946. (CCRE promoted me to research associate in November, 1946.)

Who worked with you on your return to CCRE?

I began to work with Theodore W. Anderson who had joined the CCRE as a research associate in November, 1945, in my absence. One source of inspiration for our work was a talk I heard after my return given by the biologist Sewall Wright. (He had given a general formulation for the problem of solving simultaneous stochastic equations in 1919.) I realized that factor analysis was another example of simultaneous stochastic equations and this led to a paper on it with Anderson.

Anderson and I collaborated on three papers. The first paper developed the maximum likelihood estimator of the coefficients of a single equation in a system of stochastic equations; the estimator is now known as the Limited Information Maximum Likelihood (LIML) estimator. The second paper developed the large-sample distribution theory. The LIML estimator had been developed in Anderson's 1945 dissertation. Our third joint paper developed maximum likelihood methods for factor analysis models with different identification conditions. It was a pretty innovative paper at the time.

Another source of interesting questions was Meyer A. Girshick. Early in 1946 Koopmans gave me a letter from Girshick about the problem of estimating a single equation (with more than one dependent variable) without estimating the entire complete system of equations. (A system of equations is complete if there are enough equations of the right sort so that all the coefficients could be consistently

estimated, essentially a multivariate regression problem.) I developed it somewhat and then collaborated on further aspects with T. W. Anderson. This work (with credit to Girshick) appeared finally in 1949 and 1950 in the *Annals of Mathematical Statistics.* The publication was somewhat delayed because in those days it was a major job (without the benefit of email) to communicate with the referees and my coauthor Anderson who was in Sweden during the 1947–1948 academic year. (He left CCRE in September, 1946, to go to Columbia.)

What about the Ph.D. degree?

I received the Ph.D. degree from the University of Chicago in March, 1948, at the age of 21. My official advisor was Paul Halmos in the Department of Mathematics. The dissertation topic grew out of my work at CCRE. It involved extending the original problem of Girshick of estimating a single equation to that of estimating a subsystem of equations without estimating the entire complete system of equations. The dissertation was typed up while I was on leave from CCRE as a post-doc at the Institute for Advanced Study in Princeton during the academic year of 1947–1948.

You have made major contributions to the field of asymptotics. Why do you feel asymptotics are important?

The need for asymptotics at CCRE inspired me and this culminated in my first major insights in 1949. Some of my contributions to the problem were the asymptotic theorems on limiting distributions which were never published. I introduced the idea of a random function into the generalization of the Slutsky Theorems. James Hannan and Vaclav Fabian gave the proofs in their book crediting me. For inspiration I used general topology (although metric topology is adequate). For me, the more I generalize the problem to an area of abstract mathematics the easier it is for me to understand it since I can get rid of the part which doesn't add to the meaning of the problem. Even when I computed something, if I could generalize it, then it led to insight. I know that is not how most people like to do mathematics.

You have had an abiding interest in computing. What was it like then?

At CCRE, computations were done with electromechanical desk calculators and a staff of three operated the calculators. Computations BC (Before Computers) were much slower. I was in charge of computing at CCRE until Herman Chernoff took over when I left for Princeton in August of 1947; we had some pretty funny experiences making the equipment work. (He had come there as a research assistant to CCRE in July, 1947.)

You are well known for your interests in statistical decision theory. Was it influenced by the CCRE experience?

The CCRE emphasis on economics was a factor. The idea of a utility scale for actions assuming that the state of nature is fully known, which goes back much farther, was important in quantitative economics for a long time. However, no essential progress had been made in getting a clear scale until the von Neumann – Morgenstern axioms for cardinal utility appeared in their book "Theory of Games and Economic Behavior" in 1944. One of their key contributions was the use of randomization. Researchers at the CCRE in 1947 were considering extending the ideas to unknown states of nature while I was there. I observed that adding one simple axiom made the utility for unknown states of nature a positive linear functional of the utility functions indexed by the given states of nature. (This is essentially the prior Bayes approach.)

In the early years of decision theory, the main progress was made in proving theorems and refining the concepts, and I had my share in this. Stanford was a

center of activity in this and I went there after leaving CCRE. I had various degrees of collaboration with Blackwell, Girshick, Karlin, and Chernoff, and numerous discussions with Stein. Four dissertations on decision theory were written under me.

You have worked on a variety of problems in probability, particularly stochastic integration, characterizations and infinite divisibility. You have collaborated with numerous people on these. How was that experience?

