Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES

Optimality

The Second Erich L. Lehmann Symposium

Javier Rojo, Editor

Volume 49

Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES

Volume 49

Optimality

The Second Erich L. Lehmann Symposium

Javier Rojo, Editor

Institute of Mathematical Statistics
Beachwood, Ohio, USA

Institute of Mathematical Statistics
Lecture Notes–Monograph Series

Series Editor:
Richard A. Vitale

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

Library of Congress Control Number: 2006929652

International Standard Book Number 0-940600-66-9

International Standard Serial Number 0749-2170

Copyright © 2006 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

Contents

Preface: Brief history of the Lehmann Symposia: Origins, goals and motivation
Javier Rojo v

Contributors to this volume
. vii

Scientific program
. viii

Partial list of participants
. xii

Acknowledgement of referees' services
. xix

PAPERS

Testing

On likelihood ratio tests
Erich L. Lehmann 1

Student's t-test for scale mixture errors
Gábor J. Székely 9

Multiple Testing

Recent developments towards optimality in multiple hypothesis testing
Juliet Popper Shaffer 16

On stepdown control of the false discovery proportion
Joseph P. Romano and Azeem M. Shaikh 33

An adaptive significance threshold criterion for massive multiple hypotheses testing
Cheng Cheng 51

Philosophy

Frequentist statistics as a theory of inductive inference
Deborah G. Mayo and D. R. Cox 77

Where do statistical models come from? Revisiting the problem of specification
Aris Spanos 98

Transformation Models, Proportional Hazards

Modeling inequality and spread in multiple regression
Rolf Aaberge, Steinar Bjerve and Kjell Doksum 120

Estimation in a class of semiparametric transformation models
Dorota M. Dabrowska 131

Bayesian transformation hazard models
Gousheng Yin and Joseph G. Ibrahim 170

Copulas and Decoupling

Characterizations of joint distributions, copulas, information, dependence and decoupling, with applications to time series
Victor H. de la Peña, Rustam Ibragimov and Shaturgun Sharakhmetov 183

Regression Trees

Regression tree models for designed experiments
Wei-Yin Loh 210

Competing Risks

On competing risk and degradation processes
Nozer D. Singpurwalla 229

Restricted estimation of the cumulative incidence functions corresponding to competing risks
Hammou El Barmi and Hari Mukerjee 241

Robustness

Comparison of robust tests for genetic association using case-control studies
Gang Zheng, Boris Freidlin and Joseph L. Gastwirth 253

Multiscale Stochastic Processes

Optimal sampling strategies for multiscale stochastic processes
Vinay J. Ribeiro, Rudolf H. Riedi and Richard G. Baraniuk 266

Asymptotics

The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations
Hannes Leeb 291

Local asymptotic minimax risk bounds in a locally asymptotically mixture of normal experiments under asymmetric loss
Debasis Bhattacharya and A. K. Basu 312

Density Estimation

On moment-density estimation in some biased models
Robert M. Mnatsakanov and Frits H. Ruymgaart 322

A note on the asymptotic distribution of the minimum density power divergence estimator
Sergio F. Juárez and William R. Schucany 334

Brief history of the Lehmann Symposia: Origins, goals and motivation

The idea of the Lehmann Symposia as platforms to encourage a revival of interest in fundamental questions in theoretical statistics, while keeping in focus issues that arise in contemporary interdisciplinary cutting-edge scientific problems, developed during a conversation that I had with Victor Perez Abreu during one of my visits to Centro de Investigación en Matemáticas (CIMAT) in Guanajuato, Mexico. Our goal was and has been to showcase relevant theoretical work to encourage young researchers and students to engage in such work.

The First Lehmann Symposium on Optimality took place in May of 2002 at Centro de Investigación en Matemáticas in Guanajuato, Mexico. A brief account of the Symposium has appeared in Vol. 44 of the Institute of Mathematical Statistics series of Lecture Notes and Monographs. The volume also contains several works presented during the First Lehmann Symposium. All papers were refereed. The program and a picture of the participants can be found on-line at the website `http://www.stat.rice.edu/lehmann/lst-Lehmann.html`.

The Second Lehmann Symposium on Optimality was held from May 19–May 22, 2004 at Rice University. There were close to 175 participants in the Symposium. A partial list and a photograph of participants, as well as the details of the scientific program, are provided in the next few pages. All scientific activities took place in Duncan Hall in the School of Engineering. Most of the plenary and invited speakers agreed to be videotaped and their talks may be accessed by visiting the following website: `http://webcast.rice.edu/webcast.php?action=details&event=408`. All papers presented in this volume were refereed, and one third of submitted papers were rejected.

At the time of this writing, plans are underway to hold the Third Lehmann Symposium at the Mathematical Sciences Research Institute during May of 2007.

I want to acknowledge the help from members of the Scientific Program Committee: Jane-Ling Wang (UC Davis), David W. Scott (Rice University), Juliet P. Shaffer (UC Berkeley), Deborah Mayo (Virginia Polytechnic Institute), Jef Teugels (Katholieke Universiteit Leuven), James R. Thompson (Rice University), and Javier Rojo (Chair).

The Symposia could not take place without generous financial support from various institutions. The First Symposium was financed in its entirety by CIMAT under the direction of Victor Perez Abreu. The Second Lehmann Symposium was generously funded by The National Science Foundation, Pfizer, Inc., The University of Texas MD Anderson Cancer Center, CIMAT, and Cytel. Shulamith Gross at NSF, Demissie Alemayehu at Pfizer, Gary Rosner at MD Anderson Cancer Center, Victor Perez Abreu at CIMAT, and Cyrrus Mehta at Cytel, encouraged and facilitated the process to obtain the support. The Rice University School of Engineering's wonderful physical facilities were made available for the Symposium at no charge.

Finally, thanks go the Statistics Department at Rice University for facilitating my participation in these activities.

May 15th, 2006
Javier Rojo
Rice University
Editor

Contributors to this volume

Aaberge, R., *Statistics Norway*

Baraniuk, R. G., *Rice University*
Basu, A. K., *Calcutta University*
Bhattacharya, D., *Visva-Bharati University*
Bjerve, S., *University of Oslo*

Cheng, C., *St. Jude Children's Research Hospital*
Cox, D. R., *Nuffield College, Oxford*

Dabrowska, D. M., *University of California, Los Angeles*
de la Peña, V. H., *Columbia University*
Doksum, K., *University of Wisconsin, Madison*

El Barmi, H., *Baruch College, City University of New York*

Freidlin, B., *National Cancer Institute*

Gastwirth, J. L., *The George Washington University*

Ibragimov, R., *Harvard University*
Ibrahim, J., *University of North Carolina*

Juárez, S., *Veracruzana University*

Leeb, H., *Yale University*
Lehmann, E. L., *University of California, Berkeley*
Loh, W.-Y., *University of Wisconsin, Madison*

Mayo, D. G., *Virginia Polytechnical Institute*
Mnatsakanov, R. M., *West Virginia University*
Mukerjee, H., *Wichita State University*

Ribeiro, V. J., *Rice University*
Riedi, R. H., *Rice University*
Romano, J. P., *Stanford University*
Ruymgaart, F. H., *Texas Tech University*

Schucany, W. R., *Southern Methodist University*
Shaffer, J. P., *University of California*
Shaikh, A. M., *Stanford University*
Sharakhmetov, S., *Tashkent State Economics University*
Singpurwalla, N. D., *The George Washington University*
Spanos, A., *Virginia Polytechnical Institute and State University*
Székely, G. J., *Bowling Green State University, Hungarian Academy of Sciences*

Yin, G., *MD Anderson Cancer Center*

Zheng, G., *National Heart, Lung and Blood Institute*

SCIENTIFIC PROGRAM

The Second Erich L. Lehmann Symposium
May 19–22, 2004
Rice University

Symposium Chair and Organizer Javier Rojo

Statistics Department, MS-138
Rice University
6100 Main Street
Houston, TX 77005

Co-Chair Victor Perez-Abreu

Probability and Statistics
CIMAT
Callejon Jalisco S/N
Guanajuato, Mexico

Plenary Speakers

Erich L. Lehmann UC Berkeley	*Conflicting principles in hypothesis testing*
Peter Bickel UC Berkeley	*From rank tests to semiparametrics*
Ingram Olkin Stanford University	*Probability models for survival and reliability analysis*
D. R. Cox Nuffield College Oxford	*Graphical Markov models: A tool for interpretation*
Emanuel Parzen Texas A&M University	*Data modeling, quantile/quartile functions, confidence intervals, introductory statistics reform*
Bradley Efron Stanford University	*Confidence regions and inferences for a multivariate normal mean vector*
Kjell Doksum UC Berkeley and UW Madison	*Modeling money*
Persi Diaconis Stanford University	*In praise of statistical theory*

Invited Sessions

New Investigators

Javier Rojo,	Organizer
William C. Wojciechowski,	Chair
Gabriel Huerta U of New Mexico	*Spatio-temporal analysis of Mexico city ozone levels*
Sergio Juarez U Veracruzana Mexico	*Robust and efficient estimation for the generalized Pareto distribution*
William C. Wojciechowski Rice University	*Adaptive robust estimation by simulation*
Rudolf H. Riedi Rice University	*Optimal sampling strategies for tree-based time series*

Multiple hypothesis tests: New approaches—optimality issues

Juliet P. Shaffer,	Chair
Juliet P. Shaffer UC Berkeley	*Different types of optimality in multiple testing*
Joseph Romano Stanford University	*Optimality in stepwise hypothesis testing*
Peter Westfall Texas Tech University	*Optimality considerations in testing massive numbers of hypotheses*

Robustness

James R. Thompson,	Chair
Adrian Raftery U of Washington	*Probabilistic weather forecasting using Bayesian model averaging*
James R. Thompson Rice University	*The simugram: A robust measure of market risk*
Nozer D. Singpurwalla George Washington U	*The hazard potential: An approach for specifying models of survival*

Extremes and Finance

Jef Teugels,	Chair
Richard A. Davis Colorado State University	*Regular variation and financial time series models*
Hansjoerg Albrecher University of Graz Austria	*Ruin theory in the presence of dependent claims*
Patrick L. Brockett U of Texas, Austin	*A chance constrained programming approach to pension plan management when asset returns are heavy tailed*

Recent Advances in Longitudinal Data Analysis

Naisyin Wang, Chair

Raymond J. Carroll, Texas A&M Univ. — *Semiparametric efficiency in longitudinal marginal models*

Pushing Hsieh, UC Davis — *Some issues and results on nonparametric maximum likelihood estimation in a joint model for survival and longitudinal data*

Jane-Ling Wang, UC Davis — *Functional regression and principal components analysis for sparse longitudinal data*

Semiparametric and Nonparametric Testing

David W. Scott, Chair

Jeffrey D. Hart, Texas A&M Univ. — *Semiparametric Bayesian and frequentist tests of trend for a large collection of variable stars*

Joseph Gastwirth, George Washington U. — *Efficiency robust tests for linkage or association*

Irene Gijbels, U Catholique de Louvain — *Nonparametric testing for monotonicity of a hazard rate*

Philosophy of Statistics

Persi Diaconis, Chair

David Freedman, UC Berkeley — *Some reflections on the foundations of statistics*

Sir David Cox, Nuffield College, Oxford — *Some remarks on statistical inference*

Deborah Mayo, Virginia Tech — *The theory of statistics as the "frequentist's" theory of inductive inference*

Special contributed session

Shulamith T. Gross, Chair

Victor Hugo de la Pena, Columbia University — *Pseudo maximization and self-normalized processes*

Wei-Yin Loh, U of Wisconsin, Madison — *Regression tree models for data from designed experiments*

Shulamith T. Gross, NSF and Baruch College/CUNY — *Optimizing your chances of being funded by the NSF*

Contributed papers

Aris Spanos, Virginia Tech: *Where do statistical models come from? Revisiting the problem of specification*

Hannes Leeb, Yale University: *The large-sample minimal coverage probability of confidence intervals in regression after model selection*

Jun Yan, University of Iowa: *Parametric inference of recurrent alternating event data*

Gâbor J. Székely, Bowling Green State U and Hungarian Academy of Sciences: *Student's t-test for scale mixture errors*

Jaechoul Lee, Boise State University: *Periodic time series models for United States extreme temperature trends*

Loki Natarajan, University of California, San Diego: *Estimation of spontaneous mutation rates*

Chris Ding, Lawrence Berkeley Laboratory: *Scaled principal components and correspondence analysis: clustering and ordering*

Mark D. Rothmann, Biologies Therapeutic Statistical Staff, CDER, FDA: *Inferences about a life distribution by sampling from the ages and from the obituaries*

Victor de Oliveira, University of Arkansas: *Bayesian inference and prediction of Gaussian random fields based on censored data*

Jose Aimer T. Sanqui, Appalachian State University: *The skew-normal approximation to the binomial distribution*

Guosheng Yin, The University of Texas MD Anderson Cancer Center: *A class of Bayesian shared gamma frailty models with multivariate failure time data*

Eun-Joo Lee, Texas Tech University: *An application of the Hâjek–Le Cam convolution theorem*

Daren B. H. Cline, Texas A&M University: *Determining the parameter space, Lyapounov exponents and existence of moments for threshold ARCH and GARCH time series*

Hammou El Barmi, Baruch College: *Restricted estimation of the cumulative incidence functions corresponding to K competing risks*

Asheber Abebe, Auburn University: *Generalized signed-rank estimation for nonlinear models*

Yichuan Zhao, Georgia State University: *Inference for mean residual life and proportional mean residual life model via empirical likelihood*

Cheng Cheng, St. Jude Children's Research Hospital: *A significance threshold criterion for large-scale multiple tests*

Yuan-Ji, The University of Texas MD Anderson Cancer Center: *Bayesian mixture models for complex high-dimensional count data*

K. Krishnamoorthy, University of Louisiana at Lafayette: *Inferences based on generalized variable approach*

Vladislav Karguine, Cornerstone Research: *On the Chernoff bound for efficiency of quantum hypothesis testing*

Robert Mnatsakanov, West Virginia University: *Asymptotic properties of moment-density and moment-type CDF estimators in the models with weighted observations*

Bernard Omolo, Texas Tech University: *An aligned rank test for a repeated observations model with orthonormal design*

The Second Lehmann Symposium—Optimality
Rice University, May 19–22, 2004

Partial List of Participants

Asheber Abebe
Auburn University
abebeas@auburn.edu

Hansjoerg Albrecher
Graz University
albrecher@tugraz.at

Demissie Alemayehu
Pfizer
alem@stat.columbia.edu

E. Neely Atkinson
University of Texas
MD Anderson Cancer Center
eatkinso@mdanderson.org

Scott Baggett
Rice University
baggett@rice.edu

Sarah Baraniuk
University of Texas Houston
School of Public Health
sbaraniuk@sph.uth.tmc.edu

Jose Luis Batun
CIMAT
batun@cimat.mx

Debasis Bhattacharya
Visva-Bharati, India
Debases_us@yahoo.com

Chad Bhatti
Rice University
bhatticr@rice.edu

Peter Bickel
University of California, Berkeley
bickel@stat.berkeley.edu

Sharad Borle
Rice University
sborle@rice.edu

Patrick Brockett
University of Texas, Austin
brockett@mail.utexas.edu

Barry Brown
University of Texas
MD Anderson Cancer Center
bwb@mdanderson.org

Ferry Butar Butar
Sam Houston State
University
mth_fbb@shsu.edu

Raymond Carroll
Texas A&M University
carroll@stat.tamu.edu

Wenyaw Chan
University of Texas, Houston
Health Science Center
Wenyaw.Chan@uth.tmc.edu

Jamie Chatman
Rice University
jchatman@rice.edu

Cheng Cheng
St Jude Hospital
cheng.cheng@stjude.org

Hyemi Choi
Seoul National University
hyemichoi@yahoo.com

Blair Christian
Rice University
blairc@rice.edu

Daren B. H. Cline
Texas A&M University
dcline@stat.tamu.edu

Daniel Covarrubias
Rice University
dcorvarru@stat.rice.edu

David R. Cox
Nuffield College, Oxford
david.cox@nut.ox.ac.uk

Dennis Cox
Rice University
dcox@rice.edu

Kalatu Davies
Rice University
kdavies@rice.edu

Ginger Davis
Rice University
gmdavis@rice.edu

Richard Davis
Colorado State University
rdavis@stat.colostate.edu

Victor H. de la Peña
Columbia University
vp@stat.columbia.edu

Li Deng
Rice University
lident@rice.edu

Victor De Oliveira
University of Arkansas
vdo@uark.ed

Persi Diaconis
Stanford University

Chris Ding
Lawrence Berkeley Natl Lab
chqding@lbl.gov

Kjell Doksum
University of Wisconsin
doksum@stat.wisc.edu

Joan Dong
University of Texas
MD Anderson Cancer Center
qdong@mdanderson.org

Wesley Eddings
Kenyon College
eddingsw@kenyon.edu

Brad Efron
Stanford University
brad@statistics.stanford.edu

Hammou El Barmi
Baruch College
hammou_elbarmi@baruch.cuny.edu

Kathy Ensor
Rice University
kathy@rice.edu

Alan H. Feiveson
Johnson Space Center
alan.h.feiveson@nasa.gov

Hector Flores
Rice University
hflores@rice.edu

Garrett Fox
Rice University
gfox@stat.rice.edu

David A. Freedman
University of California, Berkeley
freedman@stat.berkeley.edu

Wenjiang Fu
Texas A&M University
wfu@stat.tamu.edu

Joseph Gastwirth
George Washington University
jlgast@gwu.edu

Susan Geller
Texas A&M University
geller@math.tamu.edu

Musie Ghebremichael
Rice University
musie@rice.edu

Irene Gijbels
Catholic University of Louvin
gijbels@stat.ucl.ac.be

Nancy Glenn
University of South Carolina
nglenn@stat.sc.edu

Carlos Gonzalez Universidad
Veracruzana
cglezand@tema.cum.mx

Shulamith Gross
NSF
sgross@nsf.gov

Xiangjun Gu
University of Texas
MD Anderson Cancer Center
xgu@mdanderson.org

Rudy Guerra
Rice University
rguerra@rice.edu

Shu Han
Rice University
shuhan@rice.edu

Robert Hardy
University of Texas
Health Science Center, Houston SPH
bhardy@sph.uth.tmc.edu

Jeffrey D. Hart
Texas A&M University
hart@stat.tamu.edu

Mike Hernandez
University of Texas
MD Anderson Cancer Center
Mike@sph.uth.tmc.edu

Richard Heydorn
NASA
richard.p.heydorn@nasa.gov

Tyson Holmes
Stanford University
tholmes@stanford.edu

Charlotte Hsieh
Rice University
hsiehc@rice.edu

Pushing Hsieh
University of California, Davis
fushing@wald.ucdavis.edu

Xuelin Huang
University of Texas
MD Anderson Cancer Center
xlhuang@mdanderson.org

Gabriel Huerta
University of New Mexico
ghuerta@stat.unm.edu

Sigfrido Iglesias
Gonzalez University of Toronto
sigfrido@fisher.utstat.toronto.c

Yuan Ji
University of Texas
yuanji@mdanderson.org

Sergio Juarez
Veracruz University
Mexico
sejuarez@uv.mx

Asha Seth Kapadia
University of Texas
Health Science Center, Houston SPH
School of Public Health
akapadia@sph.uth.tmc.edu

Vladislav Karguine
Cornerstone Research
slava@bu.edu

K. Krishnamoorthy
University of Louisiana
krishna@louisiana.edu

Mike Lecocke
Rice University
mlecocke@stat.rice.edu

Eun-Joo Lee
Texas Tech University
elee@math.ttu.edu

J. Jack Lee
University of Texas
MD Anderson Cancer Center
jjlee@mdanderson.org

Jaechoul Lee
Boise State University
jaechlee@math.biostate.edu

Jong Soo Lee
Rice University
jslee@rice.edu

Young Kyung Lee
Seoul National University
itsgirl@hanmail.net

Hannes Leeb
Yale University
hannes.leeb@yale.edu

Erich Lehmann
University of California,
Berkeley
shaffer@stat.berkeley.edu

Lei Lei
University of Texas
Health Science Center, SPH
llei@sph.uth.tmc.edu

Wei-Yin Loh
University of Wisconsin
loh@stat.wisc.edu

Yen-Peng Li
University of Texas, Houston
School of Public Health
yli@sph.uth.tmc.edu

Yisheng Li
University of Texas
MD Anderson Cancer Center
ysli@mdanderson.org

Simon Lunagomez
University of Texas
MD Anderson Cancer Center
slunago@mdanderson.org

Matthias Matheas
Rice University
matze@rice.edu

Deborah Mayo
Virginia Tech
mayod@vt.edu

Robert Mnatsakanov
West Virginia University
rmnatsak@stat.wvu.edu

Jeffrey Morris
University of Texas
MD Anderson Cancer Center
jeffmo@odin.mdacc.tmc.edu

Peter Mueller
University of Texas
MD Anderson Cancer Center
pmueller@mdanderson.org

Bin Nan
University of Michigan
bnan@umich.edu

Loki Natarajan
University of California,
San Diego
loki@math.ucsd.edu

E. Shannon Neeley
Rice University
sneeley@rice.edu

Josue Noyola-Martinez
Rice University
jcnm@rice.edu

Ingram Olkin
Stanford University
iolkin@stat.stanford.edu

Peter Olofsson
Rice University

Bernard Omolo
Texas Tech University
bomolo@math.ttu.edu

Richard C. Ott
Rice University
rott@rice.edu

Galen Papkov
Rice University
gpapkov@rice.edu

Byeong U Park
Seoul National University
bupark@stats.snu.ac.kr

Emanuel Parzen
Texas A&M University
eparzen@tamu.edu

Bo Peng
Rice University
bpeng@rice.edu

Kenneth Pietz
Department of Veteran Affairs
kpietz@bcm.tmc.edu

Kathy Prewitt
Arizona State University
kathryn.prewitt@asu.edu

Adrian Raftery
University of Washington
raftery@stat.washington.edu

Vinay Ribeiro
Rice University
vinay@rice.edu

Peter Richardson
Baylor College of Medicine
peterr@bcm.tmc.edu

Rolf Riedi
Rice University
riedi@rice.edu

Javier Rojo
Rice University
jrojo@rice.edu

Joseph Romano
Stanford University
romano@stat.stanford.edu

Gary L. Rosner
University of Texas
MD Anderson Cancer Center
glrosner@mdanderson.org

Mark Rothmann
US Food and Drug
Administration
rothmann@cder.fda.gov

Chris Rudnicki
Rice University
rudnicki@stat.rice.edu

Jose Aimer Sanqui
Appalachian St. University
sanquijat@appstate.edu

William R. Schucany
Southern Methodist University
schucany@smu.edu

Alena Scott
Rice University
oetting@stat.rice.edu

David W. Scott
Rice University
scottdw@rice.edu

Juliet Shaffer
University of California, Berkeley
shaffer@stat.berkeley.edu

Yu Shen
University of Texas
MD Anderson Cancer Center
yushen@mdanderson.org

Nozer Singpurwalla
The George Washington University
nozer@gwu.edu

Tumulesh Solanky
University of New Orleans
tsolanky@uno.edu

Julianne Souchek
Department of Veteran Affairs
jsoucheck@bcm.tmc.edu

Melissa Spann
Baylor University
melissa-spann@baylor.edu

Aris Spanos
Virginia Tech
aris@vt.edu

Hsiguang Sung
Rice University
hgsung@rice.edu

Gábor Székely
Bowling Green Sate
University
gabors@bgnet.bgsu.edu

Jef Teugels
Katholieke Univ. Leuven
jef.teugels@wis.kuleuven.be

James Thompson
Rice University
thomp@rice.edu

Jack Tubbs
Baylor University
jack_tubbs@baylor.edu

Jane-Ling Wang
University of California, Davis
wang@wald.ucdavis.edu

Naisyin Wang
Texas A&M University
nwang@stat.tamu.edu

Kyle Wathen
University of Texas
MD Anderson Cancer Center
& University of Texas GSBS
jkwathen@mdanderson.org

Peter Westfall
Texas Tech University
westfall@ba.ttu.edu

William Wojciechowski
Rice University
williamc@rice.edu

Jose-Miguel Yamal
Rice University &
University of Texas
MD Anderson Cancer Center
jmy@stat.rice.edu

Jun Yan
University of Iowa
jyan@stat.wiowa.edu

Guosheng Yin
University of Texas
MD Anderson Cancer Center
gyin@odin.mdacc.tmc.edu

Zhaoxia Yu
Rice University
yu@rice.edu

Issa Zakeri
Baylor College of Medicine
izakeri@bcm.tmc.edu

Qing Zhang
University of Texas
MD Anderson Cancer Center
qzhang@mdanderson.org

Hui Zhao
University of Texas
Health Science Center
School of Public Health
hzhao@sph.uth.tmc.edu

Yichum Zhao
Georgia State University
yzhao@math.stat.gsu.edu

Acknowledgement of referees' services

The efforts of the following referees are gratefully acknowledged

Jose Luis Batun
CIMAT, Mexico

Roger Berger
Arizona State University

Prabir Burman
University of California, Davis

Ray Carroll
Texas A&M University

Cheng Cheng
St. Jude's Children's Research Hospital

David R. Cox
Nuffield College, Oxford

Dorota M. Dabrowska
University of California, Los Angeles

Victor H. de la Pena
Columbia University

Kjell Doksum
University of Wisconsin, Madison

Armando Dominguez
CIMAT, Mexico

Sandrine Dudoit
University of California, Berkeley

Richard Dykstra
University of Iowa

Bradley Efron
Stanford University

Harnmou El Barmi
The City University of New York

Luis Enrique Figueroa
Purdue University

Joseph L. Gastwirth
George Washington University

Marc G. Genton
Texas A&M University

Musie Ghebremichael
Yale University

Graciela Gonzalez
Mexico

Hannes Leeb
Yale University

Erich L. Lehmann
University of California, Berkeley

Ker-Chau Li
University of California, Los Angeles

Wei-Yin Loh
University of Wisconsin, Madison

Hari Mukerjee
Wichita State University

Loki Natarajan
University of California, San Diego

Ingram Olkin
Stanford University

Liang Peng
Georgia Institute of Technology

Joseph P. Romano
Stanford University

Louise Ryan
Harvard University

Sanat Sarkar
Temple University

William R. Schucany
Southern Methodist University

David W. Scott
Rice University

Juliet P. Shaffer
University of California, Berkeley

Nozer D. Singpurwalla
George Washington University

David Sprott
CIMAT and University of Waterloo

Jef L. Teugels
Katholieke Universiteit Leuven

Martin J. Wainwright
University of California, Berkeley

Jane-Ling Wang
University of California, Davis

Peter Westfall
Texas Tech University

Grace Yang
University of Maryland, College Park

Yannis Yatracos (2)
National University of Singapore

Guosheng Yin
University of Texas MD Anderson Cancer Center

Hongyu Zhao
Yale University

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 1–8
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000356

On likelihood ratio tests

Erich L. Lehmann[1]

University of California at Berkeley

Abstract: Likelihood ratio tests are intuitively appealing. Nevertheless, a number of examples are known in which they perform very poorly. The present paper discusses a large class of situations in which this is the case, and analyzes just how intuition misleads us; it also presents an alternative approach which in these situations is optimal.

1. The popularity of likelihood ratio tests

Faced with a new testing problem, the most common approach is the likelihood ratio (LR) test. Introduced by Neyman and Pearson in 1928, it compares the maximum likelihood under the alternatives with that under the hypothesis. It owes its popularity to a number of facts.

(i) It is intuitively appealing. The likelihood of θ,

$$L_x(\theta) = p_\theta(x)$$

i.e. the probability density (or probability) of x considered as a function of θ, is widely considered a (relative) measure of support that the observation x gives to the parameter θ. (See for example Royall [8]). Then the likelihood ratio

$$\sup_{alt}[p_\theta(x)] / \sup_{hyp}[p_\theta(x)] \tag{1.1}$$

compares the best explanation the data provide for the alternatives with the best explanations for the hypothesis. This seems quite persuasive.

(ii) In many standard problems, the LR test agrees with tests obtained from other principles (for example it is UMP unbiased or UMP invariant). Generally it seems to lead to satisfactory tests. However, counter-examples are also known in which the test is quite unsatisfactory; see for example Perlman and Wu [7] and Menéndez, Rueda, and Salvador [6].

(iii) The LR test, under suitable conditons, has good asymptotic properties.

None of these three reasons are convincing.

(iii) tells us little about small samples.

(i) has no strong logical grounding.

(ii) is the most persuasive, but in these standard problems (in which there typically exist a complete set of sufficient statistics) all principles typically lead to tests that are the same or differ only by little.

[1]Department of Statistics, 367 Evans Hall, University of California, Berkeley, CA 94720-3860, e-mail: shaffer@stat.berkeley.edu

AMS 2000 subject classifications: 62F03.

Keywords and phrases: likelihood ratio tests, average likelihood, invariance.

In view of lacking theoretical support and many counterexamples, it would be good to investigate LR tests systematically for small samples, a suggestion also made by Perlman and Wu [7]. The present paper attempts a first small step in this endeavor.

2. The case of two alternatives

The simplest testing situation is that of testing a simple hypothesis against a simple alternative. Here the Neyman-Pearson Lemma completely vindicates the LR-test, which always provides the most powerful test. Note however that in this case no maximization is involved in either the numerator or denominator of (1.1), and as we shall see, it is just these maximizations that are questionable.

The next simple situation is that of a simple hypothesis and two alternatives, and this is the case we shall now consider.

Let $\underline{X}=(X_1,\ldots,X_n)$ where the X's are iid. Without loss of generality suppose that under H the X's are uniformly distributed on $(0,1)$. Consider two alternatives f,g on $(0,1)$. To simplify further, we shall assume that the alternatives are symmetric, i.e. that

$$\begin{aligned} p_1(\underline{x}) &= f(x_1)\cdots f(x_n) \\ p_2(\underline{x}) &= f(1-x_1)\cdots f(1-x_n). \end{aligned} \tag{2.1}$$

Then it is natural to restrict attention to symmetric tests (that is the invariance principle) i.e. to rejection regions R satisfying

$$(x_1,\ldots,x_n)\in R \quad \text{if and only if } (1-x_1,\ldots,1-x_n)\in R. \tag{2.2}$$

The following result shows that under these assumptions there exists a uniformly most powerful (UMP) invariant test, i.e. a test that among all invariant tests maximizes the power against both $p_1$and p_2.

Theorem 2.1. *For testing H against the alternatives* (2.1) *there exists among all level α rejection regions R satisfying* (2.2) *one that maximizes the power against both p_1 and p_2 and it rejects H when*

$$\frac{1}{2}[p_1(\underline{x})+p_2(\underline{x})] \quad \text{is sufficiently large.} \tag{2.3}$$

We shall call the test (2.3) the *average likelihood ratio test* and from now on shall refer to (1.1) as the *maximum likelihood ratio test.*

Proof. If R satisfies (2.2), its power against $p_1$and p_2 must be the same. Hence

$$\int_R p_1 \;=\; \int_R p_2 = \int_R \frac{1}{2}(p_1+p_2). \tag{2.4}$$

□

By the Neyman–Pearson Lemma, the most powerful test of H against $\frac{1}{2}[p_1+p_2]$ rejects when (2.3) holds.

Corollary 2.1. *Under the assumptions of Theorem* 2.1*, the average LR test has power greater than or equal to that of the maximum likelihood ratio test against both p_1 and p_2.*

Proof. The maximum LR test rejects when

$$\max(p_1(\underline{x}),\, p_2(\underline{x})) \quad \text{is sufficiently large.} \tag{2.5}$$

Since this test satisfies (2.2), the result follows. □

The Corollary leaves open the possibility that the average and maximum LR tests have the same power; in particular they may coincide. To explore this possibility consider the case $n = 1$ and suppose that f is increasing. Then the likelihood ratio will be

$$f(x) \quad \text{if } x > \tfrac{1}{2} \quad \text{and } f(1-x) \text{ if } x < \frac{1}{2}. \tag{2.6}$$

The maximum LR test will therefore reject when

$$\left|x - \frac{1}{2}\right| \quad \text{is sufficiently large} \tag{2.7}$$

i.e. when x is close to either 0 or 1.

It turns out that the average LR test will depend on the shape of f and we shall consider two cases: (a) f is convex; (b) f is concave.

Theorem 2.2. *Under the assumptions of Theorem* 2.1 *and with* $n = 1$,

(i) (a) *if f is convex, the average LR test rejects when* (2.7) *holds;*
(b) *if f is concave, the average LR test rejects when*

$$\left|x - \frac{1}{2}\right| \quad \textit{is sufficiently small.} \tag{2.8}$$

(ii) (a) *if f is convex, the maximum LR test coincides with the average LR test, and hence is UMP among all tests satisfying* (2.2) *for n=1.*
(b) *if f is concave, the maximum LR test uniformly* minimizes *the power among all tests satisfying* (2.2) *for n=1, and therefore has power* $< \alpha$.

Proof. This is an immediate consequence of the fact that if $x < x' < y' < y$ then

$$\frac{f(x)+f(y)}{2} \text{ is } \begin{matrix}\geq \\ \leq\end{matrix} \frac{f(x')+f(y')}{2} \quad \text{if } f \text{ is } \begin{matrix}\text{convex.} \\ \text{concave.}\end{matrix} \tag{2.9}$$

It is clear from the argument that the superiority of the average over the likelihood ratio test in the concave case will hold even if p_1 and p_2 are not exactly symmetric. Furthermore it also holds if the two alternatives p_1 and p_2 are replaced by the family $\theta p_1 + (1-\theta)p_2$, $0 \leq \theta \leq 1$. □

3. A finite number of alternatives

The comparison of maximum and average likelihood ratio tests discussed in Section 2 for the case of two alternatives obtains much more generally. In the present section we shall sketch the corresponding result for the case of a simple hypothesis against a finite number of alternatives which exhibit a symmetry generalizing (2.1).

Suppose the densities of the simple hypothesis and the s alternatives are denoted by $p_0, p_1, \ldots, p_s$ and that there exists a group G of transformations of the sample which leaves invariant both p_0 and the set $\{p_1, \ldots, p_s\}$ (i.e. each transformation

results in a permutation of $p_1, \ldots, p_s$). Let $\bar{G}$ denote the set of these permutations and suppose that it is transitive over the set $\{p_1, \ldots, p_s\}$ i.e. that given any i and j there exists a transformation in $\bar{G}$ taking p_i into p_j. A rejection region R is said to be invariant under $\bar{G}$ if

$$x \in R \quad \text{if and only if} \quad g(x) \in R \text{ for all } g \text{ in } G. \tag{3.1}$$

Theorem 3.1. *Under these assumptions there exists a uniformly most powerful invariant test and it rejects when*

$$\frac{\sum_{i=1}^{s} p_i(x)/s}{p_0(x)} \quad \text{is sufficiently large.} \tag{3.2}$$

In generalization of the terminology of Theorem 2.1 we shall call (3.2) the average likelihood ratio test. The proof of Theorem 3.1 exactly parallels that of Theorem 2.1.

The Theorem extends to the case where G is a compact group. The average in the numerator of (3.2) is then replaced by the integral with respect to the (unique) invariant probability measure over $\bar{G}$. For details see Eaton ([3], Chapter 4). A further extension is to the case where not only the alternatives but also the hypothesis is composite.

To illustrate Theorem 3.1, let us extend the case considered in Section 2. Let (X, Y) have a bivariate distribution over the unit square which is uniform under H. Let f be a density for (X, Y) which is strictly increasing in both variables and consider the four alternatives

$$p_1 = f(x, y),\ p_2 = f(1-x, y),\ p_3 = f(x, 1-y),\ p_4 = f(1-x, 1-y).$$

The group G consists of the four transformations

$$g_1(x, y) = (x, y),\ g_2(x, y) = (1-x, y),\ g_3(x, y) = (x, 1-y),$$
$$\text{and } g_4(x, y) = (1-x, 1-y).$$

They induce in the space of $(p_1, \ldots, p_4)$ the transformations:

$\bar{g}_1 =$ the identity
$\bar{g}_2$: $p_1 \to p_2$, $p_2 \to p_1$, $p_3 \to p_4$, $p_4 \to p_3$
$\bar{g}_3$: $p_1 \to p_3$, $p_3 \to p_1$, $p_2 \to p_4$, $p_4 \to p_2$
$\bar{g}_4$: $p_1 \to p_4$, $p_4 \to p_1$, $p_2 \to p_3$, $p_3 \to p_2$.

This is clearly transitive, so that Theorem 3.1 applies. The uniformly most powerful invariant test, which rejects when

$$\sum_{i=1}^{4} p_i(x, y) \quad \text{is large}$$

is therefore uniformly at least as powerful as the maximum likelihood ratio test which rejects when

$$\max\left[p_1(x, y), p_2(x, y), p_3(x, y), p_4(x, y)\right]$$

is large.

4. Location-scale families

In the present section we shall consider some more classical problems in which the symmetries are represented by infinite groups which are not compact. As a simple example let the hypothesis H and the alternatives K be specified respectively by

$$H: \ f(x_1-\theta,\ldots,x_n-\theta) \text{ and } \quad K: \ g(x_1-\theta,\ldots,x_n-\theta) \tag{4.1}$$

where f and g are given densities and θ is an unknown location parameter. We might for example want to test a normal distribution with unknown mean against a logistic or Cauchy distribution with unknown center.

The symmetry in this problem is characterized by the invariance of H and K under the transformations

$$X_i' \quad = \quad X_i + c \ \ (i=1,\ldots,n). \tag{4.2}$$

It can be shown that there exists a uniformly most powerful invariant test which rejects H when

$$\frac{\int_{-\infty}^{\infty} g(x_1-\theta,\ldots,x_n-\theta)d\theta}{\int_{-\infty}^{\infty} f(x_1-\theta,\ldots,x_n-\theta)d\theta} \quad \text{is large.} \tag{4.3}$$

The method of proof used for Theorem 2.1 and which also works for Theorem 3.1 no longer works in the present case since the numerator (and denominator) no longer are averages. For the same reason the term average likelihood ratio is no longer appropriate and is replaced by integrated likelihood. However an easy alternative proof is given for example in Lehmann ([5], Section 6.3).

In contrast to (4.2), the maximum likelihood ratio test rejects when

$$\frac{g(x_1-\hat{\theta}_1,\ldots,x_n-\hat{\theta}_1)}{f(x_1-\hat{\theta}_0,\ldots,x_n-\hat{\theta}_0)} \quad \text{is large,} \tag{4.4}$$

where $\hat{\theta}_1$and $\hat{\theta}_0$ are the maximum likelihood estimators of θ under g and f respectively. Since (4.4) is also invariant under the transformations (4.2), it follows that the test (4.3) is uniformly at least as powerful as (4.4), and in fact more powerful unless the two tests coincide which will happen only in special cases.

The situation is quite similar for scale instead of location families. The problem (4.1) is now replaced by

$$H: \ \frac{1}{\tau^n} f(\frac{x_1}{\tau},\ldots,\frac{x_n}{\tau}) \quad \text{and} \quad K: \ \frac{1}{\tau^n} g(\frac{x_1}{\tau},\ldots,\frac{x_n}{\tau}) \tag{4.5}$$

where either the x's are all positive or f and g and symmetric about 0 in each variable.

This problem remains invariant under the transformations

$$X_i' = cX_i, \ c>0. \tag{4.6}$$

It can be shown that a uniformly most powerful invariant test exists and rejects H when

$$\frac{\int_0^{\infty} \nu^{n-1} g(\nu x_1,\ldots,\nu x_n)d\nu}{\int_0^{\infty} \nu^{n-1} f(\nu x_1,\ldots,\nu x_n)d\nu} \quad \text{is large.} \tag{4.7}$$

On the other hand, the maximum likelihood ratio test rejects when

$$\frac{g(\frac{x_1}{\hat\tau_1},\ldots,\frac{x_n}{\hat\tau_n})/\hat\tau_1^n}{f(\frac{x_1}{\hat\tau_0},\ldots,\frac{X_n}{\hat\tau_0.})/\hat\tau_0^n} \quad \text{is large} \tag{4.8}$$

where $\hat\tau_1$ and $\hat\tau_0$ are the maximum likelihood estimators of τ under g and f respectively. Since it is invariant under the transformations (4.6), the test (4.8) is less powerful than (4.7) unless they coincide.

As in (4.3), the test (4.7) involves an integrated likelihood, but while in (4.3) the parameter θ was integrated with respect to Lebesgue measure, the nuisance parameter in (4.6) is integrated with respect to $\nu^{n-1}d\nu$. A crucial feature which all the examples of Sections 2–4 have in common is that the group of transformations that leave H and K invariant is transitive i.e. that there exists a transformation which for any two members of H (or of K) takes one into the other. A general theory of this case is given in Eaton ([3], Sections 6.7 and 6.4).

Elimination of nuisance parameters through integrated likelihood is recommended very generally by Berger, Liseo and Wolpert [1]. For the case that invariance considerations do not apply, they propose integration with respect to non-informative priors over the nuisance parameters. (For a review of such prior distributions, see Kass and Wasserman [4]).

5. The failure of intuition

The examples of the previous sections show that the intuitive appeal of maximum likelihood ratio tests can be misleading. (For related findings see Berger and Wolpert ([2], pp. 125–135)). To understand just how intuition can fail, consider a family of densities p_θ and the hypothesis $H\colon \theta = 0$. The Neyman–Pearson lemma tells us that when testing p_0 against a specific p_θ, we should reject y in preference to x when

$$\frac{p_\theta(x)}{p_0(x)} < \frac{p_\theta(y)}{p_0(y)} \tag{5.1}$$

the best test therefore rejects for large values of $p_\theta(x)/p_0(x)$, i.e. is the maximum likelihood ratio test.

However, when more than one value of θ is possible, consideration of only large values of $p_\theta(x)/p_0(x)$ (as is done by the maximum likelihood ratio test) may no longer be the right strategy. Values of x for which the ratio $p_\theta(x)/p_0(x)$ is small now also become important; they may have to be included in the rejection region because $p_{\theta'}(x)/p_0(x)$ is large for some other value θ'.

This is clearly seen in the situation of Theorem 2 with f increasing and g decreasing, as illustrated in Fig. 1.

For the values of x for which f is large, g is small, and vice versa. The behavior of the test therefore depends crucially on values of x for which $f(x)$ or $g(x)$ is small, a fact that is completely ignored by the maximum likelihood ratio test.

Note however that this same phenomenon does not arise when all the alternative densities f, $g,\ldots$ are increasing. When $n = 1$, there then exists a uniformly most powerful test and it is the maximum likelihood ratio test. This is no longer true when $n > 1$, but even then all reasonable tests, including the maximum likelihood ratio test, will reject the hypothesis in a region where all the observations are large.

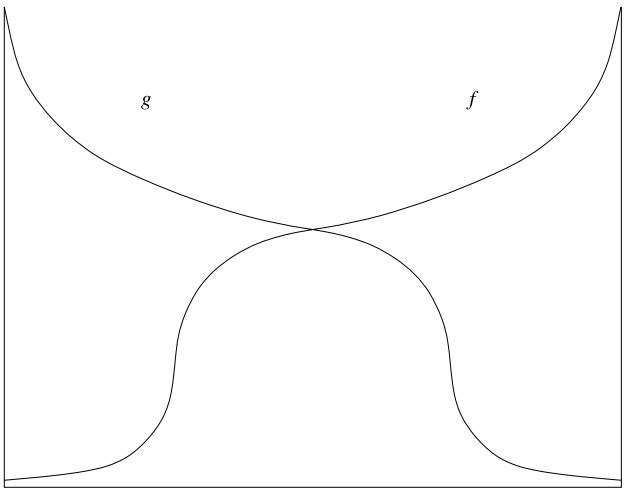


FIG 1.

6. Conclusions

For the reasons indicated in Section 1, maximum likelihood ratio tests are so widely accepted that they almost automatically are taken as solutions to new testing problems. In many situations they turn out to be very satisfactory, but gradually a collection of examples has been building up and is augmented by those of the present paper, in which this is not the case.

In particular when the problem remains invariant under a transitive group of transformations, a different principle (likelihood averaged or integrated with respect to an invariant measure) provides a test which is uniformly at least as good as the maximum likelihood ratio test and is better unless the two coincide. From the argument in Section 2 it is seen that this superiority is not restricted to invariant situations but persists in many other cases. A similar conclusion was reached from another point of view by Berger, Liseo and Wolpert [1].

The integrated likelihood approach without invariance has the disadvantage of not being uniquely defined; it requires the choice of a measure with respect to which to integrate. Typically it will also lead to more complicated test statistics. Nevertheless: In view of the superiority of integrated over maximum likelihood for large classes of problems, and the considerable unreliability of maximum likelihood ratio tests, further comparative studies of the two approaches would seem highly desirable.

References

[1] BERGER, J. O., LISEO, B. AND WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters (with discussion). *Statist. Sci.* **14**, 1–28.

[2] BERGER, J. O. AND WOLPERT, R. L. (1984). *The Likelihood Principle*. IMS Lecture Notes, Vol. 6. Institue of Mathematical Statistics, Haywood, CA.

[3] EATON, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Haywood, CA.

[4] KASS, R. E. AND WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.

[5] Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd Edition. Springer-Verlag, New York.
[6] Menéndez, J. A., Rueda, C. and Salvador, B. (1992). Dominance of likelihood ratio tests under cone constraints. *Amer. Statist.* **20**, 2087, 2099.
[7] Perlman, M. D. and Wu, L. (1999). The Emperor's new tests (with discussion). *Statist. Sci.* **14**, 355–381.
[8] Royall, R. (1997). *Statistical Evidence*. Chapman and Hall, Boca Raton.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 9–15
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000365

Student's t-test for scale mixture errors

Gábor J. Székely[1]

Bowling Green State University, Hungarian Academy of Sciences

Abstract: Generalized t-tests are constructed under weaker than normal conditions. In the first part of this paper we assume only the *symmetry* (around zero) of the error distribution (i). In the second part we assume that the error distribution is a *Gaussian scale mixture* (ii). The optimal (smallest) critical values can be computed from generalizations of Student's cumulative distribution function (cdf), $t_n(x)$. The cdf's of the generalized t-test statistics are denoted by (i) $t_n^S(x)$ and (ii) $t_n^G(x)$, resp. As the sample size $n \to \infty$ we get the counterparts of the standard normal cdf $\Phi(x)$: (i) $\Phi^S(x) := \lim_{n\to\infty} t_n^S(x)$, and (ii) $\Phi^G(x) := \lim_{n\to\infty} t_n^G(x)$. Explicit formulae are given for the underlying new cdf's. For example $\Phi^G(x) = \Phi(x)$ iff $|x| \geq \sqrt{3}$. Thus the classical 95% confidence interval for the unknown expected value of Gaussian distributions covers the center of symmetry with at least 95% probability for Gaussian scale mixture distributions. On the other hand, the 90% quantile of Φ^G is $4\sqrt{3}/5 = 1.385\cdots > \Phi^{-1}(0.9) = 1.282\ldots$.

1. Introduction

An inspiring recent paper by Lehmann [9] summarizes Student's contributions to small sample theory in the period 1908–1933. Lehmann quoted Student [10]: "The question of applicability of normal theory to non-normal material is, however, of considerable importance and merits attention both from the mathematician and from those of us whose province it is to apply the results of his labours to practical work."

In this paper we consider two important classes of distributions. The first class is the class of all symmetric distributions. The second class consists of scale mixtures of normal distributions which contains all symmetric stable distributions, Laplace, logistic, exponential power, Student's t, etc. For scale mixtures of normal distributions see Kelker [8], Efron and Olshen [5], Gneiting [7], Benjamini [1]. Gaussian scale mixtures are important in finance, bioinformatics and in many other areas of applications where the errors are heavy tailed.

First, let $X_1, X_2, \ldots, X_n$ be independent (not necessarily identically distributed) observations, and let μ be an unknown parameter with

$$X_i = \mu + \xi_i, \quad i = 1, 2, \ldots, n,$$

where the random errors $\xi_i,\ 1 \leq i \leq n$ are independent, and symmetrically distributed around zero. Suppose that

$$\xi_i = s_i \eta_i, \quad i = 1, 2, \ldots, n,$$

where $s_i, \eta_i\ i = 1, 2, \ldots, n$ are independent pairs of random variables, and the random scale, $s_i \geq 0$, is also independent of η_i. We also assume the η_i variables are

[1]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403-0221 and Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary, e-mail: gabors@bgnet.bgsu.edu

AMS 2000 subject classifications: primary 62F03; secondary 62F04.

Keywords and phrases: generalized t-tests, symmetric errors, Gaussian scale mixture errors.

identically distributed with given cdf F such that $F(x) + F(-x^-) = 1$ for all real numbers x.

Student's t-statistic is defined as $T_n = \sqrt{n}(\overline{X} - \mu)/S,\ n = 2, 3, \dots$ where $\overline{X} = \sum_{i=1}^n X_i/n$ and $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1) \neq 0$.

Introduce the notation

$$a^2 := \frac{nx^2}{x^2 + n - 1}.$$

For $x \geq 0$,

$$P\{|T_n| > x\} = P\{T_n^2 > x^2\} = P\left\{ \frac{\left(\sum_{i=1}^n \xi_i\right)^2}{\sum_{i=1}^n \xi_i^2} > a^2 \right\}. \tag{1.1}$$

(For the idea of this equation see Efron [4] p. 1279.) Conditioning on the random scales $s_1, s_2, \dots, s_n$, (1.1) becomes

$$\begin{aligned} P\{|T_n| > x\} &= EP\left\{ \frac{\left(\sum_{i=1}^n s_i\eta_i\right)^2}{\sum_{i=1}^n s_i^2\eta_i^2} > a^2 | s_1, s_2, \dots, s_n) \right\} \\ &\leq \sup_{\substack{\sigma_k \geq 0 \\ k=1,2,\dots,n}} P\left\{ \frac{\left(\sum_{i=1}^n \sigma_i\eta_i\right)^2}{\sum_{i=1}^n \sigma_i^2\eta_i^2} > a^2 \right\}, \end{aligned}$$

where $\sigma_1, \sigma_2, \dots, \sigma_n$ are arbitrary nonnegative, *non-random* numbers with $\sigma_i > 0$ for at least one $i = 1, 2, \dots, n$.

For Gaussian errors $P\{|T_n| > x\} = P(|t_{n-1}| > x)$ where t_{n-1} is a t-distributed random variable with $n-1$ degrees of freedom. The corresponding cdf is denoted by $t_{n-1}(x)$. Suppose $a \geq 0$. For scale mixtures of the cdf F introduce

$$1 - t_{n-1}^{(F)}(a) := \frac{1}{2} \sup_{\substack{\sigma_k \geq 0 \\ k=1,2,\dots,n}} P\left\{ \frac{\left(\sum_{i=1}^n \sigma_i\eta_i\right)^2}{\sum_{i=1}^n \sigma_i^2\eta_i^2} > a^2 \right\}. \tag{1.2}$$

For $a < 0$, $t_{n-1}^{(F)}(a) := 1 - t_{n-1}^{(F)}(-a)$. It is clear that if $1 - t_{n-1}^{(F)}(a) \leq \alpha/2$, then $P\{|T_n| > x\} \leq \alpha$. This is the starting point of our two excursions.

First, we assume F is the cdf of a symmetric Bernoulli random variable supported on ± 1 ($p = 1/2$). In this case the set of scale mixtures of F is the complete set of symmetric distributions around 0, and the corresponding t is denoted by t^S ($t_n^S(x) = t_n^{(F)}(x)$ when F is the Bernoulli cdf). In the second excursion we assume F is Gaussian, the corresponding t is denoted by t^G.

How to choose between these two models? If the error tails are lighter than the Gaussian tails, then of course we cannot apply the Gaussian scale mixture model. On the other hand, there are lots of models (for example the variance gamma model in finance) where the error distributions are supposed to be scale mixtures of Gaussian distributions (centered at 0). In this case it is preferable to apply the second model because the corresponding upper quantiles are smaller. For an intermediate model where the errors are symmetric and unimodal see Székely and Bakirov [11]. Here we could apply a classical theorem of Khinchin (see Feller [6]); according to this theorem all symmetric unimodal distributions are scale mixtures of symmetric uniform distributions.

2. Symmetric errors: scale mixtures of coin flipping variables

Introduce the Bernoulli random variables ε_i, $P(\varepsilon_i = \pm 1) = 1/2$. Let $\mathcal{P}$ denote the set of vectors $\mathbf{p} = (p_1, p_2, \dots, p_n)$ with Euclidean norm 1, $\sum_{k=1}^n p_k^2 = 1$. Then, according to (1.2), if the role of η_i is played by ε_i with the property that $\varepsilon_i^2 = 1$,

$$1 - t_{n-1}^S(a) = \sup_{\mathbf{p}\in\mathcal{P}} P\left\{p_1\varepsilon_1 + p_2\varepsilon_2 + \cdots + p_n\varepsilon_n \geq a\right\}.$$

The main result of this section is the following.

Theorem 2.1. *For* $0 < a \leq \sqrt{n}$, $2^{-\lceil a^2\rceil} \leq 1 - t_{n-1}^S(a) = \frac{m}{2^n}$,

where m *is the maximum number of vertices* $\mathbf{v} = (\overbrace{\pm 1, \pm 1, \dots, \pm 1}^{n})$ *of the* n*-dimensional standard cube that can be covered by an* n*-dimensional closed sphere of radius* $r = \sqrt{n - a^2}$*. (For* $a > \sqrt{n}$, $1 - t_{n-1}^S(a) = 0$*.)*

Proof. Denote by $\mathcal{P}_a$ the set of all n-dimensional vectors with Euclidean norm a. The crucial observation is the following. For all $a > 0$,

$$\begin{aligned} 1 - t_{n-1}^S(a) &= \sup_{\mathbf{p}\in\mathcal{P}} P\left\{\sum_{j=1}^n p_j\varepsilon_j \geq a\right\} \\ &= \sup_{\mathbf{p}\in\mathcal{P}_a} P\left\{\sum_{j=1}^n (\varepsilon_j - p_j)^2 \leq n - a^2\right\}. \end{aligned} \tag{2.1}$$

Here the inequality $\sum_{j=1}^n (\varepsilon_j - p_j)^2 \leq n - a^2$ means that the point

$$\mathbf{v} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n),$$

a vertex of the n dimensional standard cube, falls inside the (closed) sphere $G(\mathbf{p}, r)$ with center $\mathbf{p} \in \mathcal{P}_a$ and radius $r = \sqrt{n - a^2}$. Thus

$$1 - t_n^S(a) = \frac{m}{2^n},$$

where m is the maximal number of vertices $\mathbf{v} = (\overbrace{\pm 1, \pm 1, \dots, \pm 1}^{n})$ which can be covered by an n-dimensional closed sphere with given radius $r = \sqrt{n - a^2}$ and varying center $\mathbf{p} \in \mathcal{P}_a$. It is clear that without loss of generality we can assume that the Euclidean norm of the optimal center is a.

If $k \geq 0$ is an integer and $a^2 \leq n - k$, then $m \geq 2^k$ because one can always find 2^k vertices which can be covered by a sphere of radius $\sqrt{k}$. Take, e.g., the vertices

$$(\overbrace{1, 1, 1, \dots, 1}^{n-k}, \overbrace{\pm 1, \pm 1, \dots, \pm 1}^{k}),$$

and the sphere $G(\mathbf{c}, \sqrt{k})$ with center

$$\mathbf{c} = (\overbrace{1, 1, \dots, 1}^{n-k}, \overbrace{0, 0, \dots, 0}^{k}).$$

With a suitable constant $0 < C \leq 1$, $\mathbf{p} = C\mathbf{c}$ has norm a and since the squared distances of $\mathbf{p}$ and the vertices above are $kC \leq k$, the sphere $G(\mathbf{p}, \sqrt{k})$ covers 2^k vertices. This proves the lower bound $2^{-\lceil a^2\rceil} \leq 1 - t_n^S(a)$ in the Theorem. Thus the theorem is proved. □

Remark 1. Critical values for the t^S-test can be computed as the infima of the x-values for which $t^S_{n-1}\left(\sqrt{\frac{nx^2}{n-1+x^2}}\right) \le \alpha$.

Remark 2. Define the counterpart of the standard normal distribution as follows.

$$\Phi^S(a) \stackrel{def}{=} \lim_{n\to\infty} t^S_n(a).$$

Theorem 1 implies that for $a > 0$,

$$1 - 2^{-\lceil a^2 \rceil} \le \Phi^S(a). \tag{2.2}$$

Our computations suggest that the upper tail probabilities of Φ^S can be approximated by $2^{-\lceil a^2 \rceil}$ so well that the .9, .95, .975 quantiles of Φ^S are equal to $\sqrt{3}$, 2, $\sqrt{5}$, resp. with at least three decimal precision. We conjecture that $\Phi^S(\sqrt{3}) = .9$, $\Phi^S(2) = .95$, $\Phi^S(\sqrt{5}) = .975$. On the other hand, the .999 and higher quantiles almost coincide with the corresponding standard normal quantiles, thus in this case we do not need to pay a heavy price for dropping the condition of normality. On this problem see also the related papers by Eaton [2] and Edelman [3].

3. Gaussian scale mixture errors

An important subclass of symmetric distributions consists of the scale mixture of Gaussian distributions. In this case the errors can be represented in the form $\xi_j = s_i Z_i$ where $s_i \ge 0$ as before and independent of the standard normal Z_i. We have the equation

$$1 - t^G_{n-1}(a) = \sup_{\substack{\sigma_k \ge 0 \\ k=1,2,\dots,n}} P\left\{ \frac{\sigma_1 Z_1 + \sigma_2 Z_2 + \cdots + \sigma_n Z_n}{\sqrt{\sigma_1^2 Z_1^2 + \sigma_2^2 Z_2^2 + \cdots + \sigma_n^2 Z_n^2}} \ge a \right\}. \tag{3.1}$$

Recall that $a^2 = \frac{nx^2}{n-1+x^2}$ and thus $x = \sqrt{\frac{a^2(n-1)}{n-a^2}}$.

Theorem 3.1. *Suppose $n > 1$. Then for $0 \le a < 1$, $t^G_{n-1}(a) = 1/2$, $t^G_{n-1}(1) = 3/4$, for $a \ge \sqrt{n}$, $t^G_{n-1}(a) = 1$, and finally, for $1 < a < \sqrt{n}$,*

$$\begin{aligned} 1 - t^G_{n-1}(a) &= \max_{1\le k\le n} P\left\{ \frac{Z_1 + Z_2 + \cdots + Z_k}{\sqrt{Z_1^2 + Z_2^2 + \cdots + Z_k^2}} \ge a \right\} \\ &= \max_{a^2<k\le n} P\left(t_{k-1} > \sqrt{\frac{a^2(k-1)}{k-a^2}} \right) \end{aligned}$$

where t_{k-1} is a t-distributed random variable with $k-1$ degrees of freedom.

The point of this theorem is that $\sup_{\sigma_1,\sigma_2,\dots,\sigma_n}$ in (3.1) is taken when all nonzero $\sigma's$ are equal and here the number of zeros depends on a. For details see Székely and Bakirov [11].

Compute the intersection points of the curves

$$P\left(t_{k-1} > \sqrt{\frac{a^2(k-1)}{k-a^2}} \right)$$

for two neighboring indices. We get the following equation

$$\frac{2\Gamma\left(\frac{k}{2}\right)}{\sqrt{\pi(k-1)}\Gamma\left(\frac{k-1}{2}\right)}\int_0^{\sqrt{\frac{a^2(k-1)}{k-a^2}}}\left(1+\frac{u^2}{k-1}\right)^{-\frac{k}{2}}d\,u$$
$$=\frac{2\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\Gamma\left(\frac{k}{2}\right)}\int_0^{\sqrt{\frac{a^2k}{k+1-a^2}}}\left(1+\frac{u^2}{k}\right)^{-\frac{k+1}{2}}d\,u.$$

for the intersection point $A(k)$. It is not hard to show that $\lim_{k\to\infty}A(k)=\sqrt{3}$. This leads to the following:

Corollary 1. *There exists a sequence* $A(1):=1<A(2)<A(3)<\cdots<A(k)\longrightarrow\sqrt{3}$, *such that*

(i) *for* $a\in[A(k-1),A(k)]$, $\quad k=2,3,\ldots,n-1$,

$$t^G_{n-1}(a)=P\left(t_{k-1}>\sqrt{\frac{a^2(k-1)}{k-a^2}}\right),$$

(ii) *for* $a\geq\sqrt{3}$ *that is for* $x>\sqrt{3(n-1)/(n-3)}$,

$$t^G_{n-1}(a)=t_{n-1}(a).$$

The most surprising part of Corollary 1 is of course the nice limit, $\sqrt{3}$. This shows that above $\sqrt{3}$ the usual t-test applies even if the errors are not necessarily normals only scale mixtures of normals. Below $\sqrt{3}$, however, the 'robustness' of the t-test gradually decreases. Splus can easily compute that $A(2)=1.726,\quad A(3)=2.040$. According to our Table 1, the one sided 0.025 level critical values coincide with the classical t-critical values.

Recall that for $x\geq 0$, the Gaussian scale mixture counterpart of the standard normal cdf is

$$\Phi^G(x):=\lim_{n\to\infty}t^G_n(x) \tag{3.2}$$

(Note that in the limit, as $n\to\infty$, we have $a=x$ if both are assumed to be nonnegative; $\Phi^G(-x)=1-\Phi^G(x)$.)

Corollary 2. *For* $0\leq x<1$, $\Phi^G(x)=.5$, $\Phi^G(1)=.75$, *and for* $x\geq\sqrt{3}$, $\Phi^G(x)=\Phi(x)$, *where* $\Phi(x)$ *is the standard normal cdf (*$\Phi^G(\sqrt{3})=\Phi(\sqrt{3})=0.958$*).*

For quantiles between .5 and .875 the max in Theorem 3.1 is taken at $k=2$ and thus in this interval $\Phi^G(x)=C(x/\sqrt{(2-x^2)})$, where $C(x)$ is the standard Cauchy cdf. This is the convex section of the curve $\Phi^G(x),x\geq 0$. Interestingly the convex part is followed by a linear section: $\Phi^G(x)=x/(2\sqrt{3})+1/2$ for $1.3136\cdots<x<1.4282\ldots$. Thus the 90% quantile is exactly $\quad 4\sqrt{3}/5:\quad \Phi^G(4\sqrt{3}/5)=0.9$. The following critical values are important in applications: $0.95=\Phi(1.645)=\Phi^G(1.650)$, $0.9=\Phi(1.282)=\Phi^G(1.386)$, $0.875=\Phi(1.150)=\Phi^G(1.307)$ (see the last row of Table 1).

Remark 3. It is clear that for $a>0$ we have the inequalities $t_n(a)\geq t^G_n(a)\geq t^S_n(a)$. According to Corollary 1, the first inequality becomes an equality iff $a\geq\sqrt{3}$. In connection with the second inequality one can show that the difference of the α-quantiles of $t^G_n(a)$ and $t^S_n(a)$ tends to 0 as $\alpha\to 1$.

TABLE 1
Critical values for Gaussian scale mixture errors computed from $t_n^G(\sqrt{nx^2/(n-1+x^2)}) = \alpha$

$n-1$	0.125	0.100	0.050	0.025
2	1.625	1.886	2.920	4.303
3	1.495	1.664	2.353	3.182
4	1.440	1.579	2.132	2.776
5	1.410	1.534	2.015	2.571
6	1.391	1.506	1.943	2.447
7	1.378	1.487	1.895	2.365
8	1.368	1.473	1.860	2.306
9	1.361	1.462	1.833	2.262
10	1.355	1.454	1.812	2.228
11	1.351	1.448	1.796	2.201
12	1.347	1.442	1.782	2.179
13	1.344	1.437	1.771	2.160
14	1.341	1.434	1.761	2.145
15	1.338	1.430	1.753	2.131
16	1.336	1.427	1.746	2.120
17	1.335	1.425	1.740	2.110
18	1.333	1.422	1.735	2.101
19	1.332	1.420	1.730	2.093
20	1.330	1.419	1.725	2.086
21	1.329	1.417	1.722	2.080
22	1.328	1.416	1.718	2.074
23	1.327	1.414	1.715	2.069
24	1.326	1.413	1.712	2.064
25	1.325	1.412	1.709	2.060
100	1.311	1.392	1.664	1.984
500	1.307	1.387	1.652	1.965
1,000	1.307	1.386	1.651	1.962

Our approach can also be applied for two-sample tests. In a joint forthcoming paper with N. K. Bakirov the Behrens–Fisher problem will be discussed for Gaussian scale mixture errors with the help of our $t_n^G(x)$ function.

Acknowledgments

The author also wants to thank many helpful suggestions of N. K. Bakirov, M. Rizzo, the referees of the paper, and the editor of the volume.

References

[1] BENJAMINI, Y. (1983). Is the t-test really conservative when the parent distribution is long-tailed? *J. Amer. Statist. Assoc.* **78**, 645–654.
[2] EATON, M.L. (1974). A probability inequality for linear combinations of bounded random variables. *Ann. Statist.* **2**,3, 609–614.
[3] EDELMAN, D. (1990). An inequality of optimal order for the probabilities of the T statistic under symmetry. *J. Amer. Statist. Assoc.* **85**, 120–123.
[4] EFRON, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64**, 1278–1302.
[5] EFRON, B. AND OLSHEN, R. A. (1978). How broad is the class of normal scale mixtures? *Ann. Statist.***6**, 5, 1159–1164.

[6] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. 2. Wiley, New York.
[7] Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *J. Statist. Comput. Simul.* **59**, 375–384.
[8] Kelker, D. (1971). Infinite divisibility and variance mixtures of the normal distribution. *Ann. Math. Statist.* **42**, 2, 802–808.
[9] Lehmann, E.L. (1999). 'Student' and small sample theory. *Statistical Science* **14**, 4, 418–426.
[10] Student (1929). Statistics in biological research. *Nature* **124**, 93.
[11] Székely, G.J. and N.K. Bakirov (under review). Generalized t-tests for unimodal and normal scale mixture errors.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 16–32
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000374

Recent developments towards optimality in multiple hypothesis testing

Juliet Popper Shaffer[1]

University of California

Abstract: There are many different notions of optimality even in testing a single hypothesis. In the multiple testing area, the number of possibilities is very much greater. The paper first will describe multiplicity issues that arise in tests involving a single parameter, and will describe a new optimality result in that context. Although the example given is of minimal practical importance, it illustrates the crucial dependence of optimality on the precise specification of the testing problem. The paper then will discuss the types of expanded optimality criteria that are being considered when hypotheses involve multiple parameters, will note a few new optimality results, and will give selected theoretical references relevant to optimality considerations under these expanded criteria.

1. Introduction

There are many notions of optimality in testing a single hypothesis, and many more in testing multiple hypotheses. In this paper, consideration will be limited to cases in which there are a finite number of individual hypotheses, each of which ascribes a specific value to a single parameter in a parametric model, except for a small but important extension: consideration of directional hypothesis-pairs concerning single parameters, as described below. Furthermore, only procedures for continuous random variables will be considered, since if randomization is ruled out, multiple tests can always be improved by taking discreteness of random variables into consideration, and these considerations are somewhat peripheral to the main issues to be addressed.

The paper will begin by considering a single hypothesis or directional hypothesis-pair, where some of the optimality issues that arise can be illustrated in a simple situation. Multiple hypotheses will be treated subsequently. Two previous reviews of optimal results in multiple testing are Hochberg and Tamhane [28] and Shaffer [58]. The former includes results in confidence interval estimation while the latter is restricted to hypothesis testing.

2. Tests involving a single parameter

Two conventional types of hypotheses concerning a single parameter are

$$H : \theta \leq 0 \ \ vs.\ A : \theta > 0, \tag{2.1}$$

[1]Department of Statistics, 367 Evans Hall # 3860, Berkeley, CA 94720-3860, e-mail: shaffer@stat.berkeley.edu

AMS 2000 subject classifications: primary 62J15; secondary 62C25, 62C20.

Keywords and phrases: power, familywise error rate, false discovery rate, directional inference, Types I, II and III errors.

which will be referred to as a one-sided hypothesis, with the corresponding tests being referred to as one-sided tests, and

$$H : \theta = 0 \ \ vs, \ \ A : \theta \neq 0, \tag{2.2}$$

which will be referred to as a two-sided, or nondirectional hypothesis, with the corresponding tests being referred to as nondirectional tests. A variant of (2.1) is

$$H : \theta = 0 \ \ vs. \ A : \theta > 0, \tag{2.3}$$

which may be appropriate when the reverse inequality is considered virtually impossible. Optimality considerations in these tests require specification of optimality criteria and restrictions on the procedures to be considered. While often the distinction between (2.1) and (2.3) is unimportant, it leads to different results in some cases. See, for example, Cohen and Sackrowitz [14] where optimality results require (2.3) and Lehmann, Romano and Shaffer [41], where they require (2.1).

Optimality criteria involve consideration of two types of error: Given a hypothesis H, Type I error (rejecting $H|H$ true) and Type II error ("accepting" $H|H$ false), where the term "accepting" has various interpretations. The reverse of Type I error (accepting $H|H$ true) and Type II error (rejecting $H|H$ false, or power) are unnecessary to consider in the one-parameter case but must be considered when multiple parameters are involved and there may be both true and false hypotheses.

Experimental design often involves fixing both P(Type I error) and P(Type II error) and designing an experiment to achieve both goals. This paper will not deal with design issues; only analysis of fixed experiments will be covered.

The Neyman–Pearson approach is to minimize P(Type II error) at some specified nonnull configuration, given fixed max P(Type I error). (Alternatively, P(Type II error) can be fixed at the nonnull configuration and P(Type I error) minimized.) Lehmann [37,38] discussed the optimal choice of the Type I error rate in a Neyman-Pearson frequentist approach by specifying the losses for accepting H and rejecting H, respectively.

In the one-sided case (2.1) and/or (2.3), it is sometimes possible to find a uniformly most powerful test, in which case no restrictions need be placed on the procedures to be considered. In the two-sided formulation (2.2), this is rarely the case. When such an ideal method cannot be found, restrictions are considered (symmetry or invariance, unbiasedness, minimaxity, maximizing local power, monotonicity, stringency, etc.) under which optimality results may be achievable. All of these possibilities remain relevant with more than one parameter, in generalized form.

A Bayes approach to (2.1) is given in Casella and Berger [12], and to (2.2) in Berger and Sellke [8]; the latter requires specification of a point mass at $\theta = 0$, and is based on the posterior probability at zero. See Berger [7] for a discussion of Bayes optimality. Other Bayes approaches are discussed in later sections.

2.1. Directional hypothesis-pairs

Consider again the two-sided hypothesis (2.2). Strictly speaking, we can only either accept or reject H. However, in many, perhaps most, situations, if H is rejected we are interested in deciding whether θ is $<$ or $>$ 0. In that case, there are three possible inferences or decisions: (i) $\theta > 0$, (ii) $\theta = 0$, or (iii) $\theta < 0$, where the decision (ii) is sometimes interpreted as uncertainty about θ. An alternative formulation as

a pair of hypotheses can be useful:

$$\begin{aligned} H_1 &: \theta \leq 0 \ \ vs.\ A_1 : \theta > 0 \\ H_2 &: \theta \geq 0 \ \ vs.\ A_2 : \theta < 0 \end{aligned} \tag{2.4}$$

where the sum of the rejection probabilities of the pair of tests if $\theta = 0$ is equal to α (or at most α). Formulation (2.4) will be referred to as a directional hypothesis-pair.

2.2. Comparison of the nondirectional and directional-pair formulations

The two-sided or non-directional formulation (2.2) is appropriate, for example, in preliminary tests of model assumptions to decide whether to treat variances as equal in testing for means. It also may be appropriate in testing genes for differential expression in a microarray experiment: Often the most important goal in that case is to discover genes with differential expression, and further laboratory work will elucidate the direction of the difference. (In fact, the most appropriate hypothesis in gene expression studies might be still more restrictive: that the two distributions are identical. Any type of variation in distribution of gene expression in different tissues or different populations could be of interest.)

The directional-pair formulation (2.4) is usually more appropriate in comparing the effectiveness of two drugs, two teaching methods, etc. Or, since there might be some interest in discovering both a difference in distributions as well as the direction of the average difference or other specific distribution characteristic, some optimal method for achieving a mix of these goals might be of interest. A decision-theoretic formulation could be developed for such situations, but they do not seem to have been considered in the literature. The possible use of unequal-probability tails is relevant here (Braver [11], Mantel [44]), although these authors proposed unequal-tail use as a way of compromising between a one-sided test procedure (2.1) and a two-sided procedure (2.2).

Note that (2.4) is a multiple testing problem. It has a special feature: only one of the hypotheses can be false, and no reasonable test will reject more than one. Thus, in formulation (2.4), there are three possible types of errors:

Type I: Rejecting either H_1 or H_2 when both are true.

Type II: Accepting both H_1 and H_2 when one is false.

Type III: Rejecting H_1 when H_2 is false or rejecting H_2 when H_1 is false; i.e. rejecting $\theta = 0$, but making the wrong directional inference.

If it does not matter what conclusion is reached in (2.4) when $\theta = 0$, only Type III errors would be considered.

Shaffer [57] enumerated several different approaches to the formulation of the directional pair, variations on (2.4), and considered different criteria as they relate to these approaches. Shaffer [58] compared the three-decision and the directional hypothesis-pair formulations, noting that each was useful in suggesting analytical approaches.

Lehmann [35,37,38], Kaiser [32] and others considered the directional formulation (2.4), sometimes referring to it alternatively as a three-decision problem. Bahadur [1] treated it as deciding $\theta < 0$, $\theta > 0$, or reserving judgment. Other references are given in Finner [24].