Yes, I have collaborated with C. R. Rao, Burgess Davis, Tom Sellke, Anirban DasGupta, Steve Samuels, Prem Puri, Rick Vitale and many others on questions in probability. The results with C. R. Rao got to be known as Rao-Rubin theorems; we were both visiting Stanford that year. Burgess Davis and Jeesen Chen asked an interesting question about uniform empirical processes. Tom Sellke and I worked on several Choquet type decomposition problems in the eighties. I have always enjoyed using characteristic functions as a tool, as those works did. I am glad my book length review with Arup Bose and Anirban DasGupta on infinite divisibility got published a couple of years ago; we worked many years on that one. With Prem Puri and Steve Samuels, the works were more in applied probability, but they were good problems. And, you mention stochastic integration. Yes, I too had thought of the Stratonovich integral. I gave a talk introducing the idea behind the Stratonovich integral at the IMS meeting in Seattle in 1956. My Ph.D. student Don Fisk later wrote a thesis on it in 1961. I myself did not write it up or pursue it formally. Probability questions are always interesting.

In the fifties, you collaborated with Karlin on introducing monotone likelihood ratio. This has had a very major impact. How did that idea originate?

Steve Allen, a Ph.D. student under Girshick, had come up with a proof that in the exponential family, monotone procedures are essentially complete for one-sided testing problems. I first wrote a technical report. Karlin and I realized it works for monotone likelihood ratio. We generalized that result of Allen and gave applications. Yes, it later led to concepts such as total positivity. Karlin has written much about it.

What did you do with Chernoff?

That was the beginning of my interest in the discontinuous density problems; we had a paper together in the third Berkeley symposium. But the relationship extended beyond professional collaboration.

You have a number of publications in set theory. How did this interest arise?

I was always interested in set theory and while in graduate school at Chicago I took a course from the topologist John L. Kelley which piqued it even more. There is a version of set theory that he showed me (called the Morse-Kelley set theory) which is stronger than the usual set theory because you can prove the consistency of the usual set theories (such as the Zermelo-Frankel or the von Neuman-Bernays-Godel) in the Morse-Kelley system.

From CCRE I went to Stanford's Department of Statistics in 1949 as an Assistant Professor and eventually met Jean Hirsh when she arrived later as a mathematics Ph.D. graduate student in logic there. We married in 1952. Her interests in logic and mine in set theory eventually led to a professional collaboration.

Later Pat Suppes was teaching a class on set theory for which I gave some lectures on the axiom of choice. Professor Suppes who knew both of us suggested

that Jean and I write a book on the various equivalents of the axiom of choice. (Jean received the Ph.D. in mathematics for her work in logic in 1955 and Suppes was her advisor.) After at least eight years, two moves and two children, we finally finished the book.

With two parents with Ph.D.'s in mathematics, were the children also interested in mathematics?

Arthur was the oldest (born in 1956) and went on to get a Ph.D. in mathematics from California Institute of Technology (at the age of 22) after being a Putnam Fellow four times. Arthur and Paul Erdós wrote a paper together. Leonore who was born in 1958 received a bachelor's degree with honors jointly in mathematics and chemistry from Michigan State University and went on to get a Ph.D. in chemistry from Carnegie Mellon.

You mentioned several moves. Where did you go?

After I left Stanford in 1955, I went to the Department of Mathematics at the University of Oregon for four years. (Because of nepotism rules, Jean was not allowed to have a regular position at the same university or even paid by the State of Oregon.) I had some collaborations with Howard Tucker and A.T. Bharucha-Reid at Oregon. From Oregon I went to Michigan State University's Department of Statistics in 1959. Again, Jean could not be hired because of nepotism rules. The set theory book on the axiom of choice by Jean and me was published while we were at Michigan State.

Most of my collaborations at Michigan State were with Martin Fox in decision theory, game theory and functions of Markov states. It was also the start of a collaboration with J. Sethuraman. We did some work on what is now called moderate deviations. We also later collaborated on Bayes risk efficiency.

Then in 1967 we both came to Purdue where Jean received an offer from the Department of Mathematics that included tenure. I joined the Department of Statistics and the Department of Mathematics as a full Professor and Jean joined Math as an assistant professor. I have been here ever since and my wife Jean was a full Professor of Mathematics here until her death in 2002. She is honored by an annual seminar and remembered for her support for women faculty in academia. She started a scholarship fund for mathematics students in her will.

One of your strong ongoing interests is in prior Bayesian robustness. How do you describe it?

One of the difficulties of Bayesian analysis is coming up with a good prior and loss function. (I have been saying for years that the prior and the loss cannot be separated. The Carnegie Mellon school is doing some work on that now.) When I talk about prior Bayesian robustness I assume that one does not yet see the random observation X whose distribution depends on the unknown state of nature. One considers the choice of different priors for which one averages over the possible states of nature and over the possible random observations. This is different from posterior Bayesian robustness in which one considers the choice of different priors given the random observation X whose distribution depends on the unknown state of nature. If you can get posterior Bayesian robustness, then you automatically get prior Bayesian robustness but seldom are we so lucky as to find posterior Bayesian robustness. It is actually the axioms of utility that decree we should worry about prior Bayesian robustness. When I am faced with a choice among priors, all of which seem about the same to me, then I am very concerned about the possible alternative consequences of applying either one if it is drastically wrong. For instance, suppose I

am using squared error loss to estimate the mean of a normal random variable with variance one. The first prior for the unknown mean might be normal with mean zero and standard deviation 10 while a second prior for the unknown mean might be normal with mean zero and standard deviation 1000. Now using the first prior could be disastrous (in terms of a loss that is averaged over the values of the state of nature as well as the possible mean values) when the second prior is appropriate. Yet using the second prior would not be so bad if the first prior were appropriate. In contrast for the posterior Bayesian robustness approach, if the observation X is large, then the posterior loss is bad in either case if the wrong prior is used.

You have had a long term interest in random number generation. What inspired you?

I heard that a professor at Columbia had announced to his class that he would give a midterm in each of five three-week periods, the particular week to be chosen at random by tossing a coin. Finding an efficient way to do this was an interesting problem to me. I observed that generating all five results at once was far more efficient that generating the results one at a time from a discrete distribution. This eventually led to less trivial questions and was the start of my interest in efficient methods for generating random numbers.

The two main problems I find interesting are the following: how to get lots of random numbers which are independent and uniform; and how to turn them into independent random numbers from some other distributions.

In the case of the first main problem, when generating independent uniform random variables, most people use pseudo random numbers. It is almost impossible to prove that they have all the desired properties. Of course, they fail the test that they come from the pseudo random number generator! For physical random numbers, one can question the accuracy of the model for the physical process. (I have a technical report about paradoxes caused by the effect of dead time.) My personal preference is to use a stream of physical random numbers and a stream of pseudo random numbers to produce a stream of random numbers whose qualities should be at least as good as either of the original two streams.

In the case of the second main problem, even when you have independent uniform random variables, the problem of using them to generate variables from other distributions is sometimes hard. The basic issue of efficiency is not the question of the number of bits used but rather the computational cost... and this is a complex question. I have some technical reports on these issues.

Computing issues in probability and statistics have been a topic of research for you, too. What are your comments?

Computation is an obvious issue in the generation of random numbers. But computation of probabilities is also important and it is often best done through the use of characteristic functions (i.e. Fourier transforms). I find that reasonably efficient computational procedures require complex integration... and that requires a knowledge of analytic functions.

Another important area is the computation of Maximum Likelihood estimates (MLE). This typically requires more expertise in analysis than is usually expected in most statistics graduate programs. For Bayes procedures, integration computing problems are commonplace. Many have pointed out the difficulties in posterior Bayes computations. Simulation is another area of computing problems. In 1970, I was using simulation to compute theoretical expectations for Kolmogorov-Smirnov and Kuiper statistics under nonnull hypotheses. The finite sample distribution was approximated by a modification of a Brownian Bridge. My first observation was

that I could not just simulate at a finite number of points because the Brownian Bridge changed too rapidly. This was handled by simulating the max and the min in various intervals independently (even though the max and min are not independent) and it worked quite well. The reason it worked well was that the probability that the max and min of the whole process are both in the same interval is extremely small. This points out that it is often necessary to do analysis before numerical analysis.

How has the internet affected your work?

I contribute to newsgroups and give advice to those that ask. I also join extended discussions on how things should be done. In general, I find out things that I might not otherwise know because the internet puts me in contact with lots of people. It is much easier to collaborate and I think it is a great advantage for research. You see, you can put a hard problem on the web and get help from experts.

I have not yet directly engaged in internet communication in producing papers, but I have been a joint author with local collaborators who have. It would help if we could have an easier way to communicate mathematics notation.

What are you working on these days?

Hui Xu, a Ph.D. student, is doing some work with me on density estimation when there are discontinuities. I have had a longstanding interest in that. I am also doing some work on random number generation with Brad Johnson, another Ph.D. student. And I just finished a paper on the Binomial n problem with Anirban DasGupta; it is coming out early next year in the Chernoff Festschrift!