In decision-theoretic approaches, losses can be defined as 0 for the correct decision and 1 for the incorrect decision, or as different for Type I, Type II, and Type

III errors (Lehmann [37,38]), or as proportional to deviations from zero (magnitude of Type III errors) as in Duncan's [17] Bayesian pairwise comparison method. Duncan's approach is applicable also if (2.4) is modified to eliminate the equal sign from at least one of the two elements of the pair, so that no assumption of a point mass at zero is necessary, as it is in the Berger and Sellke [8] approach to (2.1), referred to previously.

Power can be defined for the hypothesis-pair as the probability of rejecting a false hypothesis. With this definition, power excludes Type III errors. Assume a test procedure in which the probability of no errors is $1 - \alpha$. The change from a nondirectional test to a directional test-pair makes a big difference in the performance at the origin, where power changes from α to $\alpha/2$ in an equal-tails test under mild regularity conditions. However, in most situations it has little effect on test power where the power is reasonably large, since typically the probability of Type III errors decreases rapidly with an increase in nondirectional power.

Another simple consequence is that this reformulation leads to a rationale for using equal-tails tests in asymmetric situations. A nondirectional test is unbiased if the probability of rejecting a true hypothesis is smaller than the probability of rejecting a false hypothesis. The term "bidirectional unbiased" is used in Shaffer [54] to refer to a test procedure for (2.4) in which the probability of making the wrong directional decision is smaller than the probability of making the correct directional decision. That typically requires an equal-tails test, which might not maximize power under various criteria given the formulation (2.2).

It would seem that except for this result, which can affect only the division between the tails, usually minimally, the best test procedures for (2.2) and (2.4) should be equivalent, except that in (2.2) the absolute value of a test statistic is sometimes sufficient for acceptance or rejection, whereas in (2.4) the signed value is always needed to determine the direction. However, it is possible to contrive situations in which the optimal test procedures under the two formulations are diametrically opposed, as is demonstrated in the extension of an example from Lehmann [39], described below.

2.3. An example of diametrically different optimal properties under directional and nondirectional formulations

Lehmann [39] contends that tests based on average likelihood are superior to tests based on maximum likelihood, and describes qualitatively a situation in which the best symmetric test based on average likelihood is the most-powerful test and the best symmetric test based on maximum likelihood is the least-powerful symmetric test. Although the problem is not of practical importance, it is interesting theoretically in that it illustrates the possible divergence between test procedures based on (2.2) and on (2.4). A more specific illustration of Lehmann's example can be formulated as follows:

Suppose, for $0 < X < 1$, and $\gamma > 0$, known:

$$f_0(x) \equiv 1, \text{ i.e. } f_0 \text{ is Uniform (0,1)}$$
$$f_1(x) = (1+\gamma)x^{\gamma},$$
$$f_2(x) = (1+\gamma)(1-x)^{\gamma}.$$

Assume a single observation and test $H : f_0(x)$ vs. $A : f_1(x)$ or $f_2(x)$.

One of the elements of the alternative is an increasing curve and the other is decreasing. Note that this is a nondirectional hypothesis, analogous to (2.2) above.

It seems reasonable to use a symmetric test, since the problem is symmetric in the two directions. If $\gamma > 1$ (convex f_1 and f_2), the maximum likelihood ratio test (MLR) and the average likelihood ratio test (ALR) coincide, and the test is the most powerful symmetric test:

Reject H if $0 < x < \alpha/2$ or $1 - \alpha/2 < x < 1$, i.e. an extreme-tails test. However, if $\gamma < 1$ (concave f_1 and f_2), the most-powerful symmetric test is the ALR, different from the MLR; the ALR test is:

$$\text{Reject } H \text{ if } .5 - \alpha/2 \le x \le .5 + \alpha/2, \text{ i.e. a central test.}$$

In this case, among tests based on symmetric $\alpha/2$ intervals, the MLR, using the two extreme $\alpha/2$ tails of the interval (0,1), is the *least* powerful symmetric test. In other words, regardless of the value of γ, the ALR is optimal, but coincides with the MLR only when $\gamma > 1$. Figure 1 gives the power curves for the central and extreme-tails tests over the range $0 \le \gamma \le 2$.

But suppose it is important not only to reject H but also to decide in that case whether the nonnull function is the increasing one $f_1(x) = (1 + \gamma)x^\gamma$, or the decreasing one $f_2(x) = (1 + \gamma)(1 - x)^\gamma$. Then the appropriate formulation is analogous to (2.4) above:

$$H_1 : f_0 \text{ or } f_1 \text{ vs. } A_1 : f_2$$
$$H_2 : f_0 \text{ or } f_2 \text{ vs. } A_2 : f_1.$$

If $\gamma > 1$ (convex), the most-powerful symmetric test of (2.2) (MLR and ALR) is also the most powerful symmetric test of (2.4). But if $\gamma < 1$ (concave), the most-powerful symmetric test of (2.2) (ALR) is the least-powerful while the MLR is the most-powerful symmetric test of (2.4). In both cases, the directional ALR and MLR are identical, since the alternative hypothesis consists of only a single distribution. In general, if the alternative consists of a single distribution, regardless of the dimension of the null hypothesis, the definitions of ALR and MLR coincide.

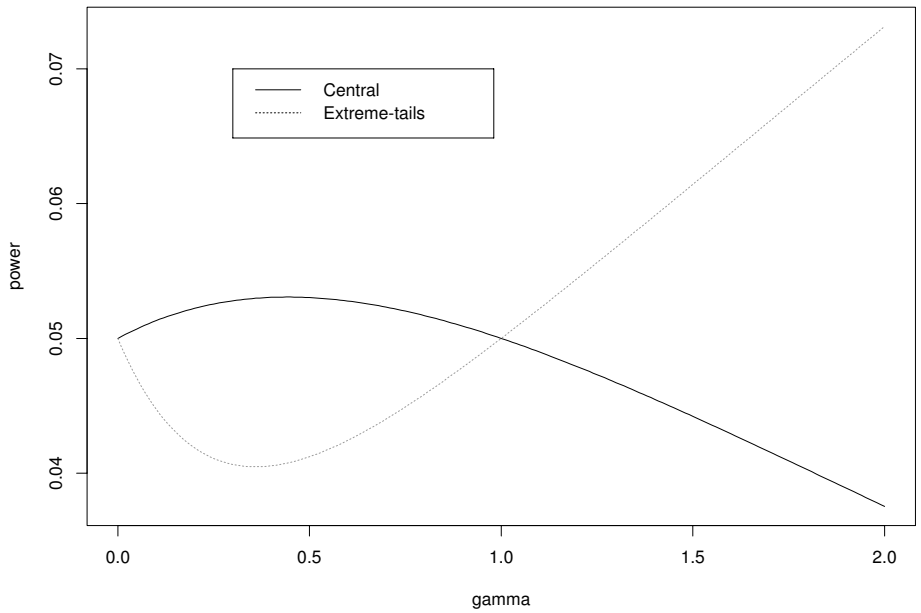


FIG 1. *Power of nondirectional central and extreme-tails tests*

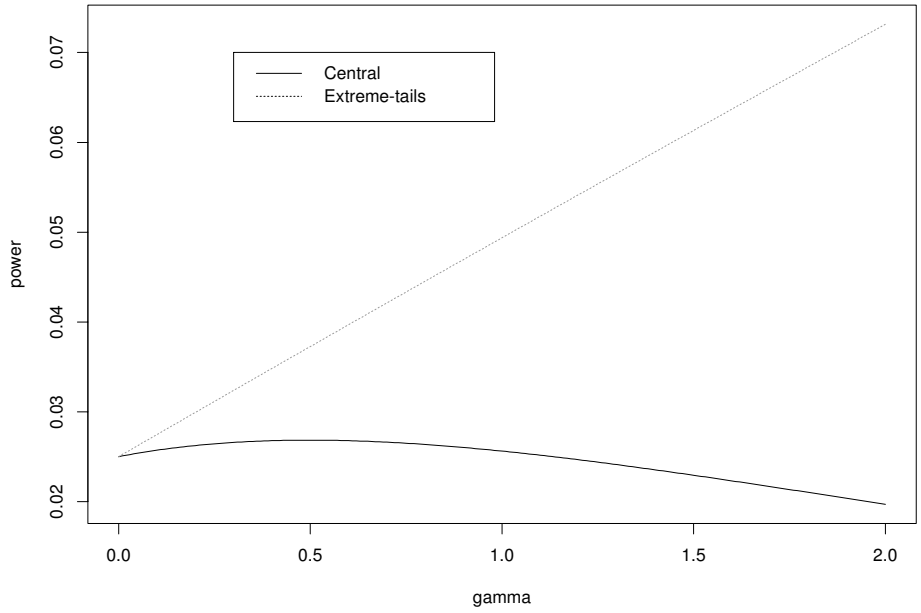


FIG 2. *Power of directional central and extreme-tails test-pairs*

Note that if γ is unknown, but is known to be < 1, the terms 'most powerful' and 'least powerful' in the example can be replaced by 'uniformly most powerful' and 'uniformly least powerful', respectively.

Figure 2 gives some power curves for the directional central and extreme-value test-pairs.

Another way to look at this directional formulaton is to note that a large part of the power of the ALR under (2.2) becomes Type III error under (2.4) when $\gamma < 1$. The Type III error not only stays high for the directional test-pair based on the central proportion α, but it actually is above the null value of $\alpha/2$ when γ is close to zero.

Of course the situation described in this example is unrealistic in many ways, and in the usual practical situations, the best test for (2.2) and for (2.4) are identical, except for the minor tail-probability difference noted. It remains to be seen whether there are realistic situations in which the two approaches diverge as radically as in this example. Note that the difference between nondirectional and directional optimality in the example generalizes to multiple parameter situations.

3. Tests involving multiple parameters

In the multiparameter case, the true and false hypotheses, and the acceptance and rejection decisions, can be represented in a two by two table (Table 1). With more than one parameter, the potential number of criteria and number of restrictions on types of tests is considerably greater than in the one-parameter case. In addition, different definitions of power, and other desirable features, can be considered. This paper will describe some of these expanded possibilities. So far optimality results have been obtained for relatively few of these conditions. The set of hypotheses to be considered jointly in defining the criteria is referred to as the family. Sometimes

TABLE 1
True states and decisions for multiple tests

Number of	Number not rejected	Number rejected	
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m_1
	m-R	R	m

the family includes all hypotheses to be tested in a given study, as is usually the case, for example, in a single-factor experiment comparing a limited number of treatments. Hypotheses tested in large surveys and multifactor experiments are usually divided into subsets (families) for error control. Discussion of choices for families can be found in Hochberg and Tamhane [28] and Westfall and Young [66].

All of the error, power, and other properties raise more complex issues when applied to tests of (2.2) than to tests of (2.1) or (2.3), and even more so to tests of (2.4) and its variants. With more than one parameter, in addition to these expanded possibilities, there are also more possible types of test procedures. For example, one may consider only stepwise tests, or, even more specifically, under appropriate distributional assumptions, only stepwise tests using either t tests or F tests or some combination. Some type of optimality might then be derived within each type. Another possibility is to derive optimal results for the sequence of probabilities to be used in a stepwise procedure without specifying the particular type of tests to be used at each stage. Optimality results may also depend on whether or not there are logical relationships among the hypotheses (for example when testing equality of all pairwise differences among a set of parameters, transitivity relationships exist). Other results are obtained under restrictions on the joint distributions of the test statistics, either independence or some restricted type of dependence. Some results are obtained under the restriction that the alternative parameters are identical.

3.1. Criteria for Type I error control

Control of Type I error with a one-sided or nondirectional hypothesis, or Type I and Type III error with a directional hypothesis-pair, can be generalized in many ways. Type II error, except for one definition below (viii), has usually been treated instead in terms of its obverse, power. Optimality results are available for only a small number of these error criteria, mainly under restricted conditions.

Until recent years, the generalized Type I error rates to be controlled were limited to the following three:

(i) The expected proportion of errors (true hypotheses rejected) among all hypotheses, or the maximum per-comparison error rate (PCER), defined as E(V/m). This criterion can be met by testing each hypothesis at the specified level, independent of the number of hypotheses; it essentially ignores the multiplicity issue, and will not be considered further.
(ii) The expected number of errors (true hypotheses rejected), or the maximum per-family error rate (PFER), where the family refers to the set of hypotheses being treated jointly, defined as E(V).
(iii) The maximum probability of one or more rejections of true hypotheses, or the familywise error rate (FWER), defined as Prob($V > 0$).

The criterion (iii) has been the most frequently adopted, as (i) is usually considered too liberal and (ii) too conservative when the same fixed conventional level

is adopted. Within the last ten years, some additional rates have been proposed to meet new research challenges, due to the emergence of new methodologies and technologies that have resulted in tests of massive numbers of hypotheses and a concomitant desire for less strict criteria.

Although there have been situations for some time in which large numbers of hypotheses are tested, such as in large surveys, and multifactor experimental designs, these hypotheses have usually been of different types and often of an indefinite number, so that error control has been restricted to subsets of hypotheses, or families, as noted above, each usually of some limited size. Within the last 20 years, there has been an explosion of interest in testing massive numbers of well-defined hypotheses in which there is no obvious basis for division into families, such as in microrarray genomic analysis, where individual hypotheses may refer to parameters of thousands of genes, to tests of coefficients in wavelet analysis, and to some types of tests in astronomy. In these cases the criterion (iii) seems to many researchers too draconian. Consequently, some new approaches to error control and power have been proposed. Although few optimal criteria have been obtained under these additional approaches, these new error criteria will be described here to indicate potential areas for optimality research.

Recently, the following error-control criteria in addition to (i)-(iii) above have been considered:

(iv) The expected proportion of falsely-rejected hypotheses among the rejected hypotheses–the false discovery rate (FDR). The proportion itself, FDP = V/R, is defined to be 0 when no hypotheses are rejected (Benjamini and Hochberg [3]; for earlier discussions of this concept see Eklund and Seeger [22], Seeger [53], Sorić [59]), so the FDR can be defined as $E(FDP|R > 0)P(R > 0)$. There are numerous publications on properties of the FDR, with more appearing continuously.

(v) The expected proportion of falsely-rejected hypotheses among the rejected hypotheses given that some are rejected (p-FDR) (Storey [62]), defined as $E(V/R)|R > 0$.

(vi) The maximum probability of at most k errors (k-FWER or g-FWER–g for generalized), given that at least k hypotheses are true, $k = 0, \ldots, m$, $P(V > k)$, (Dudoit, van der Laan, and Pollard [16], Korn, Troendle, McShane, and Simon [34], Lehmann and Romano [40], van der Laan, Dudoit, and Pollard [65], Pollard and van der Laan [48]). Some results on this measure were obtained earlier by Hommel and Hoffman [31].

(vii) The maximum proportion of falsely-rejected hypotheses among those rejected (with 0/0 defined as 0), $FDP > \gamma$ (Romano and Shaikh [51] and references listed under (vi)).

(viii) The false non-discovery rate (Genovese and Wasserman [25]), the expected proportion of nonrejected but false hypotheses among the nonrejected ones, (with 0/0 defined as 0): $FNR = E[T/(m - R)\ P(m - R > 0)]$.

(ix) The vector loss functions defined by Cohen and Sackrowitz [15], discussed below and in Section 4.

Note that the above generalizations (iv), (vi), and (vii) reduce to the same value (Type I error) when a single parameter is involved, (v) equals unity so would not be appropriate for a single test, (viii) reduces to the Type II error probability, and the FRR in (ix), defined in Section 4, is equal to (ii).

The loss function approach has been generalized as either (L_i) the sum of the loss functions for each hypothesis, (L_{ii}) a 0-1 loss function in which the loss is zero

only if all hypotheses are correctly classified, or (L_{iii}) a sum of loss functions for the FDR (iv) and the FNR ($viii$). (In connection with L_{iii}, see the discussion of Genovese and Wasserman [25] in Section 4, as well as contributions by Cheng et al [13], who also consider adjusting criteria to consider known biological results in genomic applications.) Sometimes a vector of loss functions is considered rather than a composite function when developing optimal procedures; a number of different vector approaches have been used (see the discussion of Cohen and Sackrowitz [14,15] in the section on optimality results). Many Bayes and empirical Bayes approaches involve knowing or estimating the proportion of true hypotheses, and will be discussed in that context below.

The relationships among (iv), (v), (vi) and (vii) depend partly on the variance of the number of falsely-rejected hypotheses. Owen [47] discusses previous work related to this issue, and provides a formula that takes the correlations of test statistics into account.

A contentious issue relating to generalizations (iv) to (vii) is whether rejection of hypotheses with very large p-values should be permitted in achieving control using these criteria, or whether some additional restrictions on individual p-values should be applied. For example, under (vi), k hypotheses could be rejected even if the overall test was negative, regardless of the associated p-values. Under (iv), (v), and (vii), given a sufficient number of hypotheses rejected with FWER control at α, additional hypotheses with arbitrarily large p-values can be rejected. Tukey, in a personal oral communication, suggested, in connection with (iv), that hypotheses with individual p-values greater than α might be excluded from rejection. This might be too restrictive, especially under (vi) and (vii), but some restrictions might be desirable. For example, if $\alpha = .05$, it has been suggested that hypotheses with $p \leq \alpha*$ might be considered for rejection, with $\alpha*$ possibly as large as 0.5 (Benjamini and Hochberg [4]). In some cases, it might be desirable to require nonincreasing individual rejection probabilities as m increases (with $m \geq k$ in (vi) and (vii)), which would imply Tukey's suggestion. Even the original Benjamini and Hochberg [3] FDR-controlling procedure violates this latter condition, as shown in an example in Holland and Cheung [29], who note that the adaptive method in Benjamini and Hochberg [4] violates even the Tukey suggestion. Consideration of these restrictions on error is very recent, and this issue has not yet been addressed in any serious way in the literature.

3.2. Generalizations of power, and other desirable properties

The most common generalizations of power are:

(a) probability of at least one rejection of a false hypothesis, (b) probability of rejecting all false hypotheses, (c) probability of rejecting a particular false hypothesis, and (d) average probability of rejecting false hypotheses. (The first three were initially defined in the paired-comparison situation by Ramsey [49], who called them any-pair power, all-pairs power, and per-pair power, respectively.) Generalizations (b) and (c) can be further extended to (e) the probability of rejecting more than k false hypotheses, $k = 0, \ldots, m$. Generalization (d) is also the expected proportion of false hypotheses rejected.

Two other desirable properties that have received limited attention in the literature are:

(f) complexity of the decisions. Shaffer [55] suggested the desirability, when comparing parameters, of having procedures that are close to partitioning the parameters into groups. Results that are partitions would have zero complexity; Shaffer suggested a quantitative criterion for the distance from this ideal.

(g) familywise robustness (Holland and Cheung [29]). Because the decisions on definitions of families are subjective and often difficult to make, Holland and Cheung suggested the desirability of procedures that are less sensitive to family choice, and developed some measures of this criterion.

4. Optimality results

Some optimality results under error protection criteria (i) to (iv) and under Bayesian decision-theoretic approaches were reviewed in Hochberg and Tamhane [28] and Shaffer [58]. The earlier results and some recent extensions will be reviewed below, and a few results under (vi) and (viii) will be noted. Criteria (iv) and (v) are asymptotically equivalent when there are some false hypotheses, under mild assumptions (Storey, Taylor and Siegmund [63]).

Under (ii), optimality results with additive loss functions were obtained by Lehmann [37,38], Spjøtvoll [60], and Bohrer [10], and are described in Shaffer [58] and Hochberg and Tahmane [28]. Lehmann [37,38] derived optimality under hypothesis-formulation (2.1) for each hypothesis, Spjøtvoll [60] under hypothesis-formulation (2.2), and Bohrer [10] under hypothesis-formulation (2.4), modified to remove the equality sign from one member of the pair.

Duncan [17] developed a Bayesian decision-theoretic procedure with additive loss functions under the hypothesis-formulation (2.4), applied to testing all pairwise differences between means based on normally-distributed observations, and assuming the true means are normally-distributed as well, so that the probability of two means being equal is zero. In contrast to Lehmann [37,38] and Spjøtvoll [60], Duncan uses loss functions that depend on the magnitudes of the true differences when the pair (2.4) are accepted or the wrong member of the pair (2.4) is rejected. Duncan [17] also considered an empirical Bayes version of his procedure in which the variance of the distribution of the true means is estimated from the data, pointing out that the results were almost the same as in the known-variance case when $m \geq 15$. For detailed descriptions of these decision-theoretic procedures of Lehmann, Spjøtvoll, Bohrer, and Duncan, see Hochberg and Tamhane [28].

Combining (iv) and (viii) as in L_{iii}, Genovese and Wasserman [25] consider an additive risk function combining FDR and FNR and obtain some optimality results, both finite-sample and asymptotic. If the risk δ_i for H_i is defined as 0 when the correct decision is made (either acceptance or rejection) and 1 otherwise, they define the classification risk as

$$R_m = \frac{1}{m} E\left(\sum_{i=1}^{m} |\delta_i - \hat{\delta}_i|\right), \tag{4.1}$$

equivalent to the average fraction of errors in both directions. They derive asymptotic values for R_m given various procedures and compare them under different conditions. They also consider the loss functions $FNR + \lambda FDR$ for arbitrary λ and derive both finite-sample and asymptotic expressions for minimum-risk procedures based on p-values. Further results combining (iv) and (viii) are obtained in Cheng et al [13].

Cohen and Sackrowitz [14,15] consider the one-sided formulations (2.1) and (2.3) and treat the multiple situation as a 2^m finite action problem. They assume a multivariate normal distribution for the m test statistics with a known covariance matrix of the intraclass type (equal variances, equal covariances), so the test statistics are exchangeable. They consider both additive loss functions (1 for a Type I error and an arbitrary value b for a Type II error, added over the set of hypothesis tests), the m-vector of loss functions for the m tests, and a 2-vector treating Type I and Type II errors separately, labeling the two components false rejection rate (FRR) and false acceptance rate (FAR). They investigate single-step and stepwise procedures from the points of view of admissibility, Bayes, and limits of Bayes procedures. Among a series of Bayes and decision-theoretic results, they show admissibility of single-stage and stepdown procedures, with inadmissibility of stepup procedures, in contrast to the results of Lehmann, Romano and Shaffer [41], described below, which demonstrate optimality of stepup procedures under a different loss structure.

Under error criterion (iii), early results on means are described in Shaffer [58] with references to relevant literature. Lehmann and Shaffer [42], considering multiple range tests for comparing means, found the optimal set of critical values assuming it was desirable to maximize the minimum probabilities for distinguishing among pairs of means, which implies maximizing the probabilities for comparing adjacent means. Finner [23] noted that this optimality criterion was not totally compelling, and found the optimal set under the assumption that one would want to maximize the probability of rejecting the largest range, then the next-largest, etc. He compared the resulting maximax method to the Lehmann and Shaffer [42] maximin method.

Shaffer [56] modified the empirical Bayes version of Duncan [17], decribed above, to provide control of (iii). Recently, Lewis and Thayer [43] adopted the Bayesian assumption of a normal distribution of true means as in Duncan [17]and Shaffer [56] for testing the equality of all pairwise differences among means. However, they modified the loss functions to a loss of 1 for an incorrect directional decision and α for accepting both members of the hypothesis-pair (2.4). False discoveries are rejections of the wrong member of the pair (2.4) while true discoveries are rejections of the correct member of the pair. Thus, Lewis and Thayer control what they call the directional FDR (DFDR). Under their Bayesian and loss-function assumptions, and adding the loss functions over the tests, they prove that the DFDR of the minimum Bayes-risk rule is $\leq \alpha$. They also consider an empirical Bayes variation of their method, and point out that their results provide theoretical support for an empirical finding of Shaffer [56], which demonstrated similar error properties for her modification of the Duncan [17] approach to provide control of the FWER and the Benjamini and Hochberg [3] FDR-controlling procedure. Lewis and Thayer [43] point out that the empirical Bayes version of their assumptions can be alternatively regarded as a random-effects frequentist formulation.

Recent stepwise methods have been based on Holm's [30] sequentially- rejective procedure for control of (iii). Initial results relevant to that approach were obtained by Lehmann [36], in testing one-sided hypotheses. Some recent results in Lehmann, Romano, and Shaffer [41] show optimality of stepwise procedures in testing one-sided hypotheses for controlling (iii) when the criterion is maximizing the minimum power in various ways, generalizing Lehmann's [36] results. Briefly, if rejection of at least i hypotheses are ordered in importance from $i = 1$ (most important) to $i = m$, a generalized Holm stepdown procedure is shown to be optimal, while if these rejections are ordered in importance from $i = m$ (most important) to $i = 1$, a stepup procedure generalizing Hochberg [27] is shown to be optimal.

Most of the recent literature in multiple comparisons relates to improving existing methods, rather than obtaining optimal methods, although of course such improvements indicate the directions in which optimality might be achieved. The next section is a necessarily brief and selective overview of some of this literature.

5. Literature on improvement of multiple comparison procedures

5.1. Estimating the number of true null hypotheses

Under most types of error control, if the number m_0 of true hypotheses H were known, improved procedures could be based on this knowledge. In fact, sometimes such knowledge could be important in its own right, for example in microarray analysis, where it might be of interest to estimate the number of genes differentially expressed under different conditions, or in an astronomy problem (Meinshausen and Rice [45]) in which that is the only quantity of interest.

Under error criteria (ii) and (iii), the Bonferroni method could be improved in power by carrying it out at level α/m_0 instead of α/m. FDR control with independent test statistics using the Benjamini and Hochberg [3] method is exactly equal to $\pi_0\alpha$, where $\pi_0 = m_0/m$ is the proportion of true hypotheses, (Benjamini and Hochberg [3]), so their FDR-controlling method described in that paper could be made more powerful by multiplying the criterion p-values at each stage by m/m_0. The method has been proved to be conservative under some but not all types of dependence. A modified method making use of m_0 guarantees control of the FDR at the specified level under all types of test statistic dependence (Benjamini and Yekutieli [6]).

Much recent effort has been directed towards obtaining good estimates of π_0, either for an interest in this quantity itself, or because then these improved methods and other more recent methods, including some single-step methods, could be used at level $\alpha* = \alpha/\pi_0$. There are many recent papers comparing estimates of π_0, but few optimality results are available at present. Some recent relatively theoretical references are Black [9], Genovese and Wasserman [26], Storey, Taylor, and Siegmund [63], Meinshausen and Rice [45], and Reiner, Yekutieli and Benjamini [50].

Storey, Taylor, and Siegmund [63] use empirical process theory to investigate proposed procedures for FDR control. The original Benjamini and Hochberg [3] procedure is a stepup procedure, using the ordered p-values, while the procedures proposed in Storey [61] and others are single-step procedures in which all p-values less than a criterion t are rejected. Based on the notation in Table 1, Storey, Taylor and Siegmund [63] define the empirical processes

$$\begin{aligned} V(t) &= \#(null\ \ p_i : p_i \le t) \\ S(t) &= \#(alternative\ \ p_i : p_i \le t) \\ R(t) &= V(t) + S(t) = \#(p_i : p_i \le t). \end{aligned} \tag{5.1}$$

They use empirical process theory to prove both finite-sample and asymptotic control of FDR for the Benjamini and Hochberg [3] procedure and the most conservative Storey [61] procedure, and also for new proposed procedures that involve estimation of π_0 under both independence and some forms of positive dependence.

Benjamini, Krieger, and Yekutieli [5] develop two-stage and multistage adaptive methods, and study the two-stage method analytically. That method provides an

estimate of π_0 at the first stage and takes the uncertainty about the estimate into account in modifying the second stage. It is proved to guarantee FDR control at the specified level. Based on extensive simulation results the methods proposed in Storey, Taylor and Siegmund [63] perform best when test statistics are independent, while the Benjamini, Krieger and Yekutieli [5] two-stage adaptive method appears to be the only proposed method (to this time) based on estimating m_0 that controls the FDR under the conditions of high positive dependence that are sufficient for FDR control using the original Benjamini and Hochberg [3] FDR procedure.

5.2. Resampling methods

In general, under any criterion, if appropriate aspects of joint distributions of test statistics were known (e.g. their covariance matrices), procedures based on those distributions could achieve greater power with the same error control than procedures ensuring error control but not based on such knowledge. Resampling methods are being intensively investigated from this point of view. Permutation methods, when applicable, can provide exact error control under criterion (iii) (Westfall and Young [66]) and some bootstrap methods have been shown to provide asymptotic error control, with the possibility of finding asymptotically optimal methods under such control (Dudoit, van der Laan and Pollard [16], Korn, Troendle, McShane and Simon [34], Lehmann and Romano [40], van der Laan, Dudoit and Pollard [65], Pollard and van der Laan [48], Romano and Wolf [52]).

Since the asymptotic methods are based on the assumption of large sample sizes relative to the number of tests, it is an open question how well they apply in cases of massive numbers of hypotheses in which the sample size is considerably smaller than m, and therefore how relevant any asymptotic optimal properties would be in these contexts. Some recent references in this area are Troendle, Korn, and McShane [64], Bang and Young [2]), and Muller et al [46], the latter in the context of a Bayesian decision-theoretic model.

5.3. Empirical Bayes procedures

The Bayes procedure of Berger and Sellke [8] referred to in the section on a single parameter, testing (2.1) or (2.2), requires an assumption of the prior probability that the hypothesis is true. With large numbers of hypotheses, the procedure can be replaced by empirical Bayes procedures based on estimates of this prior probability by estimating the proportion of true hypotheses. These, as well as estimates of other aspects of the prior distributions of the test statistics corresponding to true and false hypotheses, are obtained in many cases by resampling methods. Some of the references in the two preceding subsections are relevant here; see also Efron [18], and Efron and Tibshirani [21]; the latter compares an empirical Bayes method with the FDR-controlling method of Benjamini and Hochberg [3]. Kendziorski et al [33] use an empirical Bayes hierarchical mixture model with stronger parametric assumptions, enabling them to estimate the relevant parameters by log likelihood methods rather than resampling.

For an unusual approach to the choice of null hypothesis, see Efron [19], who suggests that an alternative null hypothesis distribution, based on an empirically-determined "central" value, should be used in some situations to determine "interesting" – as opposed to "significant" – results. For a novel combination of empirical Bayes hypothesis testing and estimation, related to Duncan's [17] emphasis on the magnitude of null hypothesis departure, see Efron [20].

6. Summary

Both one-sided and two-sided tests referring to a single parameter are considered. A two-sided test referring to a single parameter becomes multiple inference when the hypothesis that the parameter θ is equal to a fixed value θ_0 is reformulated as the directional hypothesis-pair (i) $\theta \leq \theta_0$ and (ii) $\theta \geq \theta_0$, a more appropriate formulation when directional inference is desired. In the first part of the paper, it is shown that optimality results in the case of a single nondirectional hypothesis can be diametrically opposite to directional optimality results. In fact, a procedure that is uniformly most powerful under the nondirectional formulation can be uniformly least powerful under the directional hypothesis-pair formulation, and vice versa.

The second part of the paper sketches the many different formulations of error rates, power, and classes of procedures when there are multiple parameters. Some of these have been utilized for many years, and some are relatively new, stimulated by the increasing number of areas in which massive sets of hypotheses are being tested. There are relatively few optimality results in multiple comparisons in general, and still fewer when these newer criteria are utilized, so there is great potential for optimality research in this area. Many existing optimality results are described in Hochberg and Tamhane [28] and Shaffer [58]). These are sketched briefly here, and some further relevant references are provided.

References

[1] Bahadur, R. R. (1952). A property of the t-statistic. *Sankhya* **12**, 79–88.

[2] Bang, S.J. and Young, S.S. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics* **6**, 157–169.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate. *J. Roy. Stat. Soc. Ser. B* **57**, 289–300.

[4] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.* **25**, 60–83.

[5] Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2005). Adaptive linear step-up procedures that control the false discovery rate. Research Paper 01-03, Department of Statistics and Operations Research, Tel Aviv University. (Available at `http://www.math.tau.ac.il/ yekutiel/papers/bkymarch9.pdf`.)

[6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate under dependency. *Ann. Statist.* **29**, 1165–1188.

[7] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Second edition). Springer, New York.

[8] Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82**, 112–139.

[9] Black, M.A. (2004). A note on the adaptive control of false discovery rates. *J. Roy. Statist. Soc. Ser. B* **66**, 297–304.

[10] Bohrer, R. (1979). Multiple three-decision rules for parametric signs. *J. Amer. Statist. Assoc.* **74**, 432–437.

[11] Braver, S. L. (1975). On splitting the tails unequally: A new perspective on one- versus two-tailed tests. *Educ. Psychol. Meas.* **35**, 283–301.

[12] Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.

[13] Cheng, C., Pounds, S., Boyett, J. M., Pei, D., Kuo, M-L. and Roussel, M. F. (2004). Statistical significance threshold criteria for analysis of microarray gene expression data. *Stat. Appl. Genet. Mol. Biol.* **3**, 1, Article 36, http://www.bepress.com/sagmb/vol3/iss1/art36.

[14] Cohen, A. and Sackrowitz, H. B. (2005a). Decision theory results for one-sided multiple comparison procedures. *Ann. Statist.* **33**, 126–144.

[15] Cohen, A. and Sackrowitz, H. B. (2005b). Characterization of Bayes procedures for multiple endpoint problems and inadmissibility of the step-up procedure. *Ann. Statist.* **33**, 145–158.

[16] Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. Part I. SIngle-step procedures for control of general Type I error rates. *Stat. Appl. Genet. Mol. Biol.* **1**, Article 13.

[17] Duncan, D. B. (1961). Bayes rules for a common multiple comparison problem and related Student-t problems. *Ann. Math. Statist.* **32**, 1013–1033.

[18] Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Ann. Statist.* **31**, 366–378.

[19] Efron, B. (2004a). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96–104.

[20] Efron, B. (2004b). Selection and estimation for large-scale simultaneous inference. (Can be downloaded from http://www-stat.stanford.edu/ brad/ – click on "papers and software".)

[21] Efron, B. and Tibshirani R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70–86.

[22] Eklund, G. and Seeger, P. (1965). Massignifikansanalys. *Statistisk Tidskrift, 3rd series* **4**, 355–365.

[23] Finner, H. (1990). Some new inequalities for the range distribution, with application to the determination of optimum significance levels of multiple range tests. *J. Amer. Statist. Assoc.* **85**, 191–194.

[24] Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Ann. Statist.* **27**, 274–289.

[25] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499–517.

[26] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–1061.

[27] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

[28] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[29] Holland, B. and Cheung, S. H. (2002). Familywise robustness criteria for multiple-comparison procedures. *J. Roy. Statist. Soc. Ser. B* **64**, 63–77.

[30] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

[31] Hommel, G. and Hoffmann, T. (1987). Controlled uncertainty. In *Medizinische Informatik und Statistik*, P. Bauer, G. Hommel, and E. Sonnemann (Eds.). Springer-Verlag, Berlin.

[32] Kaiser, H. F. (1960). Directional statistical decisions. *Psychol. Rev.* **67**, 160–167.

[33] Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22**, 3899–3914.

[34] Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124**, 379–398.

[35] Lehmann, E. L. (1950). Some principles of the theory of testing hypotheses, *Ann. Math. Statist.* **21**, 1–26.

[36] Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statist.* **23**, 541–552.

[37] Lehmann, E. L. (1957a). A theory of some multiple decision problems, Part I. *Ann. Math. Statist.* **28**, 1–25.

[38] Lehmann, E. L. (1957b). A theory of some multiple decision problems, Part II. *Ann. Math. Statist.* **28**, 547–572.

[39] Lehmann, E. L. (2006). On likelihood ratio tests. This volume.

[40] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138–1154.

[41] Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33**, 1084–1108.

[42] Lehmann, E. L. and Shaffer, J. P. (1979). Optimum significance levels for multistage comparison procedures. *Ann. Statist.* **7**, 27–45.

[43] Lewis, C. and Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *J. Statist. Plann. Inference* **125**, 49–58.

[44] Mantel, N. (1983). Ordered alternatives and the 1 1/2-tail test. *Amer. Statist.* **37**, 225–228.

[45] Meinshausen, N. and Rice, J. (2004). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34**, 373–393.

[46] Muller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99**, 990–1001.

[47] Owen, A.B. (2005). Variance of the number of false discoveries. *J. Roy. Statist. Soc. Series B* **67**, 411–426.

[48] Pollard, K. S. and van der Laan, M. J. (2003). Resampling-based multiple testing: Asymptotic control of Type I error and applications to gene expression data. U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 121.

[49] Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *J. Amer. Statist. Assoc.* **73**, 479–485.

[50] Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

[51] Romano, J. P. and Shaikh, A. M. (2004). On control of the false discovery proportion. Tech. Report 2004-31, Dept. of Statistics, Stanford University.

[52] Romano, J. P. and Wolf, M. (2004). Exact and approximate stepdown methods for multiple hypothesis testing. *Ann. Statist.* **100**, 94–108.

[53] Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics* **10**, 586–593.

[54] Shaffer, J. P. (1974). Bidirectional unbiased procedures. *J. Amer. Statist. Assoc.* **69**, 437–439.