What else do you enjoy? Have you done much traveling?

Well, I enjoy going to the concerts and the operas, although I did not so much as a child. I try to keep track of what is going on in mathematics. Previously, I could only do it by picking up the journals at the library. Now you can do some of it by using the net. I think what I enjoy the most is talking to students and people and be of any help that I can. You see, I have an open door policy.

For me, the most rewarding part of traveling is talking to people about interesting questions. I enjoyed going to the International Congress of Mathematicians at Stockholm, the Oberwolfach meeting I went to, and a meeting at Israel. I went to the ISI in 1974. Mahalanobis had just passed away, but there were a lot of people from everywhere at that meeting. I remember Persi Diaconis being there and Peter Bickel and many others. Urbanik was eating the raw jalapenos for his snacks. But even the symposium food was too spicy for me. From Calcutta, I went to Delhi. B. K. Kale invited me to come to Jaipur. It was an interesting trip. I went by a private limousine and returned on a public bus! Anirban wants to take me back there. We will see.

Have your many years of teaching influenced your ideas about statistical education?

Definitely. I believe that it is the unusual person who can go easily from the specific to the abstract. I think it is easier to go from the abstract to the specific. (Most of my colleagues disagree with me on this.) I have no objection to using examples after a concept. But going from special cases to the general still leaves the need for unlearning, which is difficult.

Because theorems and proofs are an important part of mathematical statistics, I believe that students who did not have some kind of course with theorems and proofs in high school, say Euclid-type geometry, flounder when they reach mathematics in college. We must improve quality of mathematics education in the US.

Competition is getting very strong and economic health is directly related to quality of education.

But even more important than experience with theorems and proofs are courses that emphasize concepts. For instance, thinking about integration as a limit of a sum is a crucial idea in statistics, especially for expectations. Students have difficulty if they learn integration as antidifferentiation, i.e. the "opposite of differentiation," and not as a summing process. I believe that it is possible for students to learn concepts directly if properly explained. This does not mean that a student will be able to use the concept upon hearing the words. Considerable learning may need to occur before the "light bulb" goes on.

To close the interview, what do you have to say about the future of statistics?

The biggest opportunities lie in the development of decision theoretic approaches to the problems of individual users where one considers ALL the consequences of the proposed solutions. Taking all the consequences into consideration can produce very difficult mathematical problems and provides great opportunities for those with mathematical expertise.

This is in contrast to the emphasis today on the development of general recipes that are used for solving problems and that are often used inappropriately. The latter two-thirds of the nineteenth century saw a similar emphasis. The turnaround came after World War II with people going into statistics from good mathematics programs who could attack the challenging mathematical problems. Before the turnaround there was also a rush by users, as now, to use statistical methods without understanding the assumptions and their consequences. I feel it is the user who must make the assumptions rather than just the statistician! Arguably, in a quantitative area, the user is not well prepared to do that.

Both those becoming statisticians and the users need to realize that there are underlying concepts for the field and they must use an understanding of the concepts rather than a catalog of methods. Just knowing how to compute does not help, and even being able only to prove lots of theorems would not. We CAN teach these concepts, and many of them even at fairly low level courses. The applied statistician needs to be able in many cases to invent new methods on the spot. There will be great opportunities for collaboration between applied scientists and mathematicians in the coming years. I hope neither ignores the other as an ancillary. That will be a mistake.

Thank you for your interesting views on research and teaching and for the interesting stories on your life. It was a pleasure. Good luck to you and we hope to continue to walk through the door and ask a question and get help. You have been a gracious resource to all of us. I wish you health and happiness.

You are very welcome.

References

[1] "On the distribution of the serial correlation coefficient," *Annals of Mathematical Statistics*, Vol. **16**, 1945, pp. 211–215.

[2] "The theory of decision procedures for distributions with monotone likelihood ratio," *Annals of Mathematical Statistics*, Vol. **27**, 1956, pp. 272–300. With S. Karlin.

[3] "Distributions possessing a monotone likelihood ratio," *Journal of the American Statistical Association*, Vol. **51**, 1956, pp. 637–643. With S. Karlin.

[4] "Measuring the equation systems of dynamic economics," *Cowles Commission Monograph*, No. **10**, 1950, pp. 53–237. With T. C. Koopmans and R. B., Leipnik.

[5] "Consistency of maximum-likelihood estimates in the explosive case," *Cowles Commission Monograph,* No. **10**, 1950, pp. 356–364.