[55] Shaffer, J. P. (1981). Complexity: An interpretability criterion for multiple comparisons. *J. Amer. Statist. Assoc.* **76**, 395–401.

[56] Shaffer, J. P. (1999). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *J. Statist. Plann. Inference* **82**, 197–213.

[57] Shaffer, J. P. (2002). Multiplicity, directional (Type III) errors, and the null hypothesis. *Psychol. Meth.* **7**, 356–369.

[58] Shaffer, J. P. (2004). Optimality results in multiple hypothesis testing. *The First Erich L. Lehmann Symposium – Optimality, Lecture Notes–Monograph Series*, Vol. **44**. Institute of Mathematical Statistics, 11–35.

[59] Sorić, B. (1989). Statistical "discoveries" and effect size estimation. *J. Amer. Statist. Assoc.* **84**, 608–610.

[60] Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43**, 398–411.

[61] Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.

[62] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013–2035.

[63] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. Ser. B* **66**, 187–205.

[64] Troendle, J. F., Korn, E. L. and McShane, L. M. (2004). An example of slow convergence of the bootstrap in high dimensions. *Amer. Statist.* **58**, 25–29.

[65] van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 15.

[66] Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Wiley, New York.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 33–50
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000383

On stepdown control of the false discovery proportion

Joseph P. Romano[1] **and Azeem M. Shaikh**[2]

Stanford University

Abstract: Consider the problem of testing multiple null hypotheses. A classical approach to dealing with the multiplicity problem is to restrict attention to procedures that control the familywise error rate ($FWER$), the probability of even one false rejection. However, if s is large, control of the $FWER$ is so stringent that the ability of a procedure which controls the $FWER$ to detect false null hypotheses is limited. Consequently, it is desirable to consider other measures of error control. We will consider methods based on control of the false discovery proportion (FDP) defined by the number of false rejections divided by the total number of rejections (defined to be 0 if there are no rejections). The false discovery rate proposed by Benjamini and Hochberg (1995) controls $E(FDP)$. Here, we construct methods such that, for any γ and α, $P\{FDP > \gamma\} \leq \alpha$. Based on p-values of individual tests, we consider stepdown procedures that control the FDP, without imposing dependence assumptions on the joint distribution of the p-values. A greatly improved version of a method given in Lehmann and Romano [10] is derived and generalized to provide a means by which any sequence of nondecreasing constants can be rescaled to ensure control of the FDP. We also provide a stepdown procedure that controls the FDR under a dependence assumption.

1. Introduction

In this article, we consider the problem of simultaneously testing a finite number of null hypotheses H_i $(i = 1, \ldots, s)$. We shall assume that tests based on p-values $\hat{p}_1, \ldots, \hat{p}_s$ are available for the individual hypotheses and the problem is how to combine them into a simultaneous test procedure.

A classical approach to dealing with the multiplicity problem is to restrict attention to procedures that control the familywise error rate ($FWER$), which is the probability of one or more false rejections. In addition to error control, one must also consider the ability of a procedure to detect departures from the null hypotheses when they do occur. When the number of tests s is large, control of the $FWER$ is so stringent that individual departures from the hypothesis have little chance of being detected. Consequently, alternative measures of error control have been considered which control false rejections less severely and therefore provide better ability to detect false null hypotheses.

Hommel and Hoffman [8] and Lehmann and Romano [10] considered the k-$FWER$, the probability of rejecting at least k true null hypotheses. Such an error rate with $k > 1$ is appropriate when one is willing to tolerate one or more false rejections, provided the number of false rejections is controlled. They derived single

[1]Department of Statistics, Stanford University, Stanford, CA 94305-4065, e-mail: romano@stat.stanford.edu

[2]Department of Economics, Stanford University, Stanford, CA 94305-6072, e-mail: ashaikh@stanford.edu

AMS 2000 subject classifications: 62J15.

Keywords and phrases: familywise error rate, multiple testing, p-value, stepdown procedure.

step and stepdown methods that guarantee that the k-$FWER$ is bounded above by α. Evidently, taking $k = 1$ reduces to the usual $FWER$. Lehmann and Romano [10] also considered control of the *false discovery proportion* (FDP), defined as the total number of false rejections divided by the total number of rejections (and equal to 0 if there are no rejections). Given a user specified value $\gamma \in (0, 1)$, control of the FDP means we wish to ensure that $P\{FDP > \gamma\}$ is bounded above by α. Control of the *false discovery rate* (FDR) demands that $E(FDP)$ is bounded above by α. Setting $\gamma = 0$ reduces to the usual $FWER$.

Recently, many methods have been proposed which control error rates that are less stringent than the $FWER$. For example, Genovese and Wasserman [4] study asymptotic procedures that control the FDP (and the FDR) in the framework of a random effects mixture model. These ideas are extended in Perone Pacifico, Genovese, Verdinelli and Wasserman [11], where in the context of random fields, the number of null hypotheses is uncountable. Korn, Troendle, McShane and Simon [9] provide methods that control both the k-$FWER$ and FDP; they provide some justification for their methods, but they are limited to a multivariate permutation model. Alternative methods of control of the k-$FWER$ and FDP are given in van der Laan, Dudoit and Pollard [17]. The methods proposed in Lehmann and Romano [10] are not asymptotic and hold under either mild or no assumptions, as long as p-values are available for testing each individual hypothesis. In this article, we offer an improved method that controls the FDP under no dependence assumptions of the p-values. The method is seen to be a considerable improvement in that the critical values of the new procedure can be increased by typically 50 percent over the earlier procedure, while still maintaining control of the FDP. The argument used to establish the improvement is then generalized to provide a means by which any nondecreasing sequence of constants can be rescaled (by a factor that depends on s, γ, and α) so as to ensure control of the FDP.

It is of interest to compare control of the FDP with control of the FDR, and some obvious connections between methods that control the FDP in the sense that

$$P\{FDP > \gamma\} \leq \alpha$$

and methods that control its expected value, the FDR, can be made. Indeed, for any random variable X on $[0, 1]$, we have

$$\begin{aligned} E(X) &= E(X|X \leq \gamma)P\{X \leq \gamma\} + E(X|X > \gamma)P\{X > \gamma\} \\ &\leq \gamma P\{X \leq \gamma\} + P\{X > \gamma\} \ , \end{aligned}$$

which leads to

$$\frac{E(X) - \gamma}{1 - \gamma} \leq P\{X > \gamma\} \leq \frac{E(X)}{\gamma}, \tag{1.1}$$

with the last inequality just Markov's inequality. Applying this to $X = FDP$, we see that, if a method controls the FDR at level q, then it controls the FDP in the sense $P\{FDP > \gamma\} \leq q/\gamma$. Obviously, this is very crude because if q and γ are both small, the ratio can be quite large. The first inequality in (1.1) says that if the FDP is controlled in the sense of (3.3), then the FDR is controlled at level $\alpha(1 - \gamma) + \gamma$, which is $\geq \alpha$ but typically only slightly. Therefore, in principle, a method that controls the FDP in the sense of (3.3) can be used to control the FDR and vice versa.

The paper is organized as follows. In Section 2, we describe our terminology and the general class of stepdown procedures that are examined. Results from

Lehmann and Romano [10] are summarized to motivate our choice of critical values. Control of the FDP is then considered in Section 3. The main result is presented in Theorem 3.4 and generalized in Theorem 3.5. In Section 4, we prove that a certain stepdown procedure controls the FDR under a dependence assumption.

2. A class of stepdown procedures

A formal description of our setup is as follows. Suppose data X is available from some model $P \in \Omega$. A general hypothesis H can be viewed as a subset ω of Ω. For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, let $I(P)$ denote the set of true null hypotheses when P is the true probability distribution; that is, $i \in I(P)$ if and only if $P \in \omega_i$.

We assume that p-values $\hat{p}_1, \ldots, \hat{p}_s$ are available for testing $H_1, \ldots, H_s$. Specifically, we mean that $\hat{p}_i$ must satisfy

$$P\{\hat{p}_i \leq u\} \leq u \quad \text{for any } u \in (0,1) \quad \text{and any } P \in \omega_i, \tag{2.1}$$

Note that we do not require $\hat{p}_i$ to be uniformly distributed on $(0,1)$ if H_i is true, in order to accomodate discrete situations.

In general, a p-value $\hat{p}_i$ will satisfy (2.1) if it is obtained from a nested set of rejection regions. In other words, suppose $S_i(\alpha)$ is a rejection region for testing H_i; that is,

$$P\{X \in S_i(\alpha)\} \leq \alpha \quad \text{for all } 0 < \alpha < 1,\ P \in \omega_i \tag{2.2}$$

and

$$S_i(\alpha) \subset S_i(\alpha') \quad \text{whenever } \alpha < \alpha'. \tag{2.3}$$

Then, the p-value $\hat{p}_i$ defined by

$$\hat{p}_i = \hat{p}_i(X) = \inf\{\alpha :\ X \in S_i(\alpha)\}. \tag{2.4}$$

satisfies (2.1).

In this article, we will consider the following class of *stepdown* procedures. Let

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s \tag{2.5}$$

be constants, and let $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(s)}$ denote the ordered p-values. If $\hat{p}_{(1)} > \alpha_1$, reject no null hypotheses. Otherwise,

$$\hat{p}_{(1)} \leq \alpha_1, \ldots, \hat{p}_{(r)} \leq \alpha_r, \tag{2.6}$$

and hypotheses $H_{(1)}, \ldots, H_{(r)}$ are rejected, where the largest r satisfying (2.6) is used. That is, a stepdown procedure starts with the most significant p-value and continues rejecting hypotheses as long as their corresponding p-values are small.

The Holm [6] procedure uses $\alpha_i = \alpha/(s-i+1)$ and controls the $FWER$ at level α under no assumptions on the joint distribution of the p-values. Lehmann and Romano [10] generalized the Holm procedure to control the k-$FWER$. Specifically, consider the stepdown procedure described in (2.6), where we now take

$$\alpha_i = \begin{cases} \frac{k\alpha}{s} & i \leq k \\ \frac{k\alpha}{s+k-i} & i > k \end{cases} \tag{2.7}$$

Of course, the α_i depend on s and k, but we suppress this dependence in the notation.

Theorem 2.1 (Hommel and Hoffman [8] and Lehmann and Romano [10]). *For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, suppose $\hat{p}_i$ satisfies* (2.1)*. The stepdown procedure described in* (2.6) *with α_i given by* (2.7) *controls the k-FWER; that is,*

$$P\{\text{reject at least } k \text{ hypotheses } H_i \text{ with } i \in I(P)\} \leq \alpha \quad \text{for all } P \,. \tag{2.8}$$

Moreover, one cannot increase even one of the constants α_i (for $i \geq k$) without violating control of the k-FWER. Specifically, for $i \geq k$, there exists a joint distribution of the p-values for which

$$P\{\hat{p}_{(1)} \leq \alpha_1, \hat{p}_{(2)} \leq \alpha_2, \ldots, \hat{p}_{(i-1)} \leq \alpha_{i-1}, \hat{p}_{(i)} \leq \alpha_i\} = \alpha. \tag{2.9}$$

Remark 2.1. Evidently, one can always reject the hypotheses corresponding to the smallest $k-1$ p-values without violating control of the k-$FWER$. However, it seems counterintuitive to consider a stepdown procedure whose corresponding α_i are not monotone nondecreasing. In addition, automatic rejection of $k-1$ hypotheses, regardless of the data, appears at the very least a little too optimistic. To ensure monotonicity, our stepdown procedure uses $\alpha_i = k\alpha/s$. Even if we were to adopt the more optimistic strategy of always rejecting the hypotheses corresponding to the $k-1$ smallest p-values, we could still only reject k or more hypotheses if $\hat{p}_{(k)} \leq k\alpha/s$, which is also true for the specific procedure of Theorem 2.1.

3. Control of the false discovery proportion

The number k of false rejections that one is willing to tolerate will often increase with the number of hypotheses rejected. So, it might be of interest to control not the number of false rejections (or sometimes called false discoveries) but the proportion of false discoveries. Specifically, let the *false discovery proportion* (FDP) be defined by

$$FDP = \begin{cases} \frac{\text{Number of false rejections}}{\text{Total number of rejections}} & \text{if the denominator is } > 0 \\ 0 & \text{if there are no rejections} \end{cases} \tag{3.1}$$

Thus FDP is the proportion of rejected hypotheses that are rejected erroneously. When none of the hypotheses are rejected, both numerator and denominator of that proportion are 0; since in particular there are no false rejections, the FDP is then defined to be 0.

Benjamini and Hochberg [1] proposed to replace control of the $FWER$ by control of the *false discovery rate* (FDR), defined as

$$FDR = E(FDP). \tag{3.2}$$

The FDR has gained wide acceptance in both theory and practice, largely because Benjamini and Hochberg proposed a simple stepup procedure to control the FDR. Unlike control of the k-$FWER$, however, their procedure is not valid without assumptions on the dependence structure of the p-values. Their original paper assumed the very strong assumption of independence of p-values, but this has been weakened to include certain types of dependence; see Benjamini and Yekutieli [3]. In any case, control of the FDR does not prohibit the FDP from varying, even if its average value is bounded. Instead, we consider an alternative measure of control that guarantees the FDP is bounded, at least with prescribed probability. That is, for a given γ and α in $(0, 1)$, we require

$$P\{FDP > \gamma\} \leq \alpha. \tag{3.3}$$

To develop a stepdown procedure satisfying (3.3), let f denote the number of false rejections. At step i, having rejected $i-1$ hypotheses, we want to guarantee $f/i \le \gamma$, i.e. $f \le \lfloor \gamma i \rfloor$, where $\lfloor x \rfloor$ is the greatest integer $\le x$. So, if $k = \lfloor \gamma i \rfloor + 1$, then $f \ge k$ should have probability no greater than α; that is, we must control the number of false rejections to be $\le k$. Therefore, we use the stepdown constant α_i with this choice of k (which now depends on i); that is,

$$\alpha_i = \frac{(\lfloor \gamma i \rfloor + 1)\alpha}{s + \lfloor \gamma i \rfloor + 1 - i}. \tag{3.4}$$

Lehmann and Romano [10] give two results that show the stepdown procedure with this choice of α_i satisfies (3.3). Unfortunately, some joint dependence assumption on the p-values is required. As before, $\hat{p}_1, \ldots, \hat{p}_s$ denotes the p-values of the individual tests. Also, let $\hat{q}_1, \ldots, \hat{q}_{|I|}$ denote the p-values corresponding to the $|I| = |I(P)|$ true null hypotheses. So $q_i = p_{j_i}$, where $j_1, \ldots, j_{|I|}$ correspond to the indices of the true null hypotheses. Also, let $\hat{r}_1, \ldots, \hat{r}_{s-|I|}$ denote the p-values of the false null hypotheses. Consider the following condition: for any $i = 1, \ldots, |I|$,

$$P\{\hat{q}_i \le u | \hat{r}_1, \ldots, \hat{r}_{s-|I|}\} \le u; \tag{3.5}$$

that is, conditional on the observed p-values of the false null hypotheses, a p-value corresponding to a true null hypothesis is (conditionally) dominated by the uniform distribution, as it is unconditionally in the sense of (2.1). No assumption is made regarding the unconditional (or conditional) dependence structure of the true p-values, nor is there made any explicit assumption regarding the joint structure of the p-values corresponding to false hypotheses, other than the basic assumption (3.5). So, for example, if the p-values corresponding to true null hypotheses are independent of the false ones, but have arbitrary joint dependence within the group of true null hypotheses, the above assumption holds.

Theorem 3.1 (Lehmann and Romano [10]). *Assume the condition* (3.5). *Then, the stepdown procedure with α_i given by* (3.4) *controls the FDP in the sense of* (3.3).

Lehmann and Romano [10] also show the same stepdown procedure controls the FDP in the sense of (3.3) under an alternative assumption involving the joint distribution of the p-values corresponding to true null hypotheses. We follow their approach here.

Theorem 3.2 (Lehmann and Romano [10]). *Consider testing s null hypotheses, with $|I|$ of them true. Let $\hat{q}_{(1)} \le \cdots \le \hat{q}_{(|I|)}$ denote the ordered p-values for the true hypotheses. Set $M = \min(\lfloor \gamma s \rfloor + 1, |I|)$.*
(i) *For the stepdown procedure with α_i given by* (3.4),

$$P\{FDP > \gamma\} \le P\{\bigcup_{i=1}^{M} \{\hat{q}_{(i)} \le \frac{i\alpha}{|I|}\}\}. \tag{3.6}$$

(ii) *Therefore, if the joint distribution of the p-values of the true null hypotheses satisfy Simes inequality; that is,*

$$P\{\{\hat{q}_{(1)} \le \frac{\alpha}{|I|}\} \bigcup \{\hat{q}_{(2)} \le \frac{2\alpha}{|I|}\} \bigcup \ldots \bigcup \{\hat{q}_{(|I|)} \le \alpha\}\} \le \alpha,$$

then $P\{FDP > \gamma\} \le \alpha$.

Simes inequality is known to hold for many joint distributions of positively dependent variables. For example, Sarkar and Chang [15] and Sarkar [13] have shown that the Simes inequality holds for the family of distributions which is characterized by the multivariate positive of order two condition, as well as some other important distributions.

However, we will argue that the stepdown procedure with α_i given by (3.4) does *not* control the FDP in general. First, we need to recall Lemma 3.1 of Lehmann and Romano [10], stated next for convenience (since we use it later as well). It is related to Lemma 2.1 of Sarkar [13].

Lemma 3.1. *Suppose $\hat{p}_1, \ldots, \hat{p}_t$ are p-values in the sense that $P\{\hat{p}_i \leq u\} \leq u$ for all i and u in $(0,1)$. Let their ordered values be $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(t)}$. Let $0 = \beta_0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_m \leq 1$ for some $m \leq t$.*
(i) *Then,*

$$P\{\{\hat{p}_{(1)} \leq \beta_1\} \bigcup \{\hat{p}_{(2)} \leq \beta_2\} \bigcup \cdots \bigcup \{\hat{p}_{(m)} \leq \beta_m\}\} \leq t \sum_{i=1}^{m} (\beta_i - \beta_{i-1})/i. \tag{3.7}$$

(ii) *As long as the right side of* (3.7) *is ≤ 1, the bound is sharp in the sense that there exists a joint distribution for the p-values for which the inequality is an equality.*

The following calculation illustrates the fact that the stepdown procedure with α_i given by (3.4) does *not* control the FDP in general.

Example 3.1. Suppose $s = 100$, $\gamma = 0.1$ and $|I| = 90$. Construct a joint distribution of p-values as follows. Let $\hat{q}_{(1)} \leq \cdots \leq \hat{q}_{(90)}$ denote the ordered p-values corresponding to the true null hypotheses. Suppose these 90 p-values have some joint distribution (specified below). Then, we construct the p-values corresponding to the 10 false null hypotheses conditional on the 90 p-values. First, let 8 of the p-values corresponding to false null hypotheses be identically zero (or at least less than $\alpha/100$). If $\hat{q}_{(1)} \leq \alpha/92$, let the 2 remaining p-values corresponding to false null hypotheses be identically 1; otherwise, if $\hat{q}_{(1)} > \alpha/92$, let the 2 remaining p-values also be equal to zero. For this construction, $FDP > \gamma$ if $\hat{q}_{(1)} \leq \alpha/92$ or $\hat{q}_{(2)} \leq 2\alpha/91$. The value of

$$P\{\hat{q}_{(1)} \leq \frac{\alpha}{92} \bigcup \hat{q}_{(2)} \leq \frac{2\alpha}{91}\}$$

can be bounded by Lemma 3.1. The lemma bounds this expression by

$$90\left(\frac{\alpha}{92} + \frac{\frac{2\alpha}{91} - \frac{\alpha}{92}}{2}\right) \approx 1.48\alpha > \alpha.$$

Moreover, Lemma 3.1 gives a joint distribution for the 90 p-values corresponding to true null hypotheses for which this calculation is an equality.

Since one may not wish to assume any dependence conditions on the p-values, Lehmann and Romano [10] use Theorem 3.2 to derive a method that controls the FDP without any dependence assumptions. One simply needs to bound the right hand side of (3.6). In fact, Hommel [7] has shown that

$$P\{\bigcup_{i=1}^{|I|} \{\hat{q}_{(i)} \leq \frac{i\alpha}{|I|}\}\} \leq \alpha \sum_{i=1}^{|I|} \frac{1}{i}.$$

This suggests we replace α by $\alpha(\sum_{i=1}^{|I|}(1/i))^{-1}$. But of course $|I|$ is unknown. So one possibility is to bound $|I|$ by s which then results in replacing α by α/C_s, where

$$C_j = \sum_{i=1}^{j} \frac{1}{i}. \tag{3.8}$$

Clearly, changing α in this way is much too conservative and results in a much less powerful method. However, notice in (3.6) that we really only need to bound the union over $M \leq \lfloor \gamma s \rfloor + 1$ events. This leads to the following result.

Theorem 3.3 (Lehmann and Romano [10]). *For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, suppose $\hat{p}_i$ satisfies* (2.1)*. Consider the stepdown procedure with constants $\alpha_i' = \alpha_i / C_{\lfloor \gamma s \rfloor + 1}$, where α_i is given by* (3.4) *and C_j defined by* (3.8)*. Then, $P\{FDP > \gamma\} \leq \alpha$.*

The next goal is to improve upon Theorem 3.3. In the definition of α_i', α_i is divided by $C_{\lfloor \gamma s \rfloor + 1}$. Instead, we will construct a stepdown procedure with constants $\alpha_i'' = \alpha_i / D$, where $D = D(\gamma, \alpha, s)$ is much smaller than $C_{\lfloor \gamma s \rfloor + 1}$. This procedure will also control the FDP but, since the critical values α_i'' are uniformly bigger than the α_i', the new procedure can reject more hypotheses and hence is more powerful.

To this end, define

$$\beta_m = \frac{m}{\max\{s + m - \lceil \frac{m}{\gamma} \rceil + 1, |I|\}} \qquad m = 1, \ldots, \lfloor \gamma s \rfloor \tag{3.9}$$

and

$$\beta_{\lfloor \gamma s \rfloor + 1} = \frac{\lfloor \gamma s \rfloor + 1}{|I|}. \tag{3.10}$$

where $\lceil x \rceil$ is the least integer $\geq x$.

Next, let

$$N = N(\gamma, s, |I|) = \min\{\lfloor \gamma s \rfloor + 1, |I|, \lfloor \gamma(\frac{s - |I|}{1 - \gamma} + 1) \rfloor + 1\}. \tag{3.11}$$

Then, let $\beta_0 = 0$ and set

$$S = S(\gamma, s, |I|) = |I| \sum_{i=1}^{N} \frac{\beta_i - \beta_{i-1}}{i}. \tag{3.12}$$

Finally, let

$$D = D(\gamma, s) = \max_{|I|} S(\gamma, s, |I|). \tag{3.13}$$

Theorem 3.4. *For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, suppose $\hat{p}_i$ satisfies* (2.1)*. Consider the stepdown procedure with constants $\alpha_i'' = \alpha_i / D(\gamma, s)$, where α_i is given by* (3.4) *and $D(\gamma, s)$ is defined by* (3.13)*. Then, $P\{FDP > \gamma\} \leq \alpha$.*

Proof. Let $\alpha'' = \alpha / D$. Denote by

$$\hat{q}_{(1)} \leq \cdots \leq \hat{q}_{(|I|)}$$

the ordered p-values corresponding only to true null hypotheses. Let j be the smallest (random) index where the FDP exceeds γ for the first time at step j; that is, the

number of false rejections out of the first $j-1$ rejections divided by j exceeds γ for the first time at j. Denote by $m > 0$ the unique integer satisfying $m - 1 \leq \gamma j < m$. Then, at step j, it must be the case that m true null hypotheses have been rejected. Hence,

$$\hat{q}_{(m)} \leq \alpha''_j = \frac{m\alpha''}{s+m-j}.$$

Note that the number of true hypotheses $|I|$ satisfies

$$|I| \leq s + m - j.$$

Further note that $\gamma j < m$ implies that

$$j \leq \lceil \frac{m}{\gamma} \rceil - 1. \tag{3.14}$$

Hence, α''_j is bounded above by β_m defined by (3.9) whenever $m - 1 \leq \gamma j < m$. Note that, when $m = \lfloor \gamma s \rfloor + 1$, we bound α''_j by using $j \leq s$ rather than (3.14).

The possible values of m that must be considered can be bounded. First of all, $j \leq s$ implies that $m \leq \lfloor \gamma s \rfloor + 1$. Likewise, it must be the case that $m \leq |I|$. Finally, note that $j > \frac{s-|I|}{1-\gamma}$ implies that $FDP > \gamma$. To see this, observe that

$$\frac{s-|I|}{1-\gamma} = (s - |I|) + \frac{\gamma}{1-\gamma}(s - |I|),$$

so at such a step j, it must be the case that

$$t > \frac{\gamma}{1-\gamma}(s - |I|)$$

true null hypotheses have been rejected. If we denote by $f = j - t$ the number of false null hypotheses that have been rejected at step j, it follows that

$$t > \frac{\gamma}{1-\gamma} f,$$

which in turn implies that

$$FDP = \frac{t}{t+f} > \gamma.$$

Hence, for j to satisfy the above assumption of minimality, it must be the case that

$$j - 1 \leq \frac{s-|I|}{1-\gamma},$$

from which it follows that we must also have

$$m \leq \lfloor \gamma(\frac{s-|I|}{1-\gamma} + 1) \rfloor + 1.$$

Therefore, with N defined in (3.11) and j defined as above, we have that

$$P\{FDP > \gamma\} \leq \sum_{m=1}^{N} P\left\{\{\hat{q}_{(m)} \leq \alpha''_j\} \bigcap \{m - 1 \leq \gamma j < m\}\right\}$$

$$\leq \sum_{m=1}^{N} P\left\{\hat{q}_{(m)} \leq \alpha'' \beta_m\} \bigcap \{m - 1 \leq \gamma j < m\}\right\}$$

$$\leq \sum_{m=1}^{N} P\left\{\bigcup_{i=1}^{N}\{\hat{q}_{(i)} \leq \alpha''\beta_i\}\bigcap\{m-1 \leq \gamma j < m\}\right\}$$

$$\leq P\left\{\bigcup_{i=1}^{N}\{\hat{q}_{(i)} \leq \alpha''\beta_i\right\}.$$

Note that $\beta_m \leq \beta_{m+1}$. To see this, observed that the expression $m+s-\lceil\frac{m}{\gamma}\rceil+1$ is monotone nonincreasing in m, and so the denominator of β_m, $\max\{m+s-\lceil\frac{m}{\gamma}\rceil+1, |I|\}$, is monotone nonincreasing in m as well. Also observe that $\beta_m \leq m/|I| \leq 1$ whenever $m \leq N$. We can therefore apply Lemma 3.1 to conclude that

$$P\{FDP > \gamma\} \leq \alpha''|I|\sum_{i=1}^{N}\frac{\beta_i-\beta_{i-1}}{i}$$

$$= \frac{\alpha|I|}{D}\sum_{i=1}^{N}\frac{\beta_i-\beta_{i-1}}{i} = \frac{\alpha S}{D} \leq \alpha,$$

where S and D are defined in (3.12) and (3.13), respectively. □

It is important to note that by construction the quantity $D(\gamma, s)$, which is defined to be the maximum over the possible values of $|I|$ of the quantity $S(\gamma, s, |I|)$, does not depend on the unknown number of true hypotheses. Indeed, if the number of true hypotheses, $|I|$, were known, then the smaller quantity $S(\gamma, s, |I|)$ could be used in place of $D(\gamma, s)$.

Unfortunately, a convenient formula is not available for $D(\gamma, s)$, though it is simple to program its evaluation. For example, if $s = 100$ and $\gamma = 0.1$, then $D = 2.0385$. In contrast, the constant $C_{\lfloor\gamma s\rfloor+1} = C_{11} = 3.0199$. In this case, the value of $|I|$ that maximizes S to yield D is 55. Below, in Table 1 we evaluate $D(\gamma, s)$ and $C_{\lfloor\gamma s\rfloor+1}$ for several different values of γ and s. We also compute the ratio of $C_{\lfloor\gamma s\rfloor+1}$ to $D(\gamma, s)$, from which it is possible to see the magnitude of the improvement of the Theorem 3.4 over Theorem 3.3: the constants of Theorem 3.4 are generally about 50 percent larger than those of Theorem 3.3.

Remark 3.1. The following crude argument suggests that, for critical values of the form $d\alpha_i$ for some constant d, the value of $d = D^{-1}(\gamma, s)$ is very nearly the largest possible constant one can use and still maintan control of the FDP. Consider the case where $s = 1000$ and $\gamma = .1$. In this instance, the value of $|I|$ that maximizes S is 712, yielding $N = 33$ and $D = 3.4179$. Suppose that $|I| = 712$ and construct the joint distribution of the 288 p-values corresponding to false hypotheses as follows: For $1 \leq i \leq 28$, if $\hat{q}_{(i)} \leq \alpha\beta_i$ and $\hat{q}_{(j)} > \alpha\beta_j$ for all $j < i$, then let $\lceil\frac{i}{\gamma}\rceil - 1$ of the false p-values be 0 and set the remainder equal to 1. Let the joint distribution of the 712 true p-values be constructed according to the configuration in Lemma 3.1. Note that for such a joint distribution of p-values, we have that

$$P\{FDP > \gamma\} \geq P\left\{\bigcup_{i=1}^{28}\{\hat{q}_i \leq \alpha\beta_i\}\right\} = \alpha|I|\sum_{i=1}^{28}\frac{\beta_i-\beta_{i-1}}{i} = 3.2212\alpha.$$

Hence, the largest one could possibly increase the constants by a multiple and still maintain control of the FDP is by a factor of $3.4179/3.2212 \approx 1.061$.

TABLE 1
Values of $D(\gamma, s)$ and $C_{\lfloor \gamma s \rfloor + 1}$

s	γ	$D(\gamma, s)$	$C_{\lfloor \gamma s \rfloor + 1}$	Ratio
100	0.01	1	1.5	1.5
250	0.01	1.4981	1.8333	1.2238
500	0.01	1.7246	2.45	1.4206
1000	0.01	2.0022	3.0199	1.5083
2000	0.01	2.3515	3.6454	1.5503
5000	0.01	2.8929	4.5188	1.562
25	0.05	1.4286	1.5	1.05
50	0.05	1.4952	1.8333	1.2262
100	0.05	1.734	2.45	1.4129
250	0.05	2.1237	3.1801	1.4974
500	0.05	2.4954	3.8544	1.5446
1000	0.05	2.9177	4.5188	1.5488
2000	0.05	3.3817	5.1973	1.5369
5000	0.05	4.0441	6.1047	1.5095
10	0.1	1	1.5	1.5
25	0.1	1.4975	1.8333	1.2242
50	0.1	1.7457	2.45	1.4034
100	0.1	2.0385	3.0199	1.4814
250	0.1	2.5225	3.8544	1.528
500	0.1	2.9502	4.5188	1.5317
1000	0.1	3.4179	5.1973	1.5206
2000	0.1	3.9175	5.883	1.5017
5000	0.1	4.6154	6.7948	1.4722

It is worthwhile to note that the argument used in the proof of Theorem 3.4 does not depend on the specific form of the original α_i. In fact, it can be used with any nondecreasing sequence of constants to construct a stepdown procedure that controls the FDP by scaling the constants appropriately. To see that this is the case, consider any nondecreasing sequence of constants $\delta_1 \leq \cdots \leq \delta_s$ such that $0 \leq \delta_i \leq 1$ (this restriction is without loss of generality since it can always be acheived by rescaling the constants if necessary) and redefine the constants β_m of equations (3.9) and (3.10) by the rule

$$\beta_m = \delta_{k(s,\gamma,m,|I|)} \qquad m = 1, \ldots, \lfloor \gamma s \rfloor + 1 \tag{3.15}$$

where

$$k(s, \gamma, m, |I|) = \min\{s, s + m - |I|, \lceil \frac{m}{\gamma} \rceil - 1\}.$$

Note that in the special case where $\delta_i = \alpha_i$, the definition of β_m in equation (3.15) agrees with the earlier definition of equations (3.9) and (3.10). Maintaining the definitions of N, S, and D in equations (3.11) - (3.13) (where they are now defined in terms of the β_m sequence given by equation (3.15)), we then have the following result:

Theorem 3.5. *For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, suppose $\hat{p}_i$ satisfies* (2.1). *Let $\delta_1 \leq \cdots \leq \delta_s$ be any nondecreasing sequence of constants such that $0 \leq \delta_i \leq 1$ and consider the stepdown procedure with constants $\delta_i'' = \alpha\delta_i / D(\gamma, s)$, where $D(\gamma, s)$ is defined by* (3.13). *Then, $P\{FDP > \gamma\} \leq \alpha$.*

Proof. Define j and m as in the proof of Theorem 3.4. We have, as before, that whenever $m - 1 \leq \gamma j < m$

$$|I| \leq s + m - j,$$

and

$$j \leq \lceil \frac{m}{\gamma} \rceil - 1.$$

Since $j \leq s$, it follows that

$$\hat{q}_{(m)} \leq \delta_j \leq \beta_m,$$

where β_m is as defined in (3.15). The remainder of the argument is identical to the proof of Theorem 3.4 so we do not repeat it here. □

As an illustration of this more general result, consider the nondecreasing sequence of constants given simply by $\eta_i = \frac{i}{s}$. These constants are proportional to the constants used in the procedures for controlling the FDR by Benjamini and Hochberg [1] and Benjamini and Yekutieli [3]. Applying Theorem 3.5 to this sequence of constants yields the following corollary:

Corollary 3.1. *For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, suppose $\hat{p}_i$ satisfies* (2.1)*. Then the following are true:*
(i) *The stepdown procedure with constants $\eta_i' = \alpha\eta_i / D(\gamma, s)$, where $D(\gamma, s)$ is defined by* (3.13)*, satisfies $P\{FDP > \gamma\} \leq \alpha$;*
(ii) *The stepdown procedure with constants $\eta_i'' = \gamma\alpha\eta_i / \max\{C_{\lfloor \gamma s \rfloor}, 1\}$, where C_0 is understood to equal* 0*, satisfies $P\{FDP > \gamma\} \leq \alpha$.*

Proof. The proof of (i) follows immediately from Theorem 3.5. To prove (ii), first observe that $N \leq \lfloor \gamma s \rfloor + 1$ and that for this particular sequence, we have that $\beta_m \leq \min\{\frac{m}{\gamma s}, 1\} =: \zeta_m$. Hence, we have that

$$P\{\bigcup_{i=1}^{N} \{\hat{q}_{(m)} \leq \beta_m\}\} \leq P\{\bigcup_{m=1}^{\lfloor \gamma s \rfloor + 1} \{\hat{q}_{(m)} \leq \zeta_m\}\}.$$

Using Lemma 3.1, we can bound the righthand side of this inequality by the sum

$$|I| \sum_{m=1}^{\lfloor \gamma s \rfloor + 1} \frac{\zeta_m - \zeta_{m-1}}{m}.$$

Whenever $\lfloor \gamma s \rfloor \geq 1$, we have that $\zeta_{\lfloor \gamma s \rfloor + 1} = \zeta_{\lfloor \gamma s \rfloor} = s$, so this sum can in turn be bounded by

$$\frac{|I|}{\gamma s} \sum_{m=1}^{\lfloor \gamma s \rfloor} \frac{1}{m} \leq \frac{1}{\gamma} C_{\lfloor \gamma s \rfloor}.$$

If, on the other hand, $\lfloor \gamma s \rfloor = 0$, we can simply bound the sum by $\frac{1}{\gamma}$. Therefore, if we let $C_0 = 0$, we have that

$$D(\gamma, s) \leq \frac{1}{\gamma} \max\{C_{\lfloor \gamma s \rfloor}, 1\},$$

from which the desired claim follows. □

In summary, given any nondecreasing sequence of constants δ_i, we have derived a stepdown procedure which controls the FDP, and so it is interesting to compare such FDP-controlling procedures. Clearly, a procedure with larger critical values is preferable to one with smaller ones, subject to the error constraint. The discussion from Remark 3.1 leads us to believe that the critical values from a single procedure will not uniformly dominate those from another, at least approximately. We now consider some specific comparisons which may shed light on how to choose among the various procedures.