[6] "Admissibility of quantile estimates of a single location parameter," *Annals of Mathematical Statistics*, Vol. **35**, 1964, pp. 1019–1030. With Martin Fox.

[7] *Equivalents of the Axiom of Choice II*, 1985, North Holland. With Jean E. Rubin.

[8] "How non-uniform can a uniform sample be?" *Probability Theory and Related Fields*, **73** (1986), 245–254. With Jeesen Chen and Burgess Davis.

[9] "On the distributions of sums of symmetric random variables and vectors." With Thomas Sellke. *Annals of Probability*, **14** (1986), 247–259.

[10] [10 Bose,A., DasGupta,A. and Rubin,H. (2002). A Contemporary Review of Infinitely Divisible Distributions and Processes, *Sankhȳa, Ser. A.* **64**, Part 3(ii), 763–819.

[11] "Estimation of the parameters of a single equation in a complete system of stochastic equation," *Annals of Mathematical Statistics*, Vol. **20**, 1949, pp. 46–63. With T. W. Anderson.

[12] "The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, Vol. **21**, 1950, pp. 370–382. With T.W. Anderson.

[13] "Statistical inference in factor analysis," *Proc. of the Third Berkeley Symp. on Math'l. Stat. and Prob'y.*, Vol. **5**, 1956, pp. 11–150. With T. W. Anderson.

[14] "Uniform convergence of random functions with applications to statistics," *Annals of Mathematical Statistics*, Vol. **27**, 1956, pp. 201–204.

[15] "Construction of random variables with specified distributions," *MSU Research Memorandum*, No. **88**, November 1961.

[16] "Some fast methods of generalizing random variables with preassigned distributions. I. General Acceptance–Rejection Procedures." Technical Report #466, Purdue University, Department of Statistics.

[17] "The effect of dead time on the physical generation of random digits." Technical Report #467, Purdue University, Department of Statistics.

[18] An Efficient Method of Generating Infinite-Precision Exponential Random Variables. Technical Report #86-39, Purdue University, Department of Statistics.

[19] Generating Non-Uniform Random Variables: Infinite Precision Procedures and Computational Complexity. Technical Report #86-58, Purdue University, Department of Statistics.

[20] "Probabilities of moderate deviations," *Sankhya A*, Vol. **27**, 1965, pp. 325–346. With J. Sethuraman.

[21] "Bayes risk efficiency," *Sankhya A*, Vol. **27**, 1965, pp. 347–356. With J. Sethuraman.

[22] "On a characterization of the Poisson distribution," *Sankhȳa A.* Vol. **26**, 1964, pp. 295–298. With C. R. Rao.

[23] Chen Z., Rubin H. and Vitale R. A. (1997). Independence and determination of probabilities. *Proc. Amer. Math. Soc.* **125**, No. 12, 3721–3723.

[24] "A characterization based on the absolute difference of two i.i.d. random variables," *Ann. Math. Stat.* **41** (1970) 2113–2122. With P. S. Puri.

[25] "On a characterization of the family of distributions with constant multivariate failure rates," *Annals of Probability.* **2** (1974) pp. 738–740. With P. S. Puri.

[26] "The Finite Memory Secretary Problem," Purdue University. With S. M. Samuels. *Annals of Prob.* **5**, 4 (1977) 627–635.

[27] Casella, G., DasGupta, A., Delampady, M., Rubin, H., and Strawderman, W. (2001). A new approach to default priors. *The Canadian Journal of Statistics.* **29**, No. 3, 437–450.

[28] DasGupta, A., and Rubin, H. (In Press). Estimation of the Binomial parameters when both parameters are unknown. *Special Issue of Journal of Statistical Planning and Inference in Honor of Herman Chernoff.*

[29] "Estimating the parameters of a differential process", *Annals of Mathematical Statistics*, Vol. **30**, 1959, pp. 641–658. With H. G. Tucker.

[30] "Generating functions and the semigroup theory of branching Markov process," *Proceedings of National Academy of Science*, Vol. **44**, 1958, pp. 1057–1060. With A. T. Bharucha-Reid.

[31] "The estimation of the location of a discontinuity in density," *Proceedings of the Third Berkeley Symp. of Math'l. Stat. & Prob'y.*, Vol. **1**, 1956, pp. 19–37. With H. Chernoff.

[32] Rubin, H. (1987). A weak system of axioms for "rational" behavior and the non–separability of utility from prior. *Statistics and Decisions* pp. 47–58.