TABLE 2
Values of $D(\gamma, s)$ *and* $\frac{1}{\gamma}\max\{C_{\lfloor \gamma s \rfloor}, 1\}$

s	γ	$D(\gamma, s)$	$\frac{1}{\gamma}\max\{C_{\lfloor \gamma s \rfloor}, 1\}$	Ratio
100	0.01	25.5	100	3.9216
250	0.01	60.4	150	2.4834
500	0.01	90.399	228.33	2.5258
1000	0.01	128.53	292.9	2.2788
2000	0.01	171.73	359.77	2.095
5000	0.01	235.94	449.92	1.9069
25	0.05	6.76	20	2.9586
50	0.05	12.4	30	2.4194
100	0.05	18.393	45,667	2.4828
250	0.05	28.582	62.064	2.1714
500	0.05	37.513	76.319	2.0345
1000	0.05	47.26	89.984	1.904
2000	0.05	57.666	103.75	1.7991
5000	0.05	72.126	122.01	1.6917
10	0.1	3	10	3.3333
25	0.1	6.4	15	2.3438
50	0.1	9.3867	22.833	2.4325
100	0.1	13.02	29.29	2.2496
250	0.1	18.834	38.16	2.0261
500	0.1	23.703	44.992	1.8981
1000	0.1	28.886	51.874	1.7958
2000	0.1	34.317	58.78	1.7129
5000	0.1	41.775	67.928	1.6261

To compare the constants from parts (i) and (ii) of Corollary 3.1, Table 2 displays $D(\gamma, s)$ and $\frac{1}{\gamma}\max\{C_{\lfloor \gamma s \rfloor}, 1\}$ for several different values of s and γ, as well as the ratio $\frac{1}{\gamma}\max\{C_{\lfloor \gamma s \rfloor}, 1\}/D(\gamma, s)$. In this instance, the improvement between the constants from part (i) and part (ii) is dramatic: The constants η_i' are often at least twice as large as the constants η_i''.

It is also of interest to compare the constants from part (i) of the corollary with those from Theorem 3.4. We do this for the case in which $s = 100$, $\gamma = .1$, and $\alpha = .05$ in Figure 1. The top panel displays the constants α_i'' from Theorem 3.4 and the middle panel displays the constants η_i' from Corollary 3.1 (i). Note that the scale of the top panel is much larger than the scale of the middle panel. It is therefore clear that the constants α_i'' are generally much larger than the constants η_i'. But it is important to note that the constants from Theorem 3.4 are not uniformly larger than the constants from Corollary 3.1 (i). To make this clear, the bottom panel of Figure 1 displays the ratio α_i''/η_i'. Notice that at steps 7 - 9, 15 - 19, and 25 - 29 the ratios are strictly less than 1, meaning that at those steps the η_i' are larger than the α_i''. Following our discussion in Remark 3.1 that these constants are very nearly the best possible up to a scalar multiple, we should expect that this would be the case because otherwise the constants η_i' could be multiplied by a factor larger than 1 and still retain control of the FDP. Even at these steps, however, the constants η_i' are very close to the constants α_i'' in absolute terms. Since the constants α_i'' are considerably larger than the constants η_i' at other steps, this suggests that the procedure based upon the constants α_i'' is preferrable to the procedure based on the constants η_i'.

4. Control of the FDR

Next, we construct a stepdown procedure that controls the FDR under the same conditions as Theorem 3.1. The dependence condition used is much weaker than

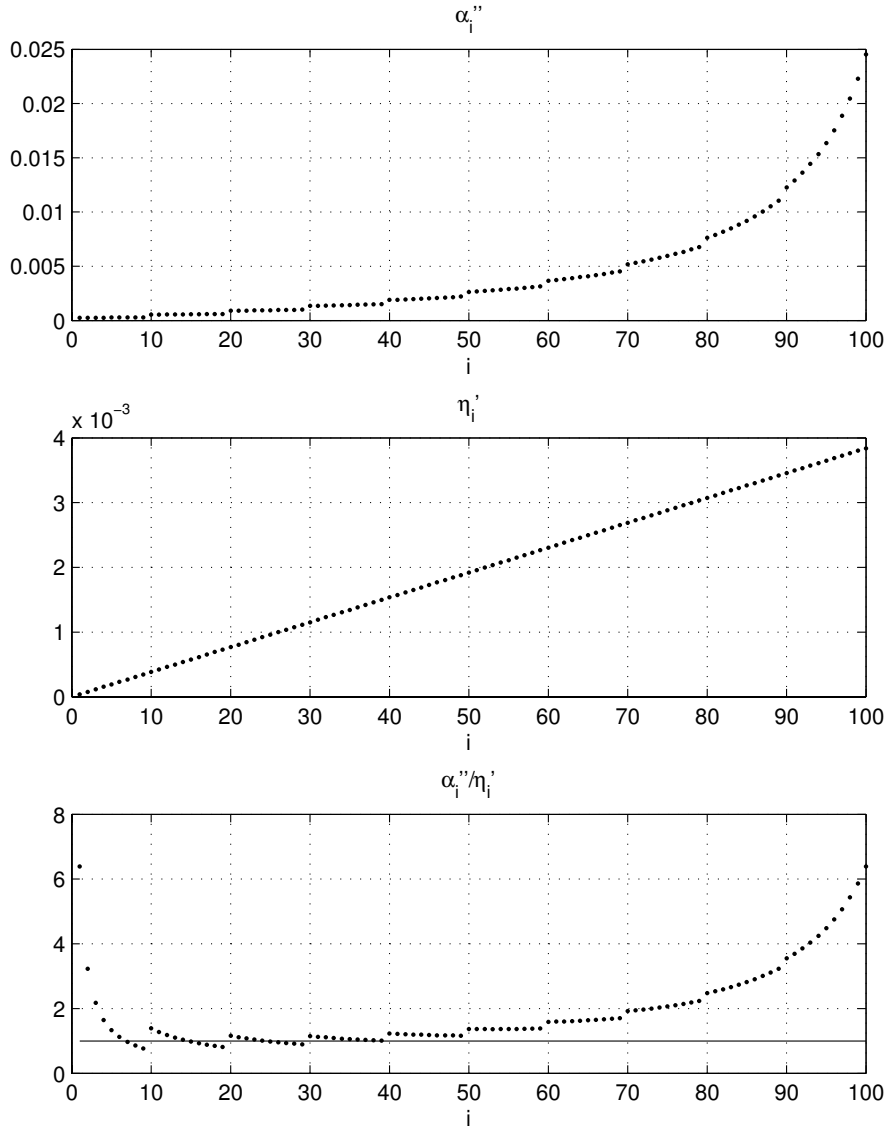


FIG 1. *Stepdown Constants for* $s = 100$, $\gamma = .1$, *and* $\alpha = .05$.

that of independence of p-values used by Benjamini and Liu [2].

Theorem 4.1. *For testing* $H_i : P \in \omega_i$, $i = 1, \ldots, s$, *suppose* $\hat{p}_i$ *satisfies* (2.1). *Consider the stepdown procedure with constants*

$$\alpha_i^* = \min\{\frac{s\alpha}{(s-i+1)^2}, 1\} \tag{4.1}$$

and assume the condition (3.5). *Then,* $FDR \leq \alpha$.

Proof. First note that if $|I| = 0$, then $FDR = 0$. Second, if $|I| = s$, then $FDR = P\{\hat{p}_{(1)} \leq \alpha_1^*\} \leq \sum_{i=1}^{s} P\{\hat{p}_i \leq \alpha_1^*\} \leq s\alpha_1^* = \alpha$.

Now suppose that $0 < |I| < s$. Define $\hat{q}_1, \ldots, \hat{q}_{|I|}$ and $\hat{r}_1, \ldots, \hat{r}_{s-|I|}$ to be the p-values corresponding, respectively, to the true and false hypotheses, and let $\hat{q}_{(1)} \leq \cdots \leq \hat{q}_{(|I|)}$ and $\hat{r}_{(1)} \leq \cdots \leq \hat{r}_{(s-|I|)}$ be their ordered values. Denote by j the largest index such that $\hat{r}_{(1)} \leq \alpha_1^*, \ldots, \hat{r}_{(j)} \leq \alpha_j^*$ (defined to be 0 if $\hat{r}_{(1)} > \alpha_1^*$). Define t to be the total number of true hypotheses rejected by the stepdown procedure and f to be the total number of false hypotheses rejected by the stepdown procedure.

Using this notation, observe that

$$E(FDP|\hat{r}_1,\ldots,\hat{r}_{s-|I|}) = E(\frac{t}{t+f}\{t+f>0\}|\hat{r}_1,\ldots,\hat{r}_{s-|I|})$$

$$\leq E(\frac{t}{t+j}\{t>0\}|\hat{r}_1,\ldots,\hat{r}_{s-|I|})$$

$$\leq \frac{|I|}{|I|+j}E(\{t>0\}|\hat{r}_1,\ldots,\hat{r}_{s-|I|})$$

$$\leq \frac{|I|}{|I|+j}P\{\hat{q}_{(1)} \leq \alpha^*_{j+1}|\hat{r}_1,\ldots,\hat{r}_{s-|I|})$$

$$\leq \frac{|I|}{|I|+j}\sum_{i=1}^{|I|}P\{\hat{q}_i \leq \alpha^*_{j+1}|\hat{r}_1,\ldots,\hat{r}_{s-|I|}\}$$

$$\leq \frac{|I|}{|I|+j}|I|\alpha^*_{j+1} \tag{4.2}$$

$$\leq \frac{|I|^2}{|I|+j}\min\{\frac{s\alpha}{(s-j)^2},1\}$$

$$\leq \frac{|I|\alpha}{(s-j)}\frac{|I|s}{(|I|+j)(s-j)}. \tag{4.3}$$

The inequality (4.2) follows from the assumption (3.5) on the joint distribution of p-values. To complete the proof, note that $|I|+j \leq s$. It follows that $\frac{|I|\alpha}{(s-j)} \leq \alpha$ and $(|I|+j)(s-j)-|I|s = j(s-|I|)-j^2 = j(s-|I|-j) \geq 0$. Combining these two inequalities, we have that the expression in (4.3) is bounded above by α. The desired bound for the FDR follows immediately. □

The following simple example illustrates the fact that the FDR is not controlled by the stepdown procedure with constants α^*_i absent the restriction (3.5) on the dependence structure of the p-values.

Example 4.1. Suppose there are $s=3$ hypotheses, two of which are true. In this case, $\alpha^*_1 = \frac{\alpha}{3}$, $\alpha^*_2 = \frac{3\alpha}{4}$, and $\alpha^*_3 = \min\{3\alpha,1\}$. Define the joint distribution of the two true p-values q_1 and q_2 as follows: Denote by I_i the half open interval $[\frac{i-1}{3},\frac{i}{3})$ and let $(q_1,q_2) \sim U(I_i \times I_j)$ with probability $\frac{1}{6}$ for all (i,j) such that $i \neq j$, $1 \leq i \leq 3$ and $1 \leq j \leq 3$. It is easy to see that $(q_{(1)},q_{(2)}) \sim U(I_i \times I_j)$ with probability $\frac{1}{3}$ for all (i,j) such that $i<j$, $1 \leq i \leq 3$ and $1 \leq j \leq 3$. Now define the distribution of the false p-value r_1 conditional on (q_1,q_2) by the following rule: If $q_{(1)} \leq \alpha/3$, then let $r_1=1$; otherwise, let $r_1=0$. For such a joint distribution of (q_1,q_2,r_1), we have that the FDP is identically one whenever $q_{(1)} \leq \frac{\alpha}{3}$ and is at least $\frac{1}{2}$ whenever $\frac{\alpha}{3} < q_{(1)} \leq \frac{3\alpha}{4}$. Hence,

$$FDR \geq P\{q_{(1)} \leq \frac{\alpha}{3}\} + \frac{1}{2}P\{\frac{\alpha}{3} < q_{(1)} \leq \frac{3\alpha}{4}\}.$$

For $\alpha < \frac{4}{9}$, we therefore have that

$$FDR \geq \frac{2\alpha}{3} + (\frac{3\alpha}{4} - \frac{\alpha}{3}) = \frac{13\alpha}{12} > \alpha.$$

Remark 4.1. Some may find it unpalatable to allow the constants to exceed α. In this case, one might consider replacing the constants α_i^* above with the more conservative values $\alpha \min\{\frac{s}{(s-i+1)^2}, 1\}$, which by construction are always less than α. Since these constants are uniformly smaller than the α_i^*, our method of proof shows that the FDR would still be controlled under the dependence condition (3.5). The above counterexample, which did not depend on the particular value of α_3^*, however, would show that it is not controlled in general.

Under the dependence condition (3.5), the constants (4.1) control the FDR in the sense $FDR \leq \alpha$, while the constants given by (3.4) control the FDP in the sense of (3.3). Utilizing (1.1), we can use the constants (4.1) to control the FDP by controlling the FDR at level $\alpha\gamma$. In Figure 2, we plot the constants (3.4) and (4.1) for the special case in which $s = 100$ and we use both constants to control the FDP for $\gamma = .1$, and $\alpha = .05$.

The top panel displays the constants α_i, the middle panel displays the constants α_i^*, and the bottom panel displays the ratio α_i/α_i^*. Since the ratios essentially always exceed 1, it is clear that in this instance the constants (3.4) are superior to

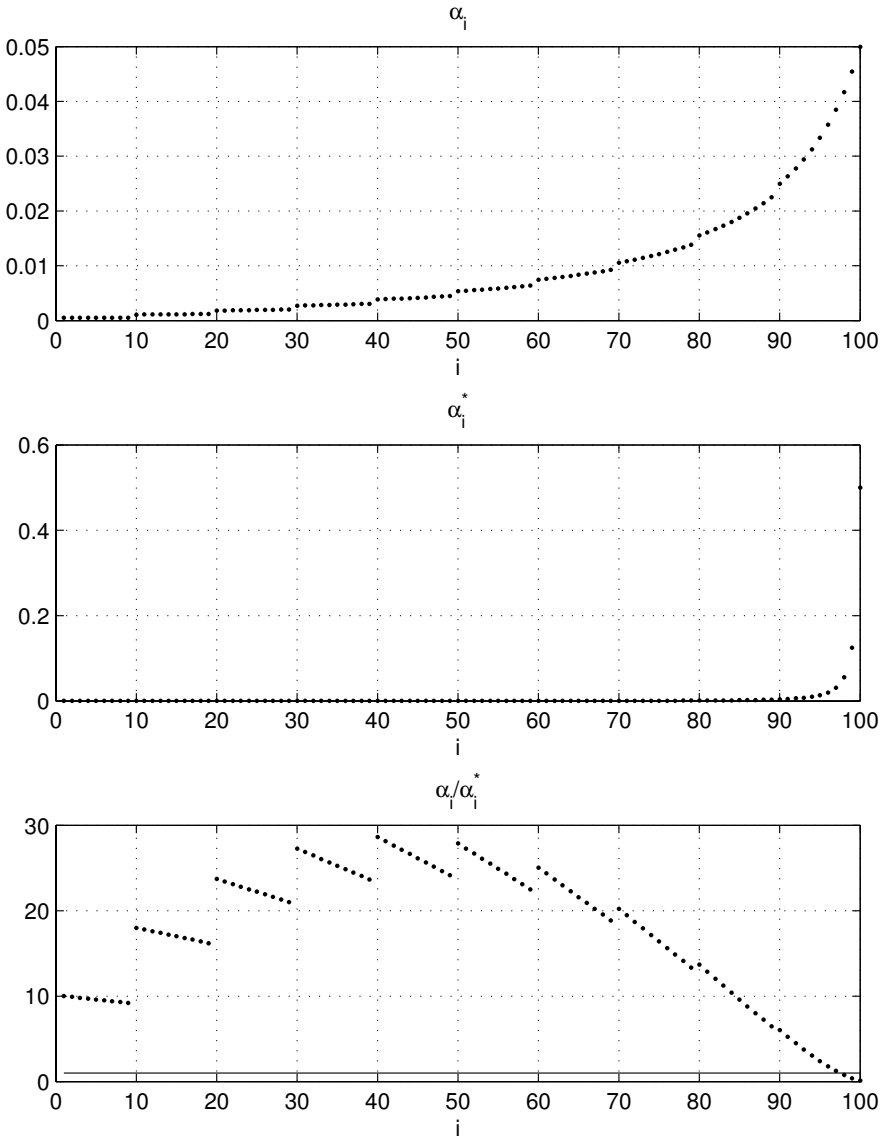


FIG 2. *FDP Control for* $s = 100$, $\gamma = .1$, *and* $\alpha = .05$.

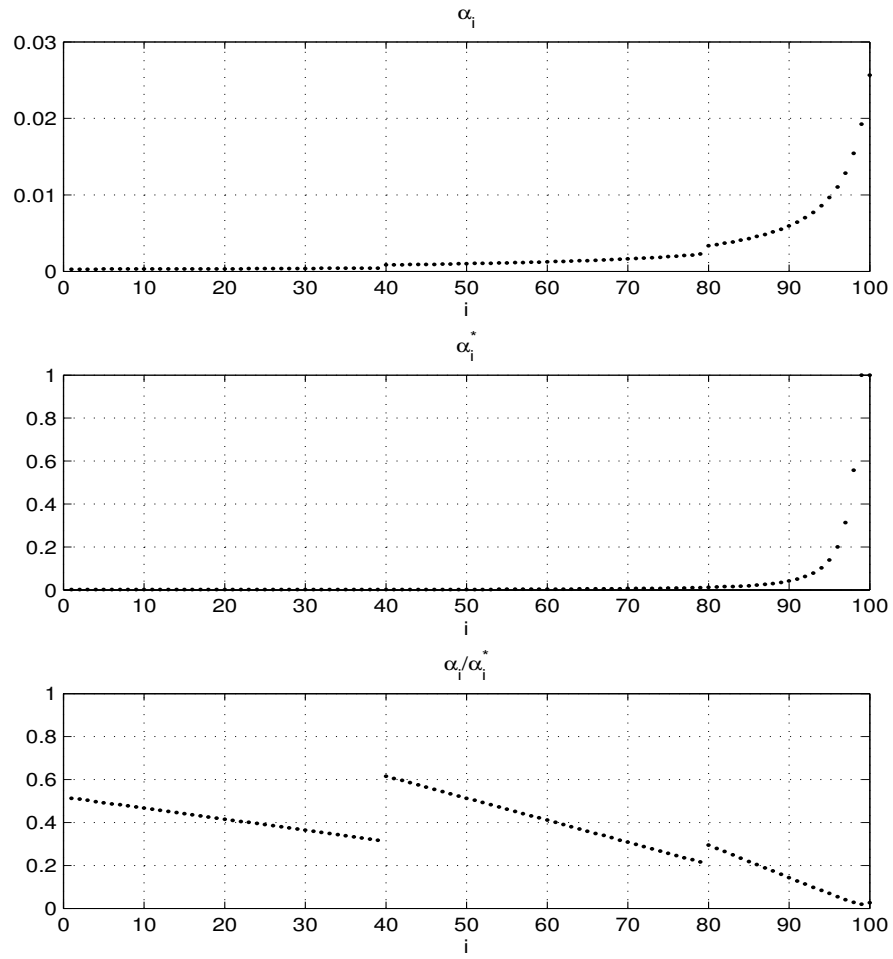


FIG 3. *FDR Control for* $s = 100$ *and* $\alpha = .05$.

the constants (4.1). If by utilizing (1.1) we use the constants (3.4) to control the FDR, on the other hand, we find that the reverse is true. Control of the FDR at level α can be achieved, for example, by controlling the FDP at level $\frac{\alpha}{2-\alpha}$ and letting $\gamma = \frac{\alpha}{2}$. Figure 3 plots the constants (3.4) and (4.1) for the special case in which $s = 100$ and we use both constants to control the FDR at level $\alpha = .05$.

As before, the top panel displays the constants α_i, the middle panel displays the constants α_i^*, and the bottom panel displays the ratio α_i/α_i^*. In this case, the ratio is always less than 1. Thus, in this instance, the constants α_i^* are preferred to the constants α_i. Of course, the argument used to establish (1.1) is rather crude, but it nevertheless suggests that it is worthwhile to consider the type of control desired when choosing critical values.

5. Conclusions

In this article we have described stepdown procedures for testing multiple hypotheses that control the FDP without any restrictions on the joint distribution of the p-values. First, we have improved upon a method proposed by Lehmann and Romano [10]. The new procedure is a considerable improvement in the sense that its critical values are generally 50 percent larger than those of the earlier procedure. Second, we have generalized the method of argument used in establishing this improvement to provide a means by which any nondecresing sequence of constants

can be rescaled so as to ensure control of the FDP. Finally, we have also described a procedure that controls the FDR, but only under an assumption on the joint distribution of the p-values.

In this article, we focused on the class of stepdown procedures. The alternative class of *stepup* procedures can be described as follows. Let

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s \tag{5.1}$$

be a nondecreasing sequence of constants. If $\hat{p}_{(s)} \leq \alpha_s$, then reject all null hypotheses; otherwise, reject hypotheses $H_{(1)}, \ldots, H_{(r)}$ where r is the smallest index satisfying

$$\hat{p}_{(s)} > \alpha_s, \ldots, \hat{p}_{(r+1)} > \alpha_{r+1}. \tag{5.2}$$

If, for all r, $\hat{p}_{(r)} > \alpha_r$, then reject no hypotheses. That is, a stepup procedure begins with the least significant p-value and continues accepting hypotheses as long as their corresponding p-values are large. If both a stepdown procedure and stepup procedure are based on the same set of constants α_i, it is clear that the stepup procedure will reject at least as many hypotheses.

For example, the well-known stepup procedure based on $\alpha_i = i\alpha/s$ controls the FDR at level α, as shown by Benjamini and Hochberg [1] under the assumption that the p-values are mutually independent. Benjamini and Yekutieli [3] generalize their result to allow for certain types of dependence; also see Sarkar [14]. Benjamini and Yekutieli [3] also derive a procedure controlling the FDR under no dependence assumptions. Romano and Shaikh [12] derive stepup procedures which control the k-$FWER$ and the FDP under no dependence assumptions, and some comparisons with stepdown procedures are made as well.

Acknowledgements

We wish to thank Juliet Shaffer for some helpful discussion and references.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and forceful approach to multiple testing. *J. Roy. Statist. Soc. Series B* **57**, 289–300.

[2] Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82**, 163–170.

[3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.

[4] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–1061.

[5] Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures.* Wiley, New York.

[6] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

[7] Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biom. J.* **25**, 423–430.

[8] Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In *Multiple Hypothesis Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.). Springer, Heidelberg, 154–161.

[9] Korn, E., Troendle, J., McShane, L. and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124**, 379–398.

[10] Lehmann, E. L. and Romano, J. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138–1154.

[11] Perone Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery rates for random fields. *J. Amer. Statist. Assoc.* **99**, 1002–1014.

[12] Romano, J. and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.*, to appear.

[13] Sarkar, S. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of Simes conjecture. *Ann. Statist.* **26**, 494–504.

[14] Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239–257.

[15] Sarkar, S. and Chang, C. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92**, 1601–1608.

[16] Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

[17] van der Laan, M., Dudoit, S., and Pollard, K. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statist. Appl. Gen. Molec. Biol.* **3**, 1, Article 15.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 51–76
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000392

An adaptive significance threshold criterion for massive multiple hypotheses testing

Cheng Cheng[1,*]

St. Jude Children's Research Hospital

Abstract: This research deals with massive multiple hypothesis testing. First regarding multiple tests as an estimation problem under a proper population model, an error measurement called *Erroneous Rejection Ratio* (ERR) is introduced and related to the False Discovery Rate (FDR). ERR is an error measurement similar in spirit to FDR, and it greatly simplifies the analytical study of error properties of multiple test procedures. Next an improved estimator of the proportion of true null hypotheses and a data adaptive significance threshold criterion are developed. Some asymptotic error properties of the significant threshold criterion is established in terms of ERR under distributional assumptions widely satisfied in recent applications. A simulation study provides clear evidence that the proposed estimator of the proportion of true null hypotheses outperforms the existing estimators of this important parameter in massive multiple tests. Both analytical and simulation studies indicate that the proposed significance threshold criterion can provide a reasonable balance between the amounts of false positive and false negative errors, thereby complementing and extending the various FDR control procedures. S-plus/R code is available from the author upon request.

1. Introduction

The recent advancement of biological and information technologies made it possible to generate unprecedented large amounts of data for just a single study. For example, in a genome-wide investigation, expressions of tens of thousands genes and markers can be generated and surveyed simultaneously for their association with certain traits or biological conditions of interest. Statistical analysis in such applications poses a massive multiple hypothesis testing problem. The traditional approaches to controlling the probability of family-wise type-I error have proven to be too conservative in such applications. Recent attention has been focused on the control of *false discovery rate* (FDR) introduced by Benjamini and Hochberg [4]. Most of the recent methods can be broadly characterized into several approaches. Mixture-distribution partitioning [2, 24, 25] views the P values as random variables and models the P value distribution to generate estimates of the FDR levels at various significance levels. Significance analysis of microarrays (SAM; [32, 35]) employs permutation tests to inference simultaneously on order statistics. Empirical Baysian approaches include for example [10, 11, 17, 23, 28]. Tsai *et al.* [34] proposed models

*Supported in part by the NIH grants U01 GM-061393 and the Cancer Center Support Grant P30 CA-21765, and the American Lebanese and Syrian Associated Charities (ALSAC).

[1]Department of Biostatistics, Mail Stop 768, St. Jude Childrens Research Hospital, 332 North Lauderdale Street, Memphis, TN 38105-2794, USA, e-mail: cheng.cheng@stjude.org

AMS 2000 subject classifications: primary 62F03, 62F05, 62F07, 62G20, 62G30, 62G05; secondary 62E10, 62E17, 60E15.

Keywords and phrases: multiple tests, false discovery rate, q-value, significance threshold selection, profile information criterion, microarray, gene expression.

and estimators of the conditional FDR, and Bickel [6] takes a decision-theoretic approach. Recent theoretical developments on FDR control include Genovese and Wasserman [13, 14], Storey *et al.* [31], Finner and Roberts [12], and Abramovich *et al.* [1]. Recent theoretical development on control of generalized family-wise type-I error includes van der Laan *et al.* [36, 37], Dudoit *et al.* [9], and the references therein.

Benjamini and Hochberg [4] argue that as an alternative to the family-wise type-I error probability, FDR is a proper measurement of the amount of false positive errors, and it enjoys many desirable properties not possessed by other intuitive or heuristic measurements. Furthermore they develop a procedure to generate a significance threshold (P value cutoff) that guarantees the control of FDR under a pre-specified level. Similar to a significance test, FDR control requires one to specify a control level *a priori*. Storey [29] takes the point of view that in discovery-oriented applications neither the FDR control level nor the significance threshold may be specified before one sees the data (P values), and often the significance threshold is so determined *a posteriori* that allows for some "discoveries" (rejecting one or more null hypotheses). These "discoveries" are then scrutinized in confirmation and validation studies. Therefore it would be more appropriate to measure the false positive errors conditional on having rejected some null hypotheses, and for this purpose the positive FDR (pFDR; Storey [29]) is a meaningful measurement. Storey [29] introduces estimators of FDR and pFDR, and the concept of q-value which is essentially a neat representation of Benjamini and Hochberg's ([4]) step-up procedure possessing a Bayesian interpretation as the posterior probability of the null hypothesis ([30]). Reiner *et al.* [26] introduce the "FDR-adjusted P value" which is equivalent to the q-value. The q-value plot ([33]) allows for visualization of FDR (or pFDR) levels in relationship to significance thresholds or numbers of null hypotheses to reject. Other closely related procedures are the adaptive FDR control by Benjamini and Hochberg [3], and the recent two-stage linear step-up procedure by Benjamini *et al.* [5] which is shown to provide sure FDR control at any pre-specified level.

In discovery-oriented exploratory studies such as genome-wide gene expression survey or association rule mining in marketing applications, it is desirable to strike a meaningful balance between the amounts false positive and false negative errors than to control the FDR or pFDR alone. Cheng *et al.* [7] argue that it is not always clear in practice how to specify the threshold for either the FDR level or the significance level. Therefore, additional statistical guidelines beyond FDR control procedures are desirable. Genovese and Wasserman [13] extend FDR control to a minimization of the "false nondiscovery rate" (FNR) under a penalty of the FDR, i.e., $FNR+\lambda FDR$, where the penalty λ is assumed to be specified *a priori*. Cheng *et al.* [7] propose to extract more information from the data (P values) and introduce three data-driven criteria for determination of the significance threshold.

This paper has two related goals: (1) develop a more accurate estimator of the proportion of true null hypotheses, which is an important parameter in all multiple hypothesis testing procedures; and (2) further develop the "profile information criterion" I_p introduced in [7] by constructing a more data-adaptive criterion and study its asymptotic error behavior (as the number of tests tends to infinity) theoretically and via simulation. For theoretical and methodological development, a new meaningful measurement of the quantity of false positive errors, the *erroneous rejection ratio* (ERR), is introduced. Just like FDR, ERR is equal to the family-wise type-I error probability when all null hypotheses are true. Under the ergodicity conditions used in recent studies ([14, 31]), ERR is equal to FDR at any significant threshold

(P value cut-off). On the other hand, ERR is much easier to handle analytically than FDR under distributional assumptions more widely satisfied in applications. Careful examination of each component in ERR gives insights into massive multiple testing in terms of the ensemble behavior of the P values. Quantities derived from ERR suggest to construct improved estimators of the null proportion (or the number of true null hypotheses) considered in [3, 29, 31], and the construction of an adaptive significance threshold criterion. The theoretical results demonstrate how the criterion can be calibrated with the Bonferroni adjustment to provide control of family-wise type-I error probability when all null hypotheses are true, and how the criterion behaves asymptotically, giving cautions and remedies in practice. The simulation results are consistent with the theory, and demonstrate that the proposed adaptive significance criterion is a useful and effective procedure complement to the popular FDR control methods.

This paper is organized as follows: Section 2 contains a brief review of FDR and the introduction of ERR; section 3 contains a brief review of the estimation of the proportion of null hypotheses, and the development of an improved estimator; section 4 develops the adaptive significance threshold criterion and studies its asymptotic error behavior (as the number of hypotheses tends to infinity) under proper distributional assumptions on the P values; section 5 contains a simulation study; and section 6 contains concluding remarks.

Notation. Henceforth, $\mathbb{R}$ denotes the real line; $\mathbb{R}^k$ denotes the k dimensional Euclidean space. The symbol $\|\cdot\|_p$ denotes the L^p or ℓ^p norm, and $:=$ indicates equal by definition. Convergence and convergence in probability are denoted by $\longrightarrow$ and $\longrightarrow_p$ respectively. A random variable is usually denoted by an upper-case letter such as P, R, V, etc. A cumulative distribution function (cdf) is usually denoted by F, G or H; an empirical distribution function (EDF) is usually indicated by a tilde, e.g., $\widetilde{F}$. A population parameter is usually denoted by a lower-case Greek letter and a hat indicates an estimator of the parameter, e.g., $\widehat{\theta}$. Equivalence is denoted by $\simeq$, e.g., "$a_n \simeq b_n$ as $n \longrightarrow \infty$" means $\lim_{n\longrightarrow\infty} a_n/b_n = 1$.

2. False discovery rate and erroneous rejection ratio

Consider testing m hypothesis pairs (H_{0i}, H_{Ai}), $i = 1, \ldots, m$. In many recent applications such as analysis of microarray gene differential expressions, m is typically on the order of 10^5. Suppose m P values, $P_1, \ldots, P_m$, one for each hypothesis pair, are calculated, and a decision on whether to reject H_{0i} is to be made. Let m_0 be the number of true null hypotheses, and let $m_1 := m - m_0$ be the number of true alternative hypotheses. The outcome of testing these m hypotheses can be tabulated as in Table 1 (Benjamini and Hochberg [4]), where V is the number of null hypotheses erroneously rejected, S is the number of alternative hypotheses correctly captured, and R is the total number of rejections.

TABLE 1
Outcome tabulation of multiple hypotheses testing.

True Hypotheses	Rejected	Not Rejected	Total
H_0	V	$m_0 - V$	m_0
H_A	S	$m_1 - S$	m_1
Total	R	$m - R$	m

Clearly only m is known and only R is observable. At least one *family-wise type-I error* is committed if $V > 0$, and procedures for multiple hypothesis testing have traditionally been produced for solely controlling the family-wise type-I error probability $\Pr(V > 0)$. It is well-known that such procedures often lack statistical power. In an effort to develop more powerful procedures, Benjamini and Hochberg ([4]) approached the multiple testing problem from a different perspective and introduced the concept of *false discovery rate* (FDR), which is, loosely speaking, the expected value of the ratio V/R. They introduced a simple and effective procedure for controlling the FDR under any pre-specified level.

It is convenient both conceptually and notationally to regard multiple hypotheses testing as an estimation problem ([7]). Define the parameter $\Theta = [\theta_1, \ldots, \theta_m]$ as $\theta_i = 1$ if H_{Ai} is true, and $\theta_i = 0$ if H_{0i} is true ($i = 1, \ldots, m$). The data consist of the P values $\{P_1, \ldots, P_m\}$, and under the assumption that each test is exact and unbiased, the population is described by the following probability model:

$$\begin{gathered} P_i \sim \mathcal{P}_{i,\theta_i}; \\ \mathcal{P}_{i,0} \text{ is } U(0,1), \text{ and } \mathcal{P}_{i,1} <_{st} U(0,1); \\ \text{each } \mathcal{P}_{i,1} \text{ has a twice continuously differentiable cdf } F_i(\cdot), \end{gathered} \tag{2.1}$$

for $i = 1, \ldots, m$, where $<_{st}$ stands for "stochastically less than." The P values are dependent in general and have a joint distribution on $\mathbb{R}^m$. By this model the marginal cdf of P_i can be written as $G_i(t) = (1 - \theta_i)t + \theta_i F_i(t)$. Note $F_i(t) \geq t$ and $G_i(t) \geq t$ for $t \in [0, 1]$.

A rejection procedure is an estimator of Θ: $\widehat{\Theta} = \widehat{\Theta}(P_1, \ldots, P_m) = [\widehat{\theta}_1, \ldots, \widehat{\theta}_m] \in \{0, 1\}^m$, where $\widehat{\theta}_i = 1$ indicates rejecting H_{0i} in favor of H_{Ai}, $i = 1, \ldots, m$. With this notation, the random variables in Table 1 can be expressed as

$$V = V_\Theta(\widehat{\Theta}) = \sum_{i=1}^{m} (1 - \theta_i)\widehat{\theta}_i; \; S = S_\Theta(\widehat{\Theta}) = \sum_{i=1}^{m} \theta_i \widehat{\theta}_i; \; R = R(\widehat{\Theta}) = \sum_{i=1}^{m} \widehat{\theta}_i. \tag{2.2}$$

A natural and perhaps the simplest procedure is the "hard-thresholding" (HT) estimator $\widehat{\Theta} = \widehat{\Theta}(\alpha)$ defined as

$$HT(\alpha): \quad \widehat{\theta}_i = 1 \text{ iff } P_i \leq \alpha, \tag{2.3}$$

where $\alpha \in (0, 1)$ is a significance threshold common to all hypotheses. Clearly for this procedure the random variables V, S, and R all depend on α. Traditional control of family-wise type-I error probability seeks to determine α so that $\Pr(V > 0) \leq \alpha^*$ for pre-specified α^*. Genovese and Wasserman [14] list several procedures to determine α. Benjamini and Hochberg [4] introduce a simple procedure to determine α so that the FDR is controlled at a given level.

2.1. False discovery rate and its control

The FDR as defined by Benjamini and Hochberg ([4]) can be expressed as

$$FDR_\Theta(\widehat{\Theta}) = E\left[\frac{\sum_{i=1}^{m} \widehat{\theta}_i(1 - \theta_i)}{\sum_{i=1}^{m} \widehat{\theta}_i + \prod_{i=1}^{m}(1 - \widehat{\theta}_i)}\right], \tag{2.4}$$

which is equivalent to $E[V/R | R > 0] \Pr(R > 0)$. Let $P_{1:m} \leq P_{2:m} \leq \cdots \leq P_{m:m}$ be the order statistics of the P values, and let $\pi_0 = m_0/m$. Benjamini and Hochberg

([4]) prove that for any specified $q^* \in (0,1)$ rejecting all the null hypotheses corresponding to $P_{1:m}, \ldots, P_{k^*:m}$ with $k^* = \max\{k : P_{k:m}/(k/m) \leq q^*\}$ controls the FDR at the level $\pi_0 q^*$, i.e., $FDR_\Theta(\widehat{\Theta}(P_{k^*:m})) \leq \pi_0 q^* \leq q^*$. Note this procedure is equivalent to applying the data-driven threshold $\alpha = P_{k^*:m}$ to all P values in (2.3), i.e., $HT(P_{k^*:m})$.

Recognizing the potential of constructing less conservative FDR controls by the above procedure, Benjamini and Hochberg ([3]) propose an estimator of m_0, $\widehat{m}_0$, (hence an estimator of π_0, $\widehat{\pi}_0 = \widehat{m}_0/m$), and replace k/m by $k/\widehat{m}_0$ in determining k^*. They call this procedure "adaptive FDR control." The estimator $\widehat{\pi}_0 = \widehat{m}_0/m$ will be discussed in Section 3. A recent development in adaptive FDR control can be found in Benjamini *et al.* [5].

Similar to a significance test, the above procedure requires the specification of an FDR control level q^* before the analysis is conducted. Storey ([29]) takes the point of view that for more discovery-oriented applications the FDR level is not specified *a priori*, but rather determined after one sees the data (P values), and it is often determined in a way allowing for some "discovery" (rejecting one or more null hypotheses). Hence a concept similar to, but different than FDR, the *positive false discovery rate* (pFDR) $E\left[V/R \middle| R > 0\right]$, is more appropriate. Storey ([29]) introduces estimators of π_0, the FDR, and the pFDR from which the *q-values* are constructed for FDR control. Storey *et al.* ([31]) demonstrate certain desirable asymptotic conservativeness of the q-values under a set of ergodicity conditions.

2.2. Erroneous rejection ratio

As discussed in [3, 4], the FDR criterion has many desirable properties not possessed by other intuitive alternative criteria for multiple tests. In order to obtain an analytically convenient expression of FDR for more in-depth investigations and extensions, such as in [13, 14, 29, 31], certain fairly strong ergodicity conditions have to be assumed. These conditions make it possible to apply classical empirical process methods to the "FDR process." However, these conditions may be too strong for more recent applications, such as genome-wide tests for gene expression–phenotype association using microarrays, in which a substantial proportion of the tests can be strongly dependent. In such applications it may not be even reasonable to assume that the tests corresponding to the true null hypotheses are independent, an assumption often used in FDR research. Without these assumptions however, the FDR becomes difficult to handle analytically. An alternative error measurement in the same spirit of FDR but easier to handle analytically is defined below.

Define the *erroneous rejection ratio* (ERR) as

$$ERR_\Theta(\widehat{\Theta}) = \frac{E[V_\Theta(\widehat{\Theta})]}{E[R(\widehat{\Theta})]} \Pr(R(\widehat{\Theta}) > 0). \tag{2.5}$$

Just like FDR, when all null hypotheses are true $ERR = \Pr(R(\widehat{\Theta}) > 0)$, which is the family-wise type-I error probability because now $V_\Theta(\widehat{\Theta}) = R(\widehat{\Theta})$ with probability one. Denote by $V(\alpha)$ and $R(\alpha)$ respectively the V and R random variables and by $ERR(\alpha)$ the ERR for the hard-thresholding procedure $HT(\alpha)$; thus

$$ERR(\alpha) = \frac{E[V(\alpha)]}{E[R(\alpha)]} \Pr(R(\alpha) > 0). \tag{2.6}$$

Careful examination of each component in $ERR(\alpha)$ reveals insights into multiple tests in terms of the ensemble behavior of the P values. Note

$$E[V(\alpha)] = \sum_{i=1}^{m}(1-\theta_i)\Pr(\widehat{\theta}_i = 1) = m_0\alpha$$
$$E[R(\alpha)] = \sum_{i=1}^{m}\Pr(\widehat{\theta}_i = 1) = m_0\alpha + \sum_{j:\theta_j=1} F_j(\alpha)$$
$$\Pr(R(\alpha) > 0) = \Pr(P_{1:m} \le \alpha).$$

Define $H_m(t) := m_1^{-1}\sum_{j:\theta_j=1} F_j(t)$ and $F_m(t) := m^{-1}\sum_{i=1}^{m} G_i(t) = \pi_0 t + (1-\pi_0)H_m(t)$. Then

$$ERR(\alpha) = \frac{\pi_0\alpha}{F_m(\alpha)}\Pr(P_{1:m} \le \alpha). \tag{2.7}$$

The functions $H_m(\cdot)$ and $F_m(\cdot)$ both are cdf's on $[0,1]$; H_m is the average of the P value marginal cdf's corresponding to the true alternative hypotheses, and F_m is the average of all P value marginal cdf's. F_m describes the ensemble behavior of all P values and H_m describes the ensemble behavior of the P values corresponding to the true alternative hypotheses. Cheng *et al.* ([7]) observe that the EDF of the P values $\widetilde{F}_m(t) := m^{-1}\sum_{i=1}^{m} I(P_i \le t)$, $t \in \mathbb{R}$ is an unbiased estimator of $F_m(\cdot)$, and if the tests $\widehat{\theta}$ $(i = 1, \ldots, m)$ are not strongly correlated asymptotically in the sense that $\sum_{i\neq j} Cov(\widehat{\theta}_i, \widehat{\theta}_j) = o(m^2)$ as $m \longrightarrow \infty$, $\widetilde{F}_m(\cdot)$ is "asymptotically consistent" for F_m in the sense that $|\widetilde{F}_m(t) - F_m(t)| \longrightarrow_p 0$ for every $t \in \mathbb{R}$. This prompts possibilities for the estimation of π_0, data-adaptive determination of α for the $HT(\alpha)$ procedure, and the estimation of FDR. The first two will be developed in detail in subsequent sections. Cheng *et al.* ([7]) and Pounds and Cheng ([25]) develop smooth FDR estimators.

Let $FDR(\alpha) := E[V(\alpha)/R(\alpha)|R(\alpha) > 0]\Pr(R(\alpha) > 0)$. $ERR(\alpha)$ is essentially $FDR(\alpha)$. Under the hierarchical (or random effect) model employed in several papers ([11, 14, 29, 31]), the two quantities are equivalent, that is, $FDR(\alpha) = ERR(\alpha)$ for all $\alpha \in (0,1]$, following from Lemma 2.1 in [14]. More generally $ERR/FDR = \{E[V]/E[R]\}/E[V/R|R > 0]$ provided $\Pr(R > 0) > 0$. Asymptotically as $m \longrightarrow \infty$, if $\Pr(R > 0) \longrightarrow 1$ then $E[V/R|R > 0] \simeq E[V/R]$; if furthermore $E[V/R] \simeq E[V]/E[R]$, then $ERR/FDR \longrightarrow 1$. Identifying reasonable sufficient (and necessary) conditions for $E[V/R] \simeq E[V]/E[R]$ to hold remains an open problem at this point.

Analogous to the relationship between FDR and pFDR, define the *positive ERR*, $pERR := E[V]/E[R]$. Both quantities are well-defined provided $\Pr(R > 0) > 0$. The relationship between pERR and pFDR is the same as that between ERR and FDR described above.

The error behavior of a given multiple test procedure can be investigated in terms of either FDR (pFDR) or ERR (pERR). The ratio $pERR = E[V]/E[R]$ can be handled easily under arbitrary dependence among the tests because $E[V]$ and $E[R]$ are simply means of sums of indicator random variables. The only possible challenging component in $ERR(\alpha)$ is $\Pr(R(\alpha) > 0) = \Pr(P_{1:m} \le \alpha)$; some assumptions on the dependence among the tests has to be made to obtain a concrete analytical form for this probability, or an upper bound for it. Thus, as demonstrated in Section 4, ERR is an error measurement that is easier to handle than FDR under more complex and application-pertinent dependence among the tests, in assessing analytically the error properties of a multiple hypothesis testing procedure.

A fine technical point is that FDR (pFDR) is always well-defined and ERR (pERR) is always well-defined under the convention $a \cdot 0 = 0$ for $a \in [-\infty, +\infty]$.

Compared to FDR (pFDR), ERR (pERR) is slightly less intuitive in interpretation. For example, FDR can be interpreted as the expected proportion of false positives among all positive findings, whereas ERR can be interpreted as the proportion of the number of false positives expected out of the total number of positive findings expected. Nonetheless, ERR (pERR) is still of practical value given its close relationship to FDR (pFDR), and is more convenient to use in analytical assessments of a multiple test procedure.

3. Estimation of the proportion of null hypotheses

The proportion of the true null hypotheses π_0 is an important parameter in all multiple test procedures. A delicate component in the control or estimation of FDR (or ERR) is the estimation of π_0. The cdf $F_m(t) = \pi_0 t + (1-\pi_0)H_m(t)$, $t \in [0,1]$, along with the fact that the EDF $\widetilde{F}_m$ is its unbiased estimator provides a clue for estimating π_0. Because for any $t \in (0,1)$ $\pi_0 = [H_m(t) - F_m(t)]/[H_m(t) - t]$, a plausible estimator of π_0 is

$$\widehat{\pi}_0 = \frac{\Lambda - \widetilde{F}_m(t_0)}{\Lambda - t_0}$$

for properly chosen Λ and t_0. Let $Q_m(u) := F_m^{-1}(u)$, $u \in [0,1]$ be the quantile function of F_m and let $\widetilde{Q}_m(u) := \widetilde{F}_m^{-1}(u) := \inf\{x : \widetilde{F}_m(x) \geq u\}$ be the *empirical quantile function* (EQF), then $\pi_0 = [H_m(Q_m(u)) - u]/[H_m(Q_m(u)) - Q_m(u)]$, for $u \in (0,1)$, and with Λ_1 and u_0 properly chosen

$$\widehat{\pi}_0 = \frac{\Lambda_1 - u_0}{\Lambda_1 - \widetilde{Q}_m(u_0)}$$

is a plausible estimator. The existing π_0 estimators take either of the above representations with minor modifications.

Clearly it is necessary to have $\Lambda_1 \geq u_0$ for a meaningful estimator. Because $Q_m(u_0) \leq u_0$ by the stochastic order assumption [cf. (2.1)], choosing Λ_1 too close to u_0 will produce an estimator much biased downward. Benjamini and Hochberg ([3]) use the heuristic that if u_0 is so chosen that all P values corresponding to the alternative hypotheses concentrate in $[0, Q_m(u_0)]$ then $H_m(Q_m(u_0)) = 1$; thus setting $\Lambda_1 = 1$. Storey ([29]) uses a similar heuristic to set $\Lambda = 1$.

3.1. Existing estimators

Taking a graphical approach Schweder and Spjøtvoll [27] propose an estimator of m_0 as $\widehat{m}_0 = m(1 - \widetilde{F}_m(\lambda))/(1-\lambda)$ for a properly chosen λ; hence a corresponding estimator of π_0 is $\widehat{\pi}_0 = \widehat{m}_0/m = (1 - \widetilde{F}_m(\lambda))/(1-\lambda)$. This is exactly Storey's ([29]) estimator. Storey observes that λ is a tuning parameter that dictates the bias and variance of the estimator, and proposes computing $\widehat{\pi}_0$ on a grid of λ values, smoothing them by a spline function, and taking the smoothed $\widehat{\pi}_0$ at $\lambda = 0.95$ as the final estimator. Storey *et al.* ([31]) propose a bootstrap procedure to estimate the mean-squared error (MSE) and pick the λ that gives the minimal estimated MSE. It will be seen in the simulation study (Section 5) that this estimator tends to be biased downward.

Approaching to the problem from the quantile perspective Benjamini and Hochberg ([3]) propose $\widehat{m}_0 = \min\{1 + (m+1-j)/(1-P_{j:m}),\, m\}$ for a properly chosen j; hence

$$\widehat{\pi}_0 = \min\left\{\frac{1}{m} + \left[\frac{1-P_{j:m}}{1-j/m+1/m}\right]^{-1},\ 1\right\}.$$

The index j is determined by examining the slopes $S_i = (1-P_{i:m})/(m+1-i)$, $i = 1, \ldots, m$, and is taken to be the smallest index such that $S_j < S_{j-1}$. Then $\widehat{m}_0 = \min\{1+1/S_j,\, m\}$. It is not difficult to see why this estimator tends to be too conservative (i.e., too much biased upward): as m gets large the event $\{S_j < S_{j-1}\}$ tends to occur early (i.e., at small j) with high probability. By definition, $S_j < S_{j-1}$ if and only if

$$\frac{1-P_{j:m}}{m+1-j} < \frac{1-P_{j-1:m}}{m+2-j},$$

if and only if

$$P_{j:m} > \frac{1}{m+2-j} + \frac{m+1-j}{m+2-j} P_{j-1:m}.$$

Thus, as $m \to \infty$,

$$\Pr\left(S_j < S_{j-1}\right) = \Pr\left(P_{j:m} > \frac{1}{m+2-j} + \frac{m+1-j}{m+2-j} P_{j-1:m}\right) \longrightarrow 1,$$

for fixed or small enough j satisfying $j/m \longrightarrow \delta \in [0,1)$. The conservativeness will be further demonstrated by the simulation study in Section 5.

Recently Mosig *et al.* ([21]) proposed an estimator of m_0 by a recursive algorithm, which is clarified and shown by Nettleton and Hwang [22] to converge under a fixed partition (histogram bins) of the P value order statistics. In essence the algorithm searches in the right tail of the P value histogram to determine a "bend point" when the histogram begins to become flat, and then takes this point for λ (or j).

For a two-stage adaptive control procedure Benjamini *et al.* ([5]) consider an estimator of m_0 derived from the first-stage FDR control at the more conservative $q/(1+q)$ level than the targeted control level q. Their simulation study indicates that with comparable bias this estimator is much less variable than the estimators by Benjamini and Hochberg [3] and Storey *et al.* [31], thus possessing better accuracy. Recently Langaas *et al.* ([19]) proposed an estimator based on nonparametric estimation of the P value density function under monotone and convex contraints.

3.2. An estimator by quantile modeling

Intuitively, the stochastic order requirement in the distributional model (2.1) implies that the cdf $F_m(\cdot)$ is approximately concave and hence the quantile function $Q_m(\cdot)$ is approximately convex. When there is a substantial proportion of true null and true alternative hypotheses, there is a "bend point" $\tau_m \in (0,1)$ such that $Q_m(\cdot)$ assumes roughly a nonlinear shape in $[0, \tau_m]$, primarily dictated by the distributions of the P values corresponding to the true alternative hypotheses, and $Q_m(\cdot)$ is essentially linear in $[\tau_m, 1]$, dictated by the $U(0,1)$ distribution for the null P values. The estimation of π_0 can benefit from properly capturing this shape characteristic by a model.

Clearly $\pi_0 \le [1-\tau_m]/[H_m(Q_m(\tau_m)) - Q_m(\tau_m)]$. Again heuristically if all P values corresponding to the alternative hypotheses concentrate in $[0, Q_m(\tau_m)]$, then

$H_m(Q_m(\tau_m)) = 1$. A strategy then is to construct an estimator of $Q_m(\cdot)$, $\widehat{Q}^*_m(\cdot)$, that possesses the desirable shape described above, along with a bend point $\widehat{\tau}_m$, and set

$$\widehat{\pi}_0 = \frac{1-\widehat{\tau}_m}{1-\widehat{Q}^*_m(\widehat{\tau}_m)}, \tag{3.1}$$

which is the inverse slope between the points $(\widehat{\tau}_m, \widehat{Q}^*_m(\widehat{\tau}_m))$ and $(1,1)$ on the unit square.

Model (2.1) implies that $Q_m(\cdot)$ is twice continuously differentiable. Taylor expansion at $t=0$ gives $Q_m(t) = q_m(0)t + \frac{1}{2}q'_m(\xi_t)t^2$ for t close to 0 and $0 < \xi_t < t$, where $q_m(\cdot)$ is the first derivative of $Q_m(\cdot)$, i.e., the *quantile density function* (qdf), and $q'_m(\cdot)$ is the second derivative of $Q_m(\cdot)$. This suggests the following definition (model) of an approximation of Q_m by a convex, two-piece function joint smoothly at τ_m. Define $\underline{Q}_m(t) := \min\{Q_m(t), t\}$, $t \in [0,1]$, define the bend point $\tau_m := \text{argmax}_t\{t - \underline{Q}_m(t)\}$ and assume that it exists uniquely, with the convention that $\tau_m = 0$ if $Q_m(t) = t$ for all $t \in [0,1]$. Define

$$Q^*_m(t;\gamma,a,d,b_1,b_0,\tau_m) = \begin{cases} at^\gamma + dt, & 0 \le t \le \tau_m \\ b_0 + b_1 t, & t \ge \tau_m \end{cases} \tag{3.2}$$

where

$$\begin{aligned} b_1 &= [1 - Q_m(\tau_m)] \,/\, (1-\tau_m) \\ b_0 &= 1 - b_1 = [Q_m(\tau_m) - \tau_m] \,/\, (1-\tau_m)\,, \end{aligned}$$

and γ, a and d are determined by minimizing $\|Q^*_m(\cdot;\gamma,a,d,b_1,b_0,\tau_m) - Q_m(\cdot)\|_1$ under the following constraints:

$$\begin{cases} \gamma \ge 1, \;\; a \ge 0, \;\; 0 \le d \le 1 \\ \gamma = a = 1, \;\; d = 0 \;\text{ if and only if }\; \tau_m = 0 \\ a\tau_m^\gamma + d\tau_m = b_0 + b_1\tau_m \;\text{ (continuity at } \tau_m) \\ a\gamma\tau_m^{\gamma-1} + d = b_1 \;\text{ (smoothness at } \tau_m). \end{cases}$$

These constraints guarantee that the two pieces are joint smoothly at τ_m to produce a convex and continuously differentiable quantile function that is the closest to Q_m on $[0,1]$ in the L^1 norm, and that there is no over-parameterization if Q_m coincides with the 45-degree line. Q^*_m will be called the *convex backbone* of Q_m.

The smoothness constraints force a, d and γ to be interdependent via b_0, b_1 and τ_m. For example,

$$\begin{cases} a = a(\gamma) = -b_0 \big/ [(\gamma-1)\tau_m] \;\; (\text{for } \gamma > 1) \\ d = d(\gamma) = b_1 - a(\gamma)\gamma\tau_m^{\gamma-1}. \end{cases}$$

Thus the above constrained minimization is equivalent to

$$\begin{aligned} &\min_\gamma \;\; \|Q^*_m(\cdot;\gamma,a(\gamma),d(\gamma),b_1,b_0,\tau_m) - Q_m(\cdot)\|_1 \\ &\text{subject to } \begin{cases} \gamma \ge 1, \;\; a(\gamma) \ge 0, \;\; 0 \le d(\gamma) \le 1 \\ \gamma = a = 1, \;\; d = 0 \;\text{ if and only if }\; \tau_m = 0. \end{cases} \end{aligned} \tag{3.3}$$

An estimator of π_0 is obtained by plugging an estimator of the convex backbone Q^*_m, $\widehat{Q}^*_m$, into (3.1). The convex backbone can be estimated by replacing Q_m

with the EQF $\widetilde{Q}_m$ in the above process. However, instead of using the raw EQF, the estimation can benefit from properly smoothing the modified EQF $\underline{\widetilde{Q}}_m(t) := \min\{\widetilde{Q}_m(t), t\}$, $t \in [0,1]$ into a smooth and approximately convex EQF, $\widehat{Q}_m(\cdot)$. This smooth and approximately convex EQF can be obtained by repeatedly smoothing the modified EQF $\underline{\widetilde{Q}}_m(\cdot)$ by the variation-diminishing spline (VD-spline; de Boor [7], P.160). Denote by $B_{j,\mathbf{t},k}$ the jth order-k B spline with extended knot sequence $\mathbf{t} = t_1, \ldots, t_{n+k}$ ($t_1 = \ldots t_k = 0 < t_{k+1} < \ldots < t_n < t_{n+1} = \ldots = t_{n+k} = 1$) and $t_j^* := \sum_{\ell=j+1}^{j+k-1} t_\ell/(k-1)$. The VD-spline approximation of a function $h\colon [0,1] \to \mathbb{R}$ is defined as

$$\widehat{h}(u) := \sum_{j=1}^{n} h(t_j^*) B_{j,\mathbf{t};k}(u), \qquad u \in [0,1]. \tag{3.4}$$

The current implementation takes $k = 5$ (thus quartic spline for $\widehat{Q}_m$ and cubic spline for its derivative, $\widehat{q}_m$), and sets the interior knots in $\mathbf{t}$ to the ordered unique numbers in $\{\frac{1}{m}, \frac{2}{m}, \frac{3}{m}, \frac{4}{m}\} \cup \{\widetilde{F}_m(t),\ t = 0.001, 0.003, 0.00625, 0.01, 0.0125, 0.025, 0.05, 0.1, 0.25\}$. The knot sequence is so designed that the variation in the quantile function in a neighborhood close to zero (corresponding to small P values) can be well captured; whereas the right tail (corresponding to large P values) is greatly smoothed. Key elements in the algorithm, such as the interior knots positions, the t_j^* positions, etc., are illustrated in Figure 1.

Upon obtaining the smooth and approximately convex EQF $\widehat{Q}_m(\cdot)$, the convex backbone estimator $\widehat{Q}_m^*(\cdot)$ is constructed by replacing $Q_m(\cdot)$ with $\widehat{Q}_m(\cdot)$ in (3.3) and numerically solving the optimization with a proper search algorithm. This algorithm produces the estimator $\widehat{\pi}_0$ in (3.1) at the same time.

Note that in general the parameters γ, a, d, b_0, b_1, $\pi_0 := m_0/m$, and their corresponding estimators all depend on m. For the sake of notational simplicity this dependency has been and continues to be suppressed in the notation. Furthermore, it is assumed that $\lim_{m\to\infty} m_0/m$ exists. For studying asymptotic properties, henceforth let $\{P_1, P_2, \ldots\}$ be an infinite sequence of P values, and let $\mathbf{P}_m := \{P_1, \ldots, P_m\}$.

4. Adaptive profile information criterion

4.1. The adaptive profile information criterion

We now develop an adaptive procedure to determine a significance threshold for the $HT(\alpha)$ procedure. The estimation perspective allows one to draw an analogy between multiple hypothesis testing and the classical variable selection problem: setting $\widehat{\theta}_i = 1$ (i.e., rejecting the ith null hypothesis) corresponds to including the ith variable in the model. A traditional model selection criterion such AIC usually consists of two terms, a model-fitting term and a penalty term. The penalty term is usually some measure of model complexity reflected by the number of parameters to be estimated. In the context of massive multiple testing a natural penalty (complexity) measurement would be the expected number of false positives $E[V(\alpha)] = \pi_0 m\alpha$ under model (2.1). When a parametric model is fully specified, the model-fitting term is usually a likelihood function or some similar quantity. In the context of massive multiple testing the stochastic order assumption in model (2.1) suggests using a proper quantity measuring the lack-off-fit from $U(0,1)$ in the ensemble distribution of the P values on the interval $[0,\alpha]$. Cheng *et al.* ([7]) considered such

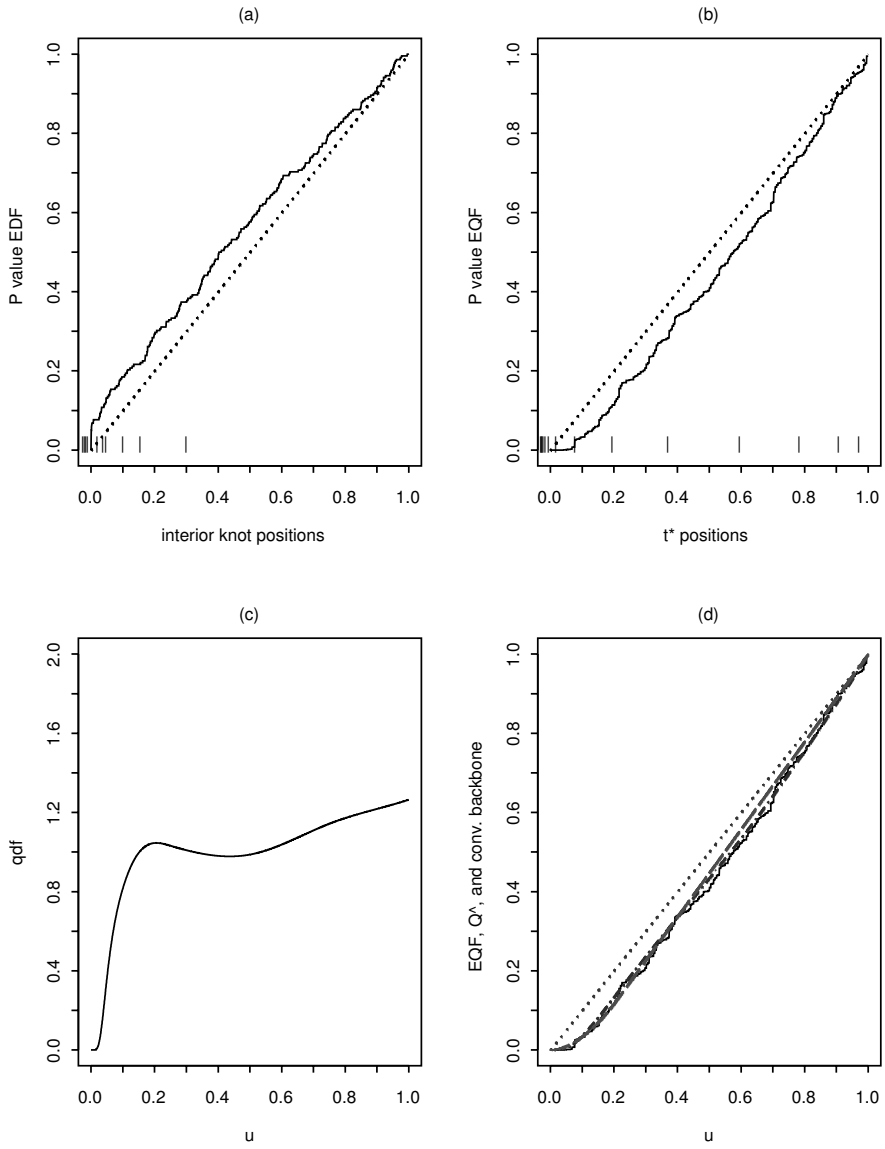


FIG 1. (a) *The interior knot positions indicated by* | *and the P value EDF;* (b) *the positions of* t_j^* *indicated by* | *and the P value EQF;* (c) $\widehat{q}_m$*: the derivative of* $\widehat{Q}_m$*;* (d) *the P value EQF (solid), the smoothed EQF* $\widehat{Q}_m$ *from Algorithm 1 (dash-dot), and the convex backbone* $\widehat{Q}_m^*$ *(long dash).*

a measurement that is an L^2 distance. The concept of convex backbone facilitates the derivation of a measurement more adaptive to the ensemble distribution of the P values. Given the convex backbone $Q_m^*(\cdot) := Q_m^*(\cdot; \gamma, a, d, b_1, b_0, \tau_m)$ as defined in (3.2), the "model-fitting" term can be defined as the L^γ distance between $Q_m^*(\cdot)$ and uniformity on $[0, \alpha]$:

$$D_\gamma(\alpha) := \left[\int_0^\alpha (t - Q_m^*(t))^\gamma \, dt \right]^{1/\gamma}, \qquad \alpha \in (0, 1].$$

The adaptivity is reflected by the use of the L^γ distance: Recall that the larger the γ, the higher concentration of small P values, and the norm inequality (Hardy, Littlewood, and Pólya [16], P.157) implies that $D_{\gamma_2}(\alpha) \geq D_{\gamma_1}(\alpha)$ for every $\alpha \in (0, 1]$ if $\gamma_2 > \gamma_1$.

Clearly $D_\gamma(\alpha)$ is non-decreasing in α. Intuitively one possibility would be to maximize a criterion like $D_\gamma(\alpha) - \lambda \pi_0 m \alpha$. However, the two terms are not on the

same order of magnitude when m is very large. The problem is circumvented by using $1/D_\gamma(\alpha)$, which also makes it possible to obtain a closed-form solution to approximately optimizing the criterion.

Thus define the *Adaptive Profile Information* (API) criterion as

$$API(\alpha) := \left[\int_0^\alpha (t - Q_m^*(t))^\gamma \, dt\right]^{-1/\gamma} + \lambda(m, \pi_0, d) m \pi_0 \alpha, \tag{4.1}$$

for $\alpha \in (0,1)$ and $Q_m^*(\cdot) := Q_m^*(\cdot; \gamma, a, d, b_1, b_0, \tau_m)$ as defined in (3.2). One seeks to minimize $API(\alpha)$ to obtain an adaptive significance threshold for the $HT(\alpha)$ procedure.

With $\gamma > 1$, the integral can be approximated by $\int_0^\alpha ((1-d)t)^\gamma dt = (1-d)(\gamma+1)^{-1}\alpha^{\gamma+1}$. Thus

$$API(\alpha) \approx \overline{API}(\alpha) := (1-d)^{-1}\left[\frac{1}{\gamma+1}\alpha^{\gamma+1}\right]^{-1/\gamma} + \lambda(m, \pi_0, d) m \pi_0 \alpha.$$

Taking the derivative of $\overline{API}(\cdot)$ and setting it to zero gives

$$\alpha^{-(2\gamma+1)/\gamma} = (1-d)(\gamma+1)^{-1/\gamma}\frac{\gamma}{\gamma+1}\lambda(m, \pi_0, d) m \pi_0.$$

Solving for α gives

$$\alpha^* = \left[\frac{(\gamma+1)^{(1+1/\gamma)}}{(1-d)\pi_0\gamma}\right]^{\gamma/(2\gamma+1)} [\lambda(m, \pi_0, d) m]^{-\gamma/(2\gamma+1)},$$

which is an approximate minimizer of API. Setting $\lambda(m, \pi_0, d) = m^{\beta\pi_0}/(1-d)$ and $\beta = 2\pi_0/\gamma$ gives

$$\alpha^* = \left[\frac{(\gamma+1)^{(1+1/\gamma)}}{\pi_0\gamma}\right]^{\gamma/(2\gamma+1)} m^{-(1+2\pi_0^2/\gamma)\gamma/(2\gamma+1)}.$$

This particular choice for λ is motivated by two facts. When most of the P values have the $U(0,1)$ distribution (equivalently, $\pi_0 \approx 1$), the d parameter of the convex backbone can be close to 1; thus with $1-d$ in the denominator, α^* can be unreasonably high in such a case. This issue is circumvented by putting $1-d$ in the denominator of λ, which eliminates $1-d$ from the denominator of α^*. Next, it is instructive to compare α^* with the Bonferroni adjustment $\alpha_{Bonf}^* = \alpha_0/m$ for a pre-specified α_0. If γ is large, then $\alpha_{Bonf}^* < \alpha^* \approx O(m^{-1/2})$ as $m \longrightarrow \infty$. Although the derivation required $\gamma > 1$, α^* is still well defined even if $\pi_0 = 1$ (implying $\gamma = 1$), and in this case $\alpha^* = 4^{1/3}m^{-1}$ is comparable to α_{Bonf}^* as $m \longrightarrow \infty$. This in fact suggests the following significance threshold calibrated with the Bonferroni adjustment:

$$\alpha_{cal}^* := 4^{-1/3}\left(\frac{\gamma}{\pi_0}\right)\alpha_0\alpha^* = A(\pi_0, \gamma) m^{-B(\pi_0,\gamma)}, \tag{4.2}$$

which coincides with the Bonferroni threshold $\alpha_0 m^{-1}$ when $\pi_0 = 1$, where

$$\begin{aligned} A(x,y) &:= \left[y/(4^{1/3}x)\right]\alpha_0\left[(y+1)^{(1+1/y)}/(xy)\right]^{y/(2y+1)} \\ B(x,y) &:= (1 + 2x^2/y)y/(2y+1). \end{aligned} \tag{4.3}$$

The factor α_0 serves asymptotically as a calibrator of the adaptive significance threshold to the Bonferroni threshold in the least favorable scenario $\pi_0 = 1$, i.e., all null hypotheses are true. Analysis of the asymptotic ERR of the $HT(\alpha^*_{cal})$ procedure suggests a few choices of α_0 in practice.

4.2. Asymptotic ERR of $HT(\alpha^*_{cal})$

Recall from (2.7) that

$$ERR(\alpha) = \big[\pi_0\alpha/F_m(\alpha)\big]\Pr(P_{1:m} \le \alpha).$$

The probability $\Pr(P_{1:m} \le \alpha)$ is not tractable in general, but an upper bound can be obtained under a reasonable assumption on the set $\mathbf{P}_m$ of the m P values. Massive multiple tests are mostly applied in exploratory studies to produce "inference-guided discoveries" that are either subject to further confirmation and validation, or helpful for developing new research hypotheses. For this reason often all the alternative hypotheses are two-sided, and hence so are the tests. It is instructive to first consider the case of m two-sample t tests. Conceptually the data consist of n_1 i.i.d. observations on $\mathbb{R}^m$ $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{im}]$, $i = 1, \ldots, n_1$ in the first group, and n_2 i.i.d. observations $\mathbf{Y}_i = [Y_{i1}, Y_{i2}, \ldots, Y_{im}]$, $i = 1, \ldots, n_2$ in the second group. The hypothesis pair (H_{0k}, H_{Ak}) is tested by the two-sided two-sample t statistic $T_k = |T(\mathcal{X}_k, \mathcal{Y}_k, n_1, n_2)|$ based on the data $\mathcal{X}_k = \{X_{1k}, \ldots, X_{n_1k}\}$ and $\mathcal{Y}_k = \{Y_{1k}, \ldots, Y_{n_2k}\}$. Often in biological applications that study gene signaling pathways (see e.g., Kuo *et al.* [18], and the simulation model in Section 5), X_{ik} and $X_{ik'}$ $(i = 1, \ldots, n_1)$ are either positively or negatively correlated for certain $k \ne k'$, and the same holds for Y_{ik} and $Y_{ik'}$ $(i = 1, \ldots, n_2)$. Such dependence in data raises positive association between the two-sided test statistics T_k and $T_{k'}$ so that $\Pr(T_k \le t | T_{k'} \le t) \ge \Pr(T_k \le t)$, implying $\Pr(T_k \le t,\ T_{k'} \le t) \ge \Pr(T_k \le t)\Pr(T_{k'} \le t)$, $t \ge 0$. Then the P values in turn satisfy $\Pr(P_k > \alpha,\ P_{k'} > \alpha) \ge \Pr(P_k > \alpha)\Pr(P_{k'} > \alpha)$, $\alpha \in [0, 1]$. It is straightforward to generalize this type of dependency to more than two tests. Alternatively, a direct model for the P values can be constructed.

Example 4.1. Let $\mathcal{J} \subseteq \{1, \ldots, m\}$ be a nonempty set of indices. Assume $P_j = P_0^{X_j}$, $j \in \mathcal{J}$, where P_0 follows a distribution F_0 on $[0, 1]$, and X_j's are i.i.d. continuous random variables following a distribution H on $[0, \infty)$, and are independent of the P values. Assume that the P_i's for $i \notin \mathcal{J}$ are either independent or related to each other in the same fashion. This model mimics the effect of an activated gene signaling pathway that results in gene differential expression as reflected by the P values: the set $\mathcal{J}$ represents the genes involved in the pathway, P_0 represents the underlying activation mechanism, and X_j represents the noisy response of gene j resulting in P_j. Because $P_i > \alpha$ if and only if $X_j < \log\alpha/\log P_0$, direct calculations using independence of the X_j's show that

$$\Pr\left(\bigcap_{j\in\mathcal{J}}\{P_j > \alpha\}\right) = \int_0^1 \Pr\left(\bigcap_{j\in\mathcal{J}}\left\{X_j < \frac{\log\alpha}{\log t}\right\}\right) dF_0(t) = E\left[\left[H\left(\frac{\log\alpha}{\log P_0}\right)\right]^{|\mathcal{J}|}\right],$$

where $|\mathcal{J}|$ is the cardinality $\mathcal{J}$. Next

$$\prod_{j\in\mathcal{J}} \Pr\left(P_j > \alpha\right) = \prod_{j\in\mathcal{J}} \int_0^1 \left[H\left(\frac{\log\alpha}{\log t}\right)\right] dF_0(t) = \left[E\left[H\left(\frac{\log\alpha}{\log P_0}\right)\right]\right]^{|\mathcal{J}|}.$$

Finally $\Pr(\cap_{j\in\mathcal{J}}\{P_j > \alpha\}) \geq \prod_{j\in\mathcal{J}} \Pr(P_j > \alpha)$, following from Jensen's inequality.

The above considerations lead to the following definition.

Definition 4.1. The set of P values $\mathbf{P}_m$ has the *positive orthant dependence property* if for any $\alpha \in [0,1]$

$$\Pr\left(\bigcap_{i=1}^{m}\{P_i > \alpha\}\right) \geq \prod_{i=1}^{m} \Pr(P_i > \alpha).$$

This type of dependence is similar to the positive quadrant dependence introduced by Lehmann [20].

Now define the upper envelope of the cdf's of the P values as

$$\overline{F}_m(t) := \max_{i=1,\ldots,m}\{G_i(t)\}, \qquad t \in [0,1],$$

where G_i is the cdf of P_i. If $\mathbf{P}_m$ has the positive orthant dependence property then

$$\Pr(P_{1:m} \leq \alpha) = 1 - \Pr\left(\bigcap_{i=1}^{m}\{P_i > \alpha\}\right) \leq 1 - \prod_{i=1}^{m} \Pr(P_i > \alpha) \leq 1 - (1 - \overline{F}_m(\alpha))^m,$$

implying

$$ERR(\alpha^*_{cal}) \leq \frac{\pi_0 \alpha^*_{cal}}{\pi_0 \alpha^*_{cal} + (-\pi_0) H_m(\alpha^*_{cal})} \left[1 - (1 - \overline{F}_m(\alpha^*_{cal}))^m\right]. \tag{4.4}$$

Because $\alpha^*_{cal} \longrightarrow 0$ as $m \longrightarrow \infty$, the asymptotic magnitude of the above ERR can be established by considering the magnitude of $\overline{F}_m(t_m)$ and $H_m(t_m)$ as $t_m \longrightarrow 0$. The following definition makes this idea rigorous.

Definition 4.2. The set of m P values $\mathbf{P}_m$ is said to be *asymptotically stable* as $m \longrightarrow \infty$ if there exists sequences $\{\beta_m\}$, $\{\eta_m\}$, $\{\psi_m\}$, $\{\xi_m\}$ and constants β^*, β_*, η, ψ^*, ψ_*, and ξ such that

$$\overline{F}_m(t) \simeq \beta_m t^{\eta_m}, \quad H_m(t) \simeq \psi_m t^{\xi_m}, \quad t \longrightarrow 0$$

and

$$\begin{aligned} &0 < \beta_* \leq \beta_m \leq \beta^* < \infty, \quad 0 < \eta \leq \eta_m \leq 1 \\ &0 < \psi_* \leq \psi_m \leq \psi^* < \infty, \quad 0 < \xi \leq \xi_m \leq 1 \end{aligned}$$

for sufficiently large m.

This definition essentially says that $\mathbf{P}_m$ is regarded as asymptotically stable if the ensemble distribution functions $F_m(\cdot)$ and $H_m(\cdot)$ vary in the left tail similarly to Beta distributions.

The following theorem establishes the asymptotic magnitude of an upper bound of $ERR(\alpha^*_{cal})$ – the ERR of applying α^*_{cal} in the hard-thresholding procedure (2.3).

Theorem 4.1. *Let ψ_*, ξ_m, β^*, and η be as given in Definition 4.2, and let $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ be as defined in (4.3). If the set of P values $\mathbf{P}_m$ is asymptotically stable and has the positive orthant dependence property for sufficiently large m, then $ERR(\alpha^*_{cal}) \leq \Psi(\alpha^*_{cal})$, and $\Psi(\alpha^*_{cal})$ satisfies*

(a) *if $\pi_0 = 1$ for all m, $\lim_{m\to\infty} \Psi(\alpha^*_{cal}) = 1 - e^{-\alpha_0}$;*

(b) *if $\pi_0 < 1$ and $A(\pi_0,\gamma) \leq \overline{A} < \infty$ for some $\overline{A}$ and sufficiently large m, then*

$$\Psi(\alpha^*_{cal}) \simeq \frac{\pi_0}{1-\pi_0} \psi_*^{-1} \left[A(\pi_0,\gamma)\right]^{1-\xi_m} m^{-(1-\xi_m)B(\pi_0,\gamma)}, \quad as \quad m \longrightarrow \infty.$$

Proof. See Appendix. □

There are two important consequences from this theorem. First, the level α_0 can be chosen to bound ERR (and FDR) asymptotically in the least favorable situation $\pi_0 = 1$. In this case both ERR and FDR are equal to the family-wise type-I error probability. Note that $1-e^{-\alpha_0}$ is also the limiting family-wise type-I error probability corresponding to the Bonferroni significance threshold $\alpha_0 m^{-1}$. In this regard the adaptive threshold α^*_{cal} is calibrated to the conservative Bonferroni threshold when $\pi_0 = 1$. If one wants to bound the error level at α_1, then set $\alpha_0 = -\log(1-\alpha_1)$. Of course $\alpha_0 \approx \alpha_1$ for small α_1; for example, $\alpha_0 \approx 0.05129, 0.1054, 0.2231$ for $\alpha_1 = 0.05, 0.1, 0.2$ respectively.

Next, Part (b) demonstrates that if the "average power" of rejecting the false null hypotheses remains visible asymptotically in the sense that $\xi_m \leq \overline{\xi} < 1$ for some $\overline{\xi}$ and sufficiently large m, then the upper bound

$$\Psi(\alpha^*_{cal}) \simeq \frac{\pi_0}{1-\pi_0}\psi_*^{-1}\left[A(\pi_0,\gamma)\right]^{1-\xi_m} m^{-(1-\overline{\xi})B(\pi_0,\gamma)} \longrightarrow 0;$$

therefore $ERR(\alpha^*_{cal})$ diminishes asymptotically. However, the convergence can be slow if the power is weak in the sense $\overline{\xi} \approx 1$ (hence $H_m(\cdot)$ is close to the $U(0,1)$ cdf in the left tail). Moreover, Ψ can be considerably close to 1 in the unfavorable scenario $\pi_0 \approx 1$ and $\overline{\xi} \approx 1$. On the other hand, increase in the average power in the sense of decrease in $\overline{\xi}$ makes Ψ (hence the ERR) diminishes faster asymptotically.

Note from (4.3) that as long as π_0 is bounded away from zero (i.e., there is always some null hypotheses remain true) and γ is bounded, the quantity $A(\pi_0,\gamma)$ is bounded. Because the positive ERR does not involve the probability $\Pr(R>0)$, part (b) holds for $pERR(\alpha^*_{cal})$ under *arbitrary* dependence among the P values (tests).

4.3. Data-driven adaptive significance threshold

Plugging $\widehat{\pi}_0$ and $\widehat{\gamma}$ generated by optimizing (3.3) into (4.2) produces a data-driven significance threshold:

$$\widehat{\alpha}^*_{cal} := A(\widehat{\pi}_0,\widehat{\gamma})m^{-B\left(\widehat{\pi}_0,\widehat{\gamma}\right)}. \tag{4.5}$$

Now consider the ERR of the procedure $HT(\widehat{\alpha}^*_{cal})$ with $\widehat{\alpha}^*_{cal}$ as above. Define

$$ERR^* := \frac{E\left[V(\widehat{\alpha}^*_{cal})\right]}{E\left[R(\widehat{\alpha}^*_{cal})\right]}\Pr\left(R(\widehat{\alpha}^*_{cal})>0\right).$$

The interest here is the asymptotic magnitude of ERR^* as $m \longrightarrow \infty$. A major difference here from Theorem 4.1 is that the threshold $\widehat{\alpha}^*_{cal}$ is random. A similar result can be established with some moment assumptions on $A(\widehat{\pi}_0,\widehat{\gamma})$, where $A(\cdot,\cdot)$ is defined in (4.3) and $\widehat{\pi}_0$, $\widehat{\gamma}$ are generated by optimizing (3.3). Toward this end, still assume that $\mathbf{P}_m$ is asymptotically stable, and let η_m, η, and ξ_m be as in Definition 4.2. Let ν_m be the joint cdf of $[\widehat{\pi}_0,\widehat{\gamma}]$, and let

$$a_m := \int_{\mathbb{R}^2} A(s,t)^{\eta_m} d\nu_m(s,t)$$

$$a_{1m} := \int_{\mathbb{R}^2} A(s,t) d\nu_m(s,t)$$

$$a_{2m} := \int_{\mathbb{R}^2} A(s,t)^{\xi_m} d\nu_m(s,t).$$

All these moments exist as long as $\widehat{\pi}_0$ is bounded away from zero and $\widehat{\gamma}$ is bounded with probability one.

Theorem 4.2. *Suppose that* $\mathbf{P}_m$ *is asymptotically stable and has the positive orthant dependence property for sufficiently large* m*. Let* β^**,* η*,* ψ_**, and* ξ_m *be as in Definition* 4.2*. If* a_m*,* a_{1m} *and* a_{2m} *all exist for sufficiently large* m*, then* $ERR^* \le \Psi_m$ *and there exist* $\delta_m \in [\eta/3, \eta]$*,* $\varepsilon_m \in [1/3, 1]$*, and* $\varepsilon'_m \in [\xi_m/3, \xi_m]$ *such that as* $m \longrightarrow \infty$

$$\Psi_m \simeq \begin{cases} K(\beta^*, a_m, \delta_m), & \text{if } \pi_0 = 1, \text{ all } m \\ \frac{\pi_0}{1-\pi_0}\left(\frac{a_{1m}}{a_{2m}}\right)\psi_*^{-1}\frac{K(\beta^*, a_m, \delta_m)}{m^{(\varepsilon_m - \varepsilon'_m)}}, & \text{if } \pi_0 < 1, \text{ sufficiently large } m, \end{cases}$$

where $K(\beta^*, a_m, \delta_m) = 1 - \left(1 - \beta^* a_m m^{-\delta_m}\right)^m$

Proof. See Appendix. □

Although less specific than Theorem 4.1, this result still is instructive. First, if the "average power" sustains asymptotically in the sense that $\xi_m < 1/3$ so that $\varepsilon_m > \varepsilon'_m$ for sufficiently large m, or if $\lim_{m\to\infty} \xi_m = \overline{\xi} < 1/3$, then ERR^* diminishes as $m \longrightarrow \infty$. The asymptotic behavior of ERR^* in the case of $\overline{\xi} \ge 1/3$ is indefinite from this theorem, and obtaining a more detailed upper bound for ERR^* in this case remains an open problem. Next, ERR^* can be potentially high if $\pi_0 = 1$ always or $\pi_0 \approx 1$ and the average power is weak asymptotically. The reduced specificity in this result compared to Theorem 4.1 is due to the random variations in $A(\widehat{\pi}_0, \widehat{\gamma})$ and $B(\widehat{\pi}_0, \widehat{\gamma})$, which are now random variables instead of deterministic functions. Nonetheless Theorem 4.2 and its proof (see Appendix) do indicate that when $\pi_0 \approx 1$ and the average power is weak (i.e., $H_m(\cdot)$ is small), for sake of ERR (and FDR) reduction the variability in $A(\widehat{\pi}_0, \widehat{\gamma})$ and $B(\widehat{\pi}_0, \widehat{\gamma})$ should be reduced as much as possible in a way to make δ_m and ε_m as close to 1 as possible. In practice one should make an effort to help this by setting $\widehat{\pi}_0$ and $\widehat{\gamma}$ to 1 when the smoothed empirical quantile function $\widehat{Q}_m$ is too close to the $U(0,1)$ quantile function. On the other hand, one would like to have a reasonable level of false negative errors when true alternative hypotheses do exist even if $\pi_0 \approx 1$; this can be helped by setting α_0 at a reasonably liberal level. The simulation study (Section 5) indicates that $\alpha_0 = 0.22$ is a good choice in a wide variety of scenarios.

Finally note that, just like in Theorem 4.1, the bound when $\pi_0 < 1$ holds for the positive ERR $pERR^* := E[V(\widehat{\alpha}^*_{cal})]/E[R(\widehat{\alpha}^*_{cal})]$ under *arbitrary* dependence among the tests.

5. A Simulation study

To better understand and compare the performance and operating characteristics of $HT(\widehat{\alpha}^*_{cal})$, a simulation study is performed using models that mimic a gene signaling pathway to generate data, as proposed in [7]. Each simulation model is built from a network of 7 related "genes" (random variables), X_0, X_1, X_2, X_3, X_4, X_{190}, and X_{221}, as depicted in Figure 2, where X_0 is a latent variable. A number of other variables are linear functions of these random variables.

Ten models (scenarios) are simulated. In each model there are m random variables, each observed in K groups with n_k independent observations in the kth group ($k = 1, \ldots K$). Let μ_{ik} be the mean of variable i in group k. Then m ANOVA hypotheses, one for each variable (H_{0i}: $\mu_{i1} = \cdots = \mu_{iK}$, $i = 1 \ldots, m$), are tested.

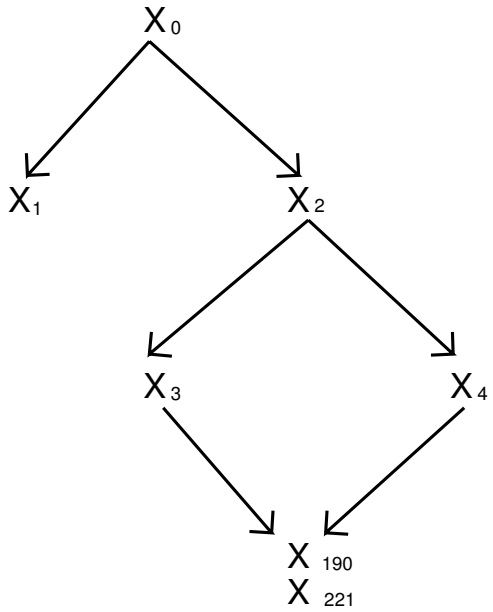


FIG 2. *A seven-variable framework to simulate differential gene expressions in a pathway.*

TABLE 2
Relationships among $X_0, X_1, \ldots, X_4$, X_{190} and X_{221}: X_{ikj} denote the jth observation of the ith variable in group k; $N(0, \sigma^2)$ denotes normal random noise. The index j always runs through 1,2,3

X_{01j} i.i.d. $N(0, \sigma^2)$; X_{0kj} i.i.d. $N(8, \sigma^2)$, $k = 2, 3, 4$
$X_{1kj} = X_{0kj}/4 + N(0, 0.0784)$ (X_1 is highly correlated with X_0; $\sigma = 0.28$.)
$X_{2kj} = X_{0kj} + N(0, \sigma^2)$, $k = 1, 2$; $X_{23j} = X_{03j} + 6 + N(0, \sigma^2)$; $X_{24j} = X_{04j} + 14 + N(0, \sigma^2)$
$X_{3kj} = X_{2kj} + N(0, \sigma^2)$, $k = 1, 2, 3, 4$
$X_{4kj} = X_{2kj} + N(0, \sigma^2)$, $k = 1, 2$; $X_{43j} = X_{23j} - 6 + N(0, \sigma^2)$; $X_{44j} = X_{24j} - 8 + N(0, \sigma^2)$
$X_{190,1j} = X_{31j} + 24 + N(0, \sigma^2)$; $X_{190,2j} = X_{32j} + X_{42j} + N(0, \sigma^2)$;
$X_{190,3j} = X_{33j} - X_{43j} - 6 + N(0, \sigma^2)$; $X_{190,4j} = X_{34j} - 14 + N(0, \sigma^2)$
$X_{221,kj} = X_{3kj} + 24 + N(0, \sigma^2)$, $k = 1, 2$;
$X_{221,3j} = X_{33j} - X_{43j} + N(0, \sigma^2)$; $X_{221,4j} = X_{34j} + 2 + N(0, \sigma^2)$

Realizations are drawn from Normal distributions. For all ten models the number of groups $K = 4$ and the sample size $n_k = 3$, $k = 1, 2, 3, 4$. The usual one-way ANOVA F test is used to calculate P values. Table 2 contains a detailed description of the joint distribution of $X_0, \ldots, X_4$, X_{190} and X_{221} in the ANOVA set up. The ten models comprised of different combinations of m, π_0, and the noise level σ are detailed in Table 3, Appendix. The odd numbered models represent the high-noise (thus weak power) scenario and the even numbered models represent the low-noise (thus substantial power) scenario. In each model variables not mentioned in the table are i.i.d. $N(0, \sigma^2)$. Performance statistics under each model are calculated from 1,000 simulation runs.

First, the π_0 estimators by Benjamini and Hochberg [3], Storey *et al.* [31], and (3.1) are compared on several models. Root mean square error (MSE) and bias are plotted in Figure 3. In all cases the root MSE of the estimator (3.1) is either the smallest or comparable to the smallest. In the high noise case ($\sigma = 3$) Benjamini and Hochberg's estimator tends to be quite conservative (upward biased), especially for relatively low true π_0 (0.83 and 0.92, Models 1 and 3); whereas Storey's estimator is biased downward slightly in all cases. The proposed estimator (3.1) is biased in the conservative direction, but is less conservative than Benjamini and Hochberg's estimator. In the low noise case ($\sigma = 1$) the root MSE of all three estimators

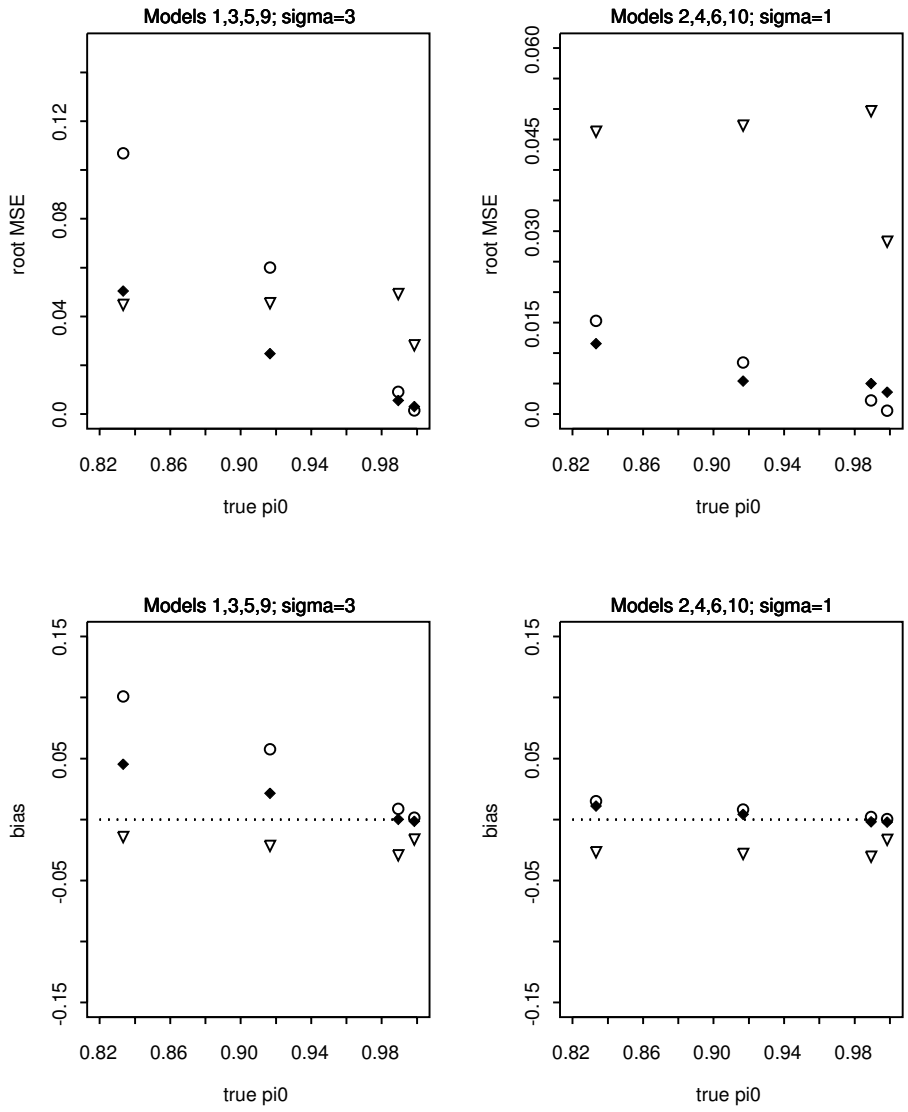


FIG 3. *Root MSE and bias of the π_0 estimators by Benjamini and Hochberg* [3] *(circle), Storey et al.* [31] *(triangle), and* (3.1) *(diamond)*

and the bias of the proposed and the Benjamini and Hochberg's estimators are reduced substantially while the small downward bias of Storey's bootstrap estimator remains. Overall the proposed estimator (3.1) outperforms the other two estimators in terms of MSE and bias.

Next, operating characteristics of the adaptive FDR control ([3]) and q-value FDR control ([31]) at the 1%, 5%, 10%, 15%, 20%, 30%, 40%, 60%, and 70% levels, the criteria API (i.e., the $HT(\widehat{\alpha}^*_{cal})$ procedure) and I_p ([7]), are simulated and compared. The performance measures are the estimated FDR ($\widehat{FDR}$) and the estimated false nondiscovery proportion ($\widehat{FNDP}$) defined as follows. Let m_1 be the number of true alternative hypotheses according to the simulation model, let R_l be the total number of rejections in simulation trial l, and let S_l be the number of correct rejections. Define

$$\widehat{FDR} = \tfrac{1}{1000}\textstyle\sum_{l=1}^{1000} I(R_l > 0)(R_l - S_l)/R_l$$
$$\widehat{FNDP} = \tfrac{1}{1000}\textstyle\sum_{l=1}^{1000}(m_1 - S_l)/m_1,$$

where $I(\cdot)$ is the indicator function. These are the Monte Carlo estimators of the

FDR and the $FNDP := E\left[m_1 - S\right]/m_1$ (cf. Table 1). In other words FNDP is the expected proportion of true alternative hypotheses not captured by the procedure. A measurement of the average power is $1 - FNDP$.

Following the discussions in Section 4, the parameter α_0 required in the API procedure should be set at a reasonably liberal level. A few values of α_0 were examined in a preliminary simulation study, which suggested that $\alpha_0 = 0.22$ is a level that worked well for the variety of scenarios covered by the ten models in Table 3, Appendix.

Results corresponding to $\alpha_0 = 0.22$ are reported here. The results are first summarized in Figure 4. In the high noise case ($\sigma = 3$, Models 1, 3, 5, 7, 9), compared to I_p, API incurs no or little increase in FNDP but substantially lower FDR when π_0 is high (Models 5, 7, 9), and keeps the same FDR level and a slightly reduced FNDP when π_0 is relatively low (Models 1, 3); thus API is more adaptive than I_p. As expected, it is difficult for all methods to have substantial power (low FNDP) in the high noise case, primarily due to the low power in each individual test to reject a false null hypothesis. For the FDR control procedures, no substantial number of false null hypotheses can be rejected unless the FDR control level is raised to a relatively high level of $\geq 30\%$, especially when π_0 is high.

In the low noise case ($\sigma = 1$, Models 2, 4, 6, 8, 10), API performs similarly to I_p, although it is slightly more liberal in terms of higher FDR and lower FNDP when π_0 is relatively low (Models 2, 4). Interestingly, when π_0 is high (Models 6, 8, 10), FDR control by q-value (Storey *et al.* [31]) is less powerful than the adaptive FDR procedure (Benjamini and Hochberg [3]) at low FDR control levels (1%, 5%, and 10%), in terms of elevated FNDP levels.

The methods are further compared by plotting $\widehat{FNDP}$ vs. $\widehat{FDR}$ for each model in Figure 5. The results demonstrate that in low-noise (model 2, 4, 6, 8, 10) and high-noise, high-π_0 (models 5, 7, 9) cases, the adaptive significance threshold determined from API gives very reasonable balance between the amounts of false positive and false negative errors, as indicated by the position of the diamond ($\widehat{FNDP}$ vs. $\widehat{FDR}$ of API) relative to the curves of the FDR-control procedures. It is noticeable that in the low noise cases the adaptive significance threshold corresponds well to the maximum FDR level for which there is no longer substantial gain in reducing FNDP by controlling the FDR at higher levels. There is some loss of efficiency for using API in high-noise, low-π_0 cases (model 1, 3) – its FNDP is higher than the control procedures at comparable FDR levels. This is a price to pay for not using a prespecified, fixed FDR control level.

The simulation results on API are very consistent with the theoretical results in Section 4. They indicate that API can provide a reasonable, data-adaptive significance threshold that balances the amounts of false positive and false negative errors: it is reasonably conservative in the high π_0 and high noise (hence low power) cases, and is reasonably liberal in the relatively low π_0 and low noise cases.

6. Concluding remarks

In this research an improved estimator of the null proportion and an adaptive significance threshold criterion API for massive multiple tests are developed and studied, following the introduction of a new measurement of the level of false positive errors, ERR, as an alternative to FDR for theoretical investigation. ERR allows for obtaining insights into the error behavior of API under more application-pertinent distributional assumptions that are widely satisfied by the data in many recent

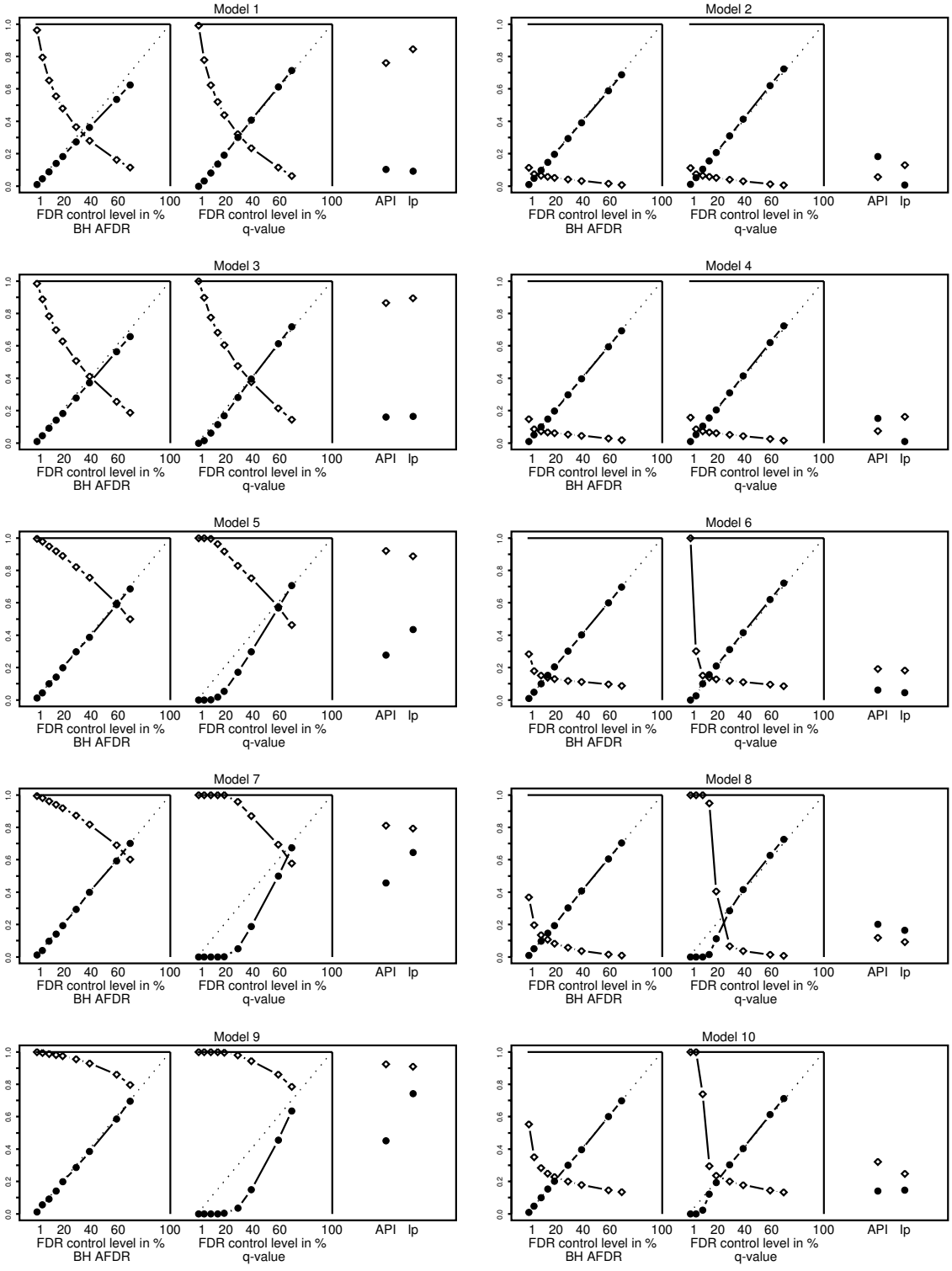


FIG 4. *Simulation results on the rejection criteria. Each panel corresponds to a model configuration. Panels in the left column correspond to the "high noise" case* $\sigma = 3$*, and panels in the right column correspond to the "low noise" case* $\sigma = 1$*. The performance statistics* $\widehat{FDR}$ *(bullet) and* $\widehat{FNDP}$ *(diamond) are plotted against each criteria. Each panel has three sections. The left section shows FDR control with the Benjamini & Hochberg* [3] *adaptive procedure (BH AFDR), and the middle section shows FDR control by q-value, all at the* 1%, 5%, 10%, 15%, 20%, 30%, 40%, 60%, *and* 70% *levels. The right section shows* $\widehat{FDR}$ *and* $\widehat{FNDP}$ *of API and* I_p*.*

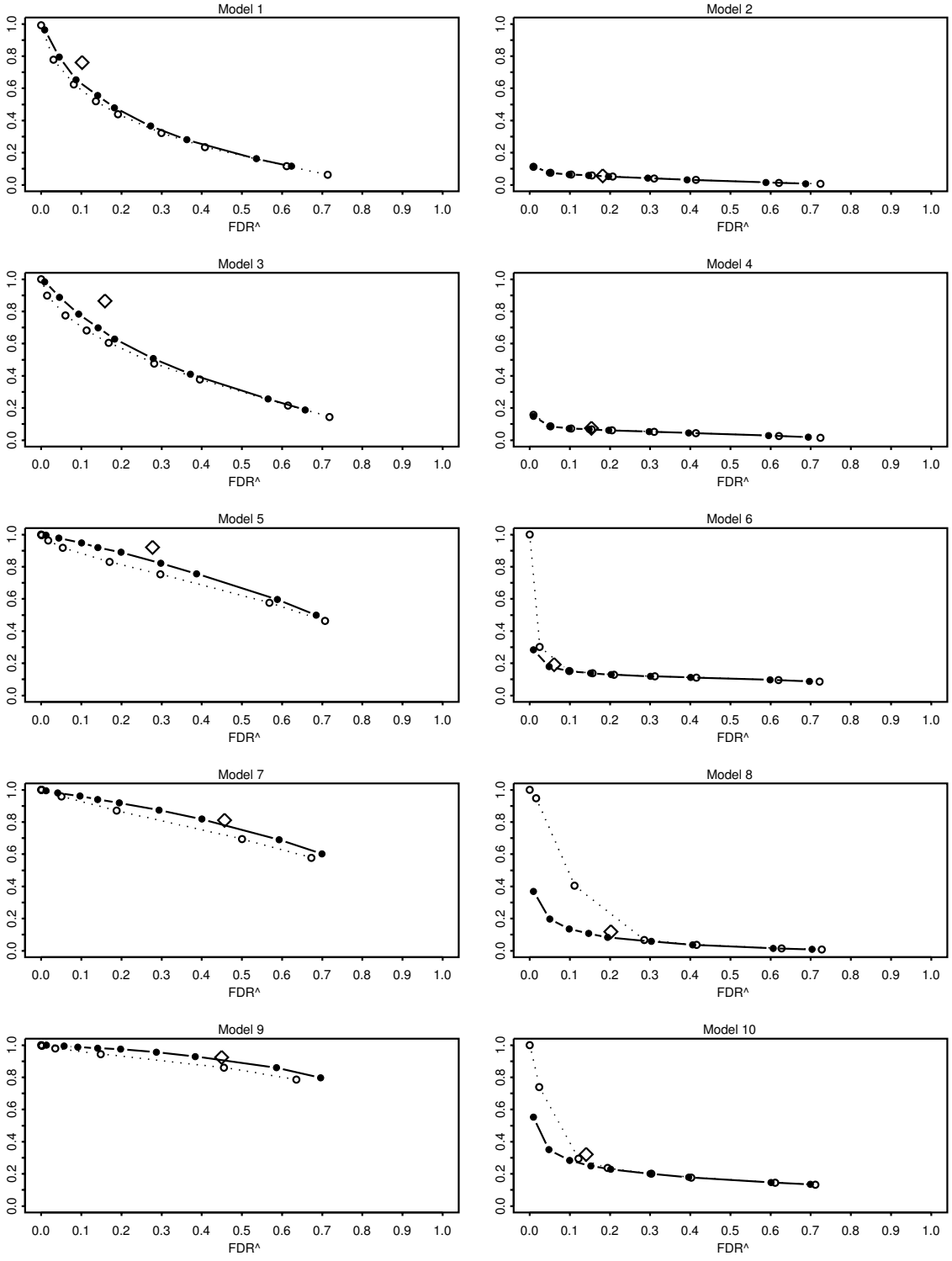


FIG 5. *$\widehat{FNDP}$ vs. $\widehat{FDR}$ for Benjamini and Hochberg* [3] *adaptive FDR control (solid line and bullet) and q-value FDR control (dotted line and circle) when FDR control levels are set at* 1%, 5%, 10%, 15%, 20%, 30%, 40%, 60%, *and* 70%. *For each model $\widehat{FNDP}$ vs. $\widehat{FDR}$ of the adaptive API procedure occupies one point on the plot, indicated by a diamond.*

applications. Under these assumptions, for the first time the asymptotic ERR level (and the FDR level under certain conditions) is explicitely related to the ensemble behavior of the P values described by the upper envelope cdf $\overline{F}_m$ and the "average power" H_m. Parallel to positive FDR, the concept of positive ERR is also useful. Asymptotic pERR properties of the proposed adaptive method can be established under arbitrary dependence among the tests. The theoretical understanding provides cautions and remedies to the application of API in practice.

Under proper ergodicity conditions such as those used in [31, 14], FDR and ERR are equivalent for the hard-thresholding procedure (2.3); hence Theorems 4.1 and 4.2 hold for FDR as well.

The simulation study shows that the proposed estimator of the null proportion by quantile modeling is superior to the two popular estimators in terms of reduced MSE and bias. Not surprisingly, when there is little power to reject each individual false null hypothesis (hence little average power), FDR control and API both incur high level of false negative errors in terms of FNDP. When there is a reasonable amount of power, API can produce a reasonable balance between the false positives and false negatives, thereby complementing and extending the widely used FDR-control approach to massive multiple tests.

In exploratory type applications where it is desirable to provide "inference-guided discoveries", the role of α_0 is to provide a protection in the situation where no true alternative hypothesis exits ($\pi_0 = 1$). On the other hand it is not advisable to choose the significance threshold too conservatively in such applications because the "discoveries" will be scrutinized in follow up investigations. Even if setting $\alpha_0 = 1$ the calibrated adaptive significance threshold is m^{-1}, giving the limiting ERR (or FDR, or family-wise type-I error probability) $1 - e^{-1} \approx 0.6321$ when $\pi_0 = 1$.

At least two open problems remain. First, although there has been empirical evidence from the simulation study that the π_0 estimator (3.1) outperforms the existing ones, there is lack of analytical understating of this estimator, in terms of MSE for example. Second, the bounds obtained in Theorems 4.1 and 4.2 are not sharp, a more detailed characterization of the upper bound of ERR^* (Theorem 4.2) is desirable for further understanding of the asymptotic behavior of the adaptive procedure.

Appendix

Proof of Theorem 4.1. For (a), from (4.3) and (4.4), if $\pi_0 = 1$ for all m, then the first factor on the right-hand side of (4.4) is 1, and the second factor is now equal to

$$1 - (1 - \alpha^*_{cal})^m = 1 - (1 - A(1,1)m^{-B(1,1)})^m = 1 - (1 - \alpha_0 m^{-1})^m \longrightarrow 1 - e^{-\alpha_0}$$

because $A(1,1) = \alpha_0$ and $B(1,1) = 1$. For (b), first

$$1 - (1 - \overline{F}_m(\alpha^*_{cal}))^m \simeq 1 - (1 - \beta_m \alpha^{*\eta_m}_{cal})^m \leq 1 - (1 - \beta^* \overline{A} \alpha_0 m^{-\eta B(\pi_0,\gamma)})^m := \varepsilon_m,$$

and $\varepsilon_m \simeq 1 - \exp\left(-\beta^* \overline{A} \alpha_0 m^{1-\eta B(\pi_0,\gamma)}\right) \longrightarrow 1$ because $B(\pi_0,\gamma) \leq (\gamma+2)/(2\gamma+1) < 1$ so that $\eta B(\pi_0,\gamma) < 1$. Next, let $\omega_m := \psi_m^{-1} \frac{[A(\pi_0,\gamma)]^{1-\xi_m}}{m^{(1-\xi_m)B(\pi_0,\gamma)}}$.

Then

$$\frac{\pi_0\alpha^*_{cal}}{\pi_0\alpha^*_{cal}+(1-\pi_0)H_m(\alpha^*_{cal})} \simeq \frac{\pi_0\omega_m}{1-\pi_0} \leq \frac{\pi_0}{1-\pi_0}\psi_*^{-1}\frac{[A(\pi_0,\gamma)]^{1-\xi_m}}{m^{(1-\xi_m)B(\pi_0,\gamma)}}$$

for sufficiently large m. Multiplying this upper bound and the limit of ε_m gives (b). □

Proof of Theorem 4.2. First, for sufficiently large m,

$$\begin{aligned}\Pr\left(R(\widehat{\alpha}^*_{cal})>0\right) &\leq 1-\int_{\mathbb{R}^2}\left[\beta_m A(s,t)^{\eta_m}m^{-\eta_m B(s,t)}\right]^m d\nu_m(s,t)\\ &\leq 1-\left[1-\beta^*\int_{\mathbb{R}^2}A(s,t)^{\eta_m}m^{-\eta B(s,t)}d\nu_m(s,t)\right]^m\end{aligned}$$

Because $1/3 \leq B(\widehat{\pi}_0,\widehat{\gamma}) \leq 1$ with probability 1, so that $m^{-\eta} \leq m^{\eta B\left(\widehat{\pi}_0,\widehat{\gamma}\right)} \leq m^{-\eta/3}$ with probability 1, by the mean value theorem of integration (Halmos [15], P.114), there exists some $\mathcal{E}_m \in [m^{-\eta}, m^{-\eta/3}]$ such that

$$\int_{\mathbb{R}^2}A(s,t)^{\eta_m}m^{-\eta B(s,t)}d\nu_m(s,t)=\mathcal{E}_m a_m,$$

and $\mathcal{E}_m$ can be written equivalently as $m^{-\delta_m}$ for some $\delta_m \in [\eta/3,\eta]$, giving, for sufficiently large m,

$$\Pr\left(R(\widehat{\alpha}^*_{cal})>0\right) \leq 1-\left(1-\beta^* a_m m^{-\delta_m}\right)^m.$$

This is the upper bound Ψ_m of ERR^* if $\pi_0=1$ for all m because now $V(\widehat{\alpha}^*_{cal}) = R(\widehat{\alpha}^*_{cal})$ with probability 1. Next,

$$E\left[V(\widehat{\alpha}^*_{cal})\right]=E\left[E\left[V(\widehat{\alpha}^*_{cal})\middle|\widehat{\alpha}^*_{cal}\right]\right]=E\left[\pi_0\widehat{\alpha}^*_{cal}\right]=\pi_0\int_{\mathbb{R}^2}\frac{A(s,t)}{m^{B(s,t)}}d\nu_m(s,t).$$

Again by the mean value theorem of integration there exists $\varepsilon_m \in [1/3,1]$ such that $E\left[V(\widehat{\alpha}^*_{cal})\right]=\pi_0 a_{1m}m^{-\varepsilon_m}$. Similarly,

$$E\left[H_m(\widehat{\alpha}^*_{cal})\right] \simeq \psi_m\int_{\mathbb{R}^2}A(s,t)^{\xi_m}m^{-\xi_m B(s,t)}d\nu_m(s,t) \geq \psi_* a_{2m}m^{-\varepsilon'_m}$$

for some $\varepsilon'_m \in [\xi_m/3,\xi_m]$. Finally, because

$$E\left[R(\widehat{\alpha}^*_{cal})\right]=E\left[E\left[R(\widehat{\alpha}^*_{cal})\middle|\widehat{\alpha}^*_{cal}\right]\right]=\pi_0 E\left[V(\widehat{\alpha}^*_{cal})\right]+(1-\pi_0)E\left[H_m(\widehat{\alpha}^*_{cal})\right],$$

if $\pi_0<1$ for sufficiently large m, then

$$\begin{aligned}ERR^* &\leq \left[1-\left(1-\beta^* a_m m^{-\delta_m}\right)^m\right]\frac{\pi_0 E\left[\widehat{\alpha}^*_{cal}\right]}{(1-\pi_0)E\left[H_m(\widehat{\alpha}^*_{cal})\right]}\\ &\leq \frac{\pi_0}{1-\pi_0}\left(\frac{a_{1m}}{a_{2m}}\right)\psi_*^{-1}m^{-(\varepsilon_m-\varepsilon'_m)}\left[1-\left(1-\beta^* a_m m^{-\delta_m}\right)^m\right].\end{aligned}$$

□

TABLE 3

Ten models: Model configuration in terms of (m, m_1, σ) and determination of true alternative hypotheses by $X_1, \ldots, X_4$, X_{190} and X_{221}, where m_1 is the number of true alternative hypotheses; hence $\pi_0 = 1 - m_1/m$. $n_k = 3$ and $K = 4$ for all models. $N(0, \sigma^2)$ denotes normal random noise

Model	m	m_1	σ	True H_A's in addition to $X_1, \ldots X_4$, X_{190}, X_{221}
1	3000	500	3	$X_i = X_1 + N(0, \sigma^2)$, $i = 5, \ldots, 16$
				$X_i = -X_1 + N(0, \sigma^2)$, $i = 17, \ldots, 25$
				$X_i = X_2 + N(0, \sigma^2)$, $i = 26, \ldots, 60$
				$X_i = -X_2 + N(0, \sigma^2)$, $i = 61, \ldots, 70$
				$X_i = X_3 + N(0, \sigma^2)$, $i = 71, \ldots, 100$
				$X_i = -X_3 + N(0, \sigma^2)$, $i = 101, \ldots, 110$
				$X_i = X_4 + N(0, \sigma^2)$, $i = 111, \ldots, 150$
				$X_i = -X_4 + N(0, \sigma^2)$, $i = 151, \ldots, 189$
				$X_i = X_{190} + N(0, \sigma^2)$, $i = 191, \ldots, 210$
				$X_i = -X_{190} + N(0, \sigma^2)$, $i = 211, \ldots, 220$
				$X_i = X_{221} + N(0, \sigma^2)$, $i = 222, \ldots, 250$
				$X_i = 2X_{i-250} + N(0, \sigma^2)$, $i = 251, \ldots, 500$
2	3000	500	1	the same as Model 1
3	3000	250	3	the same as Model 1 except only the first 250 are true H_A's
4	3000	250	1	the same as Model 3
5	3000	32	3	$X_i = X_1 + N(0, \sigma^2)$, $i = 5, \ldots, 8$
				$X_i = X_2 + N(0, \sigma^2)$, $i = 9, \ldots, 12$
				$X_i = X_3 + N(0, \sigma^2)$, $i = 13, \ldots, 16$
				$X_i = X_4 + N(0, \sigma^2)$, $i = 17, \ldots, 20$
				$X_i = X_{190} + N(0, \sigma^2)$, $i = 191, \ldots, 195$
				$X_i = X_{221} + N(0, \sigma^2)$, $i = 222, \ldots, 226$
6	3000	32	1	the same as Model 5
7	3000	6	3	none, except $X_1, \ldots X_4$, X_{190}, X_{221}
8	3000	6	1	the same as Model 7
9	10000	15	3	$X_i = X_1 + N(0, \sigma^2)$, $i = 5, 6$
				$X_i = X_2 + N(0, \sigma^2)$, $i = 7, 8$
				$X_i = X_3 + N(0, \sigma^2)$, $i = 9, 10$
				$X_i = X_4 + N(0, \sigma^2)$, $i = 11, 12$
				$X_{191} = X_{190} + N(0, \sigma^2)$
10	10000	15	1	the same as Model 9

Acknowledgments. I am grateful to Dr. Stan Pounds, two referees, and Professor Javier Rojo for their comments and suggestions that substantially improved this paper.

References

[1] ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. AND JOHNSTONE, I. (2000). Adapting to unknown sparsity by controlling the false discover rate. Technical Report 2000-19, Department of Statistics, Stanford University, Stanford, CA.

[2] ALLISON, D. B., GADBURY, G. L., HEO, M. FERNANDEZ, J. R., LEE, C-K, PROLLA, T. A. AND WEINDRUCH, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.* **39**, 1–20.

[3] BENJAMINI, Y. AND HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

[4] BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300.

[5] Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2005). Adaptive linear step-up procedures that control the false discovery rate. Research Paper 01-03, Dept. of Statistics and Operations Research, Tel Aviv University.

[6] Bickel, D. R. (2004). Error-rate and decision-theoretic methods of multiple testing: which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology* **3**, Article 8. URL `//www.bepress.com/sagmb/vol3/iss1/art8.`

[7] Cheng, C., Pounds, S., Boyett, J. M., Pei, D., Kuo, M-L., Roussel, M. F. (2004). Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 36. URL `//www.bepress.com/sagmb/vol3/iss1/art36.`

[8] de Boor (1987). *A Practical Guide to Splines.* Springer, New York.

[9] Dudoit, S., van der Laan, M., Pollard, K. S. (2004). Multiple Testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 13. URL `//www.bepress.com/sagmb/vol3/iss1/art13.`

[10] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96–104.

[11] Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151–1160.

[12] Finner, H. and Roberts, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30**, 220–238.

[13] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 499–517.

[14] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery rates. *Ann. Statist.* **32**, 1035–1061.

[15] Halmos, P. R. (1974). *Measure Theory.* Springer, New York.

[16] Hardy, G., Littlewood, J. E. and Pólya, G. (1952). *Inequalities.* Cambridge University Press, Cambridge, UK.

[17] Ishawaran, H. and Rao, S. (2003). Detecting differentially genes in microarrays using Baysian model selection. *J. Amer. Statist. Assoc.* **98**, 438–455.

[18] Kuo, M.-L., Duncavich, E., Cheng, C., Pei, D., Sherr, C. J. and Roussel M. F. (2003). Arf induces p53-dependent and in-dependent antiproliferative genes. *Cancer Research* **1**, 1046–1053.

[19] Langaas, M., Ferkingstady, E. and Lindqvist, B. H. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 555–572.

[20] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist* **37**, 1137–1153.

[21] Mosig, M. O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M. and Friedmann, A. (2001). A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157**, 1683–1698.

[22] Nettleton, D. and Hwang, G. (2003). Estimating the number of false null hypotheses when conducting many tests. Technical Report 2003-09, Department of Statistics, Iowa State University, `http://www.stat.iastate.edu/preprint/articles/2003-09.pdf`

[23] Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method, *Biostatistics* **5**, 155–176.

[24] Pounds, S. and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242.

[25] Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20**, 1737–1745.

[26] Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

[27] Schweder, T. and Spjøtvoll, E. (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika* **69** 493-502.

[28] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 3. URL: //www.bepress.com/sagmb/vol3/iss1/art3.

[29] Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 479–498.

[30] Storey, J. D. (2003). The positive false discovery rate: a Baysian interpretation and the q-value. *Ann. Statis.* **31**, 2103–2035.

[31] Storey, J. D., Taylor, J. and Siegmund, D. (2003). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 187–205.

[32] Storey, J. D. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data* (Parmigiani, G. et al., eds.). Springer, New York.

[33] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445.

[34] Tsai, C-A., Hsueh, H-M. and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics* **59**, 1071–1081.

[35] Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121.

[36] van der Laan, M., Dudoit, S. and Pollard, K. S. (2004a). Multiple Testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 14. URL: //www.bepress.com/sagmb/vol3/iss1/art14.

[37] van der Laan, M., Dudoit, S. and Pollard, K. S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 15. URL: //www.bepress.com/sagmb/vol3/iss1/art15.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 77–97
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000400

Frequentist statistics as a theory of inductive inference

Deborah G. Mayo[1] and D. R. Cox[2]

Viriginia Tech and Nuffield College, Oxford

Abstract: After some general remarks about the interrelation between philosophical and statistical thinking, the discussion centres largely on significance tests. These are defined as the calculation of p-values rather than as formal procedures for "acceptance" and "rejection." A number of types of null hypothesis are described and a principle for evidential interpretation set out governing the implications of p-values in the specific circumstances of each application, as contrasted with a long-run interpretation. A variety of more complicated situations are discussed in which modification of the simple p-value may be essential.

1. Statistics and inductive philosophy

1.1. What is the Philosophy of Statistics?

The philosophical foundations of statistics may be regarded as the study of the epistemological, conceptual and logical problems revolving around the use and interpretation of statistical methods, broadly conceived. As with other domains of philosophy of science, work in statistical science progresses largely without worrying about "philosophical foundations". Nevertheless, even in statistical practice, debates about the different approaches to statistical analysis may influence and be influenced by general issues of the nature of inductive-statistical inference, and thus are concerned with foundational or philosophical matters. Even those who are largely concerned with applications are often interested in identifying general principles that underlie and justify the procedures they have come to value on relatively pragmatic grounds. At one level of analysis at least, statisticians and philosophers of science ask many of the same questions.

- What should be observed and what may justifiably be inferred from the resulting data?
- How well do data confirm or fit a model?
- What is a good test?
- Does failure to reject a hypothesis H constitute evidence "confirming" H?
- How can it be determined whether an apparent anomaly is genuine? How can blame for an anomaly be assigned correctly?
- Is it relevant to the relation between data and a hypothesis if looking at the data influences the hypothesis to be examined?
- How can spurious relationships be distinguished from genuine regularities?

[1]Department of Philosophy and Economics, Virginia Tech, Blacksburg, VA 24061-0126, e-mail: mayod@vt.edu

[2]Nuffield College, Oxford OX1 1NF, UK, e-mail: david.cox@nuffield.ox.ac.uk

AMS 2000 subject classifications: 62B15, 62F03.

Keywords and phrases: statistical inference, significance test, confidence interval, test of hypothesis, Neyman–Pearson theory, selection effect, multiple testing.

- How can a causal explanation and hypothesis be justified and tested?
- How can the gap between available data and theoretical claims be bridged reliably?

That these very general questions are entwined with long standing debates in philosophy of science helps explain why the field of statistics tends to cross over, either explicitly or implicitly, into philosophical territory. Some may even regard statistics as a kind of "applied philosophy of science" (Fisher [10]; Kempthorne [13]), and statistical theory as a kind of "applied philosophy of inductive inference". As Lehmann [15] has emphasized, Neyman regarded his work not only as a contribution to statistics but also to inductive philosophy. A core question that permeates "inductive philosophy" both in statistics and philosophy is: What is the nature and role of probabilistic concepts, methods, and models in making inferences in the face of limited data, uncertainty and error?

Given the occasion of our contribution, a session on philosophy of statistics for the second Lehmann symposium, we take as our springboard the recommendation of Neyman ([22], p. 17) that we view statistical theory as essentially a "Frequentist Theory of Inductive Inference". The question then arises as to what conception(s) of inductive inference would allow this. Whether or not this is the only or even the most satisfactory account of inductive inference, it is interesting to explore how much progress towards an account of inductive inference, as opposed to inductive behavior, one might get from frequentist statistics (with a focus on testing and associated methods). These methods are, after all, often used for inferential ends, to learn about aspects of the underlying data generating mechanism, and much confusion and criticism (e.g., as to whether and why error rates are to be adjusted) could be avoided if there was greater clarity on the roles in inference of hypothetical error probabilities.

Taking as a backdrop remarks by Fisher [10], Lehmann [15] on Neyman, and by Popper [26] on induction, we consider the roles of significance tests in bridging inductive gaps in traditional hypothetical deductive inference. Our goal is to identify a key principle of evidence by which hypothetical error probabilities may be used for inductive inference from specific data, and to consider how it may direct and justify (a) different uses and interpretations of statistical significance levels in testing a variety of different types of null hypotheses, and (b) when and why "selection effects" need to be taken account of in data dependent statistical testing.

1.2. The role of probability in frequentist induction

The defining feature of an inductive inference is that the premises (evidence statements) can be true while the conclusion inferred may be false without a logical contradiction: the conclusion is "evidence transcending". Probability naturally arises in capturing such evidence transcending inferences, but there is more than one way this can occur. Two distinct philosophical traditions for using probability in inference are summed up by Pearson ([24], p. 228):

"For one school, the degree of confidence in a proposition, a quantity varying with the nature and extent of the evidence, provides the basic notion to which the numerical scale should be adjusted." The other school notes the relevance in ordinary life and in many branches of science of a knowledge of the relative frequency of occurrence of a particular class of events in a series of repetitions, and suggests that "it is through its link with relative frequency that probability has the most direct meaning for the human mind".

Frequentist induction, whatever its form, employs probability in the second manner. For instance, significance testing appeals to probability to characterize the proportion of cases in which a null hypothesis H_0 would be rejected in a hypothetical long-run of repeated sampling, an error probability. This difference in the role of probability corresponds to a difference in the form of inference deemed appropriate: The former use of probability traditionally has been tied to the view that a probabilistic account of induction involves quantifying a degree of support or confirmation in claims or hypotheses.

Some followers of the frequentist approach agree, preferring the term "inductive behavior" to describe the role of probability in frequentist statistics. Here the inductive reasoner "decides to infer" the conclusion, and probability quantifies the associated risk of error. The idea that one role of probability arises in science to characterize the "riskiness" or probativeness or severity of the tests to which hypotheses are put is reminiscent of the philosophy of Karl Popper [26]. In particular, Lehmann ([16], p. 32) has noted the temporal and conceptual similarity of the ideas of Popper and Neyman on "finessing" the issue of induction by replacing inductive reasoning with a process of hypothesis testing.

It is true that Popper and Neyman have broadly analogous approaches based on the idea that we can speak of a hypothesis having been well-tested in some sense, quite distinct from its being accorded a degree of probability, belief or confirmation; this is "finessing induction". Both also broadly shared the view that in order for data to "confirm" or "corroborate" a hypothesis H, that hypothesis would have to have been subjected to a test with high probability or power to have rejected it if false. But despite the close connection of the ideas, there appears to be no reference to Popper in the writings of Neyman (Lehmann [16], p. 3) and the references by Popper to Neyman are scant and scarcely relevant. Moreover, because Popper denied that any inductive claims were justifiable, his philosophy forced him to deny that even the method he espoused (conjecture and refutations) was reliable. Although H might be true, Popper made it clear that he regarded corroboration at most as a report of the past performance of H: it warranted no claims about its reliability in future applications. By contrast, a central feature of frequentist statistics is to be able to assess and control the probability that a test would have rejected a hypothesis, if false. These probabilities come from formulating the data generating process in terms of a statistical model.

Neyman throughout his work emphasizes the importance of a probabilistic model of the system under study and describes frequentist statistics as modelling the phenomenon of the stability of relative frequencies of results of repeated "trials", granting that there are other possibilities concerned with modelling psychological phenomena connected with intensities of belief, or with readiness to bet specified sums, etc. citing Carnap [2], de Finetti [8] and Savage [27]. In particular Neyman criticized the view of "frequentist" inference taken by Carnap for overlooking the key role of the stochastic model of the phenomenon studied. Statistical work related to the inductive philosophy of Carnap [2] is that of Keynes [14] and, with a more immediate impact on statistical applications, Jeffreys [12].

1.3. Induction and hypothetical-deductive inference

While "hypothetical-deductive inference" may be thought to "finesse" induction, in fact inductive inferences occur throughout empirical testing. Statistical testing ideas may be seen to fill these inductive gaps: If the hypothesis were deterministic

we could find a relevant function of the data whose value (i) represents the relevant feature under test and (ii) can be predicted by the hypothesis. We calculate the function and then see whether the data agree or disagree with the prediction. If the data conflict with the prediction, then either the hypothesis is in error or some auxiliary or other background factor may be blamed for the anomaly (Duhem's problem).

Statistical considerations enter in two ways. If H is a statistical hypothesis, then usually no outcome strictly contradicts it. There are major problems involved in regarding data as inconsistent with H merely because they are highly improbable; all individual outcomes described in detail may have very small probabilities. Rather the issue, essentially following Popper ([26], pp. 86, 203), is whether the possibly anomalous outcome represents some systematic and reproducible effect.

The focus on falsification by Popper as the goal of tests, and falsification as the defining criterion for a scientific theory or hypothesis, clearly is strongly redolent of Fisher's thinking. While evidence of direct influence is virtually absent, the views of Popper agree with the statement by Fisher ([9], p. 16) that every experiment may be said to exist only in order to give the facts the chance of disproving the null hypothesis. However, because Popper's position denies ever having grounds for inference about reliability, he denies that we can ever have grounds for inferring reproducible deviations.

The advantage in the modern statistical framework is that the probabilities arise from defining a probability model to represent the phenomenon of interest. Had Popper made use of the statistical testing ideas being developed at around the same time, he might have been able to substantiate his account of falsification.

The second issue concerns the problem of how to reason when the data "agree" with the prediction. The argument from H entails data y, and that y is observed, to the inference that H is correct is, of course, deductively invalid. A central problem for an inductive account is to be able nevertheless to warrant inferring H in some sense. However, the classical problem, even in deterministic cases, is that many rival hypotheses (some would say infinitely many) would also predict y, and thus would pass as well as H. In order for a test to be probative, one wants the prediction from H to be something that at the same time is in some sense very surprising and not easily accounted for were H false and important rivals to H correct. We now consider how the gaps in inductive testing may bridged by a specific kind of statistical procedure, the significance test.

2. Statistical significance tests

Although the statistical significance test has been encircled by controversies for over 50 years, and has been mired in misunderstandings in the literature, it illustrates in simple form a number of key features of the perspective on frequentist induction that we are considering. See for example Morrison and Henkel [21] and Gibbons and Pratt [11]. So far as possible, we begin with the core elements of significance testing in a version very strongly related to but in some respects different from both Fisherian and Neyman-Pearson approaches, at least as usually formulated.

2.1. General remarks and definition

We suppose that we have empirical data denoted collectively by y and that we treat these as observed values of a random variable Y. We regard y as of interest only in so far as it provides information about the probability distribution of

Y as defined by the relevant statistical model. This probability distribution is to be regarded as an often somewhat abstract and certainly idealized representation of the underlying data generating process. Next we have a hypothesis about the probability distribution, sometimes called the hypothesis under test but more often conventionally called the null hypothesis and denoted by H_0. We shall later set out a number of quite different types of null hypotheses but for the moment we distinguish between those, sometimes called simple, that completely specify (in principle numerically) the distribution of Y and those, sometimes called composite, that completely specify certain aspects and which leave unspecified other aspects.

In many ways the most elementary, if somewhat hackneyed, example is that Y consists of n independent and identically distributed components normally distributed with unknown mean μ and possibly unknown standard deviation σ. A simple hypothesis is obtained if the value of σ is known, equal to σ_0, say, and the null hypothesis is that $\mu = \mu_0$, a given constant. A composite hypothesis in the same context might have σ unknown and again specify the value of μ.

Note that in this formulation it is required that some unknown aspect of the distribution, typically one or more unknown parameters, is precisely specified. The hypothesis that, for example, $\mu \leq \mu_0$ is not an acceptable formulation for a null hypothesis in a Fisherian test; while this more general form of null hypothesis is allowed in Neyman-Pearson formulations.

The immediate objective is to test the conformity of the particular data under analysis with H_0 in some respect to be specified. To do this we find a function $t = t(y)$ of the data, to be called the test statistic, such that

- the larger the value of t the more inconsistent are the data with H_0;
- the corresponding random variable $T = t(Y)$ has a (numerically) known probability distribution when H_0 is true.

These two requirements parallel the corresponding deterministic ones. To assess whether there is a genuine discordancy (or reproducible deviation) from H_0 we define the so-called p-value corresponding to any t as

$$p = p(t) = P(T \geq t; H_0),$$

regarded as a measure of concordance with H_0 in the respect tested. In at least the initial formulation alternative hypotheses lurk in the undergrowth but are not explicitly formulated probabilistically; also there is no question of setting in advance a preassigned threshold value and "rejecting" H_0 if and only if $p \leq \alpha$. Moreover, the justification for tests will not be limited to appeals to long run-behavior but will instead identify an inferential or evidential rationale. We now elaborate.

2.2. Inductive behavior vs. inductive inference

The reasoning may be regarded as a statistical version of the valid form of argument called in deductive logic *modus tollens*. This infers the denial of a hypothesis H from the combination that H entails E, together with the information that E is false. Because there was a high probability $(1-p)$ that a less significant result would have occurred were H_0 true, we may justify taking low p-values, properly computed, as evidence against H_0. Why? There are two main reasons:

Firstly such a rule provides low error rates (i.e., erroneous rejections) in the long run when H_0 is true, a behavioristic argument. In line with an error- assessment view of statistics we may give any particular value p, say, the following hypothetical

interpretation: suppose that we were to treat the data as just decisive evidence against H_0. Then in hypothetical repetitions H_0 would be rejected in a long-run proportion p of the cases in which it is actually true. However, knowledge of these hypothetical error probabilities may be taken to underwrite a distinct justification.

This is that such a rule provides a way to determine whether a specific data set is evidence of a discordancy from H_0.

In particular, a low p-value, so long as it is properly computed, provides evidence of a discrepancy from H_0 in the respect examined, while a p-value that is not small affords evidence of accordance or consistency with H_0 (where this is to be distinguished from positive evidence for H_0, as discussed below in Section 2.3). Interest in applications is typically in whether p is in some such range as $p \geq 0.1$ which can be regarded as reasonable accordance with H_0 in the respect tested, or whether p is near to such conventional numbers as 0.05, 0.01, 0.001. Typical practice in much applied work is to give the observed value of p in rather approximate form. A small value of p indicates that (i) H_0 is false (there is a discrepancy from H_0) or (ii) the basis of the statistical test is flawed, often that real errors have been underestimated, for example because of invalid independence assumptions, or (iii) the play of chance has been extreme.

It is part of the object of good study design and choice of method of analysis to avoid (ii) by ensuring that error assessments are relevant.

There is no suggestion whatever that the significance test would typically be the only analysis reported. In fact, a fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results. Although the complexity of the story makes it more difficult to set out neatly, as, for example, if a single algorithm is thought to capture the whole of inductive inference, the payoff is an account that approaches the kind of full-bodied arguments that scientists build up in order to obtain reliable knowledge and understanding of a field.

Amidst the complexity, significance test reasoning reflects a fairly straightforward conception of evaluating evidence anomalous for H_0 in a statistical context, the one Popper perhaps had in mind but lacked the tools to implement. The basic idea is that error probabilities may be used to evaluate the "riskiness" of the predictions H_0 is required to satisfy, by assessing the reliability with which the test discriminates whether (or not) the actual process giving rise to the data accords with that described in H_0. Knowledge of this probative capacity allows determining if there is strong evidence of discordancy The reasoning is based on the following frequentist principle for identifying whether or not there is evidence against H_0:

FEV (i) y is (strong) evidence against H_0, i.e. (strong) evidence of discrepancy from H_0, if and only if, where H_0 a correct description of the mechanism generating y, then, with high probability, this would have resulted in a less discordant result than is exemplified by y.

A corollary of **FEV** is that y is not (strong) evidence against H_0, if the probability of a more discordant result is not very low, even if H_0 is correct. That is, if there is a moderately high probability of a more discordant result, even were H_0 correct, then H_0 accords with y in the respect tested.

Somewhat more controversial is the interpretation of a failure to find a small p-value; but an adequate construal may be built on the above form of **FEV**.

2.3. Failure and confirmation

The difficulty with regarding a modest value of p as evidence in favour of H_0 is that accordance between H_0 and y may occur even if rivals to H_0 seriously different from H_0 are true. This issue is particularly acute when the amount of data is limited. However, sometimes we can find evidence for H_0, understood as an assertion that a particular discrepancy, flaw, or error is absent, and we can do this by means of tests that, with high probability, would have reported a discrepancy had one been present. As much as Neyman is associated with automatic decision-like techniques, in practice at least, both he and E. S. Pearson regarded the appropriate choice of error probabilities as reflecting the specific context of interest (Neyman[23], Pearson [24]).

There are two different issues involved. One is whether a particular value of p is to be used as a threshold in each application. This is the procedure set out in most if not all formal accounts of Neyman-Pearson theory. The second issue is whether control of long-run error rates is a justification for frequentist tests or whether the ultimate justification of tests lies in their role in interpreting evidence in particular cases. In the account given here, the achieved value of p is reported, at least approximately, and the "accept- reject" account is purely hypothetical to give p an operational interpretation. E. S. Pearson [24] is known to have disassociated himself from a narrow behaviourist interpretation (Mayo [17]). Neyman, at least in his discussion with Carnap (Neyman [23]) seems also to hint at a distinction between behavioural and inferential interpretations.

In an attempt to clarify the nature of frequentist statistics, Neyman in this discussion was concerned with the term "degree of confirmation" used by Carnap. In the context of an example where an optimum test had failed to "reject" H_0, Neyman considered whether this "confirmed" H_0. He noted that this depends on the meaning of words such as "confirmation" and "confidence" and that in the context where H_0 had not been "rejected" it would be "dangerous" to regard this as confirmation of H_0 if the test in fact had little chance of detecting an important discrepancy from H_0 even if such a discrepancy were present. On the other hand if the test had appreciable power to detect the discrepancy the situation would be "radically different".

Neyman is highlighting an inductive fallacy associated with "negative results", namely that if data y yield a test result that is not statistically significantly different from H_0 (e.g., the null hypothesis of 'no effect'), and yet the test has small probability of rejecting H_0, even when a serious discrepancy exists, then y is not good evidence for inferring that H_0 is confirmed by y. One may be confident in the absence of a discrepancy, according to this argument, only if the chance that the test would have correctly detected a discrepancy is high.

Neyman compares this situation with interpretations appropriate for inductive behaviour. Here confirmation and confidence may be used to describe the choice of action, for example refraining from announcing a discovery or the decision to treat H_0 as satisfactory. The rationale is the pragmatic behavioristic one of controlling errors in the long-run. This distinction implies that even for Neyman evidence for deciding may require a distinct criterion than evidence for believing; but unfortunately Neyman did not set out the latter explicitly. We propose that the needed evidential principle is an adaption of **FEV**(**i**) for the case of a p-value that is not small:

FEV(**ii**): A moderate p value is evidence of the absence of a discrepancy δ from

H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller p value) were a discrepancy δ to exist. **FEV**(**ii**) especially arises in the context of "embedded" hypotheses (below).

What makes the kind of hypothetical reasoning relevant to the case at hand is not solely or primarily the long-run low error rates associated with using the tool (or test) in this manner; it is rather what those error rates reveal about the data generating source or phenomenon. The error-based calculations provide reassurance that incorrect interpretations of the evidence are being avoided in the particular case. To distinguish between this "evidential" justification of the reasoning of significance tests, and the "behavioristic" one, it may help to consider a very informal example of applying this reasoning "to the specific case". Thus suppose that weight gain is measured by well-calibrated and stable methods, possibly using several measuring instruments and observers and the results show negligible change over a test period of interest. This may be regarded as grounds for inferring that the individual's weight gain is negligible within limits set by the sensitivity of the scales. Why?

While it is true that by following such a procedure in the long run one would rarely report weight gains erroneously, that is not the rationale for the particular inference. The justification is rather that the error probabilistic properties of the weighing procedure reflect what is actually the case in the specific instance. (This should be distinguished from the evidential interpretation of Neyman–Pearson theory suggested by Birnbaum [1], which is not data-dependent.)

The significance test is a measuring device for accordance with a specified hypothesis calibrated, as with measuring devices in general, by its performance in repeated applications, in this case assessed typically theoretically or by simulation. Just as with the use of measuring instruments, applied to a specific case, we employ the performance features to make inferences about aspects of the particular thing that is measured, aspects that the measuring tool is appropriately capable of revealing.

Of course for this to hold the probabilistic long-run calculations must be as relevant as feasible to the case in hand. The implementation of this surfaces in statistical theory in discussions of conditional inference, the choice of appropriate distribution for the evaluation of p. Difficulties surrounding this seem more technical than conceptual and will not be dealt with here, except to note that the exercise of applying (or attempting to apply) **FEV** may help to guide the appropriate test specification.

3. Types of null hypothesis and their corresponding inductive inferences

In the statistical analysis of scientific and technological data, there is virtually always external information that should enter in reaching conclusions about what the data indicate with respect to the primary question of interest. Typically, these background considerations enter not by a probability assignment but by identifying the question to be asked, designing the study, interpreting the statistical results and relating those inferences to primary scientific ones and using them to extend and support underlying theory. Judgments about what is relevant and informative must be supplied for the tools to be used non- fallaciously and as intended. Nevertheless, there are a cluster of systematic uses that may be set out corresponding to types of test and types of null hypothesis.

3.1. Types of null hypothesis

We now describe a number of types of null hypothesis. The discussion amplifies that given by Cox ([4], [5]) and by Cox and Hinkley [6]. Our goal here is not to give a guide for the panoply of contexts a researcher might face, but rather to elucidate some of the different interpretations of test results and the associated p-values. In Section 4.3, we consider the deeper interpretation of the corresponding inductive inferences that, in our view, are (and are not) licensed by p-value reasoning.

1. *Embedded null hypotheses.* In these problems there is formulated, not only a probability model for the null hypothesis, but also models that represent other possibilities in which the null hypothesis is false and, usually, therefore represent possibilities we would wish to detect if present. Among the number of possible situations, in the most common there is a parametric family of distributions indexed by an unknown parameter θ partitioned into components $\theta = (\phi, \lambda)$, such that the null hypothesis is that $\phi = \phi_0$, with λ an unknown nuisance parameter and, at least in the initial discussion with ϕ one-dimensional. Interest focuses on alternatives $\phi > \phi_0$.

This formulation has the technical advantage that it largely determines the appropriate test statistic $t(y)$ by the requirement of producing the most sensitive test possible with the data at hand.

There are two somewhat different versions of the above formulation. In one the full family is a tentative formulation intended not to so much as a possible base for ultimate interpretation but as a device for determining a suitable test statistic. An example is the use of a quadratic model to test adequacy of a linear relation; on the whole polynomial regressions are a poor base for final analysis but very convenient and interpretable for detecting small departures from a given form. In the second case the family is a solid base for interpretation. Confidence intervals for ϕ have a reasonable interpretation.

One other possibility, that arises very rarely, is that there is a simple null hypothesis and a single simple alternative, i.e. only two possible distributions are under consideration. If the two hypotheses are considered on an equal basis the analysis is typically better considered as one of hypothetical or actual discrimination, i.e. of determining which one of two (or more, generally a very limited number) of possibilities is appropriate, treating the possibilities on a conceptually equal basis.

There are two broad approaches in this case. One is to use the likelihood ratio as an index of relative fit, possibly in conjunction with an application of Bayes theorem. The other, more in accord with the error probability approach, is to take each model in turn as a null hypothesis and the other as alternative leading to an assessment as to whether the data are in accord with both, one or neither hypothesis. Essentially the same interpretation results by applying **FEV** to this case, when it is framed within a Neyman–Pearson framework.

We can call these three cases those of a formal family of alternatives, of a well-founded family of alternatives and of a family of discrete possibilities.

2. *Dividing null hypotheses.* Quite often, especially but not only in technological applications, the focus of interest concerns a comparison of two or more conditions, processes or treatments with no particular reason for expecting the outcome to be exactly or nearly identical, e.g., compared with a standard a new drug may increase or may decrease survival rates.

One, in effect, combines two tests, the first to examine the possibility that $\mu > \mu_0$,

say, the other for $\mu < \mu_0$. In this case, the two- sided test combines both one-sided tests, each with its own significance level. The significance level is twice the smaller p, because of a "selection effect" (Cox and Hinkley [6], p. 106). We return to this issue in Section 4. The null hypothesis of zero difference then divides the possible situations into two qualitatively different regions with respect to the feature tested, those in which one of the treatments is superior to the other and a second in which it is inferior.

3. *Null hypotheses of absence of structure.* In quite a number of relatively empirically conceived investigations in fields without a very firm theory base, data are collected in the hope of finding structure, often in the form of dependencies between features beyond those already known. In epidemiology this takes the form of tests of potential risk factors for a disease of unknown aetiology.

4. *Null hypotheses of model adequacy.* Even in the fully embedded case where there is a full family of distributions under consideration, rich enough potentially to explain the data whether the null hypothesis is true or false, there is the possibility that there are important discrepancies with the model sufficient to justify extension, modification or total replacement of the model used for interpretation. In many fields the initial models used for interpretation are quite tentative; in others, notably in some areas of physics, the models have a quite solid base in theory and extensive experimentation. But in all cases the possibility of model misspecification has to be faced even if only informally.

There is then an uneasy choice between a relatively focused test statistic designed to be sensitive against special kinds of model inadequacy (powerful against specific directions of departure), and so-called omnibus tests that make no strong choices about the nature of departures. Clearly the latter will tend to be insensitive, and often extremely insensitive, against specific alternatives. The two types broadly correspond to chi-squared tests with small and large numbers of degrees of freedom. For the focused test we may either choose a suitable test statistic or, almost equivalently, a notional family of alternatives. For example to examine agreement of n independent observations with a Poisson distribution we might in effect test the agreement of the sample variance with the sample mean by a chi-squared dispersion test (or its exact equivalent) or embed the Poisson distribution in, for example, a negative binomial family.

5. *Substantively-based null hypotheses.* In certain special contexts, null results may indicate substantive evidence for scientific claims in contexts that merit a fifth category. Here, a theory $\mathcal{T}$ for which there is appreciable theoretical and/or empirical evidence predicts that H_0 is, at least to a very close approximation, the true situation.

(a) In one version, there may be results apparently anomalous for $\mathcal{T}$, and a test is designed to have ample opportunity to reveal a discordancy with H_0 if the anomalous results are genuine.

(b) In a second version a rival theory $\mathcal{T}^*$ predicts a specified discrepancy from H_0. and the significance test is designed to discriminate between $\mathcal{T}$ and the rival theory $\mathcal{T}^*$ (in a thus far not tested domain).

For an example of (a) physical theory suggests that because the quantum of energy in nonionizing electro-magnetic fields, such as those from high voltage transmission lines, is much less than is required to break a molecular bond, there should be no carcinogenic effect from exposure to such fields. Thus in a randomized ex-

periment in which two groups of mice are under identical conditions except that one group is exposed to such a field, the null hypothesis that the cancer incidence rates in the two groups are identical may well be exactly true and would be a prime focus of interest in analysing the data. Of course the null hypothesis of this general kind does not have to be a model of zero effect; it might refer to agreement with previous well-established empirical findings or theory.

3.2. Some general points

We have in the above described essentially one-sided tests. The extension to two-sided tests does involve some issues of definition but we shall not discuss these here.

Several of the types of null hypothesis involve an incomplete probability specification. That is, we may have only the null hypothesis clearly specified. It might be argued that a full probability formulation should always be attempted covering both null and feasible alternative possibilities. This may seem sensible in principle but as a strategy for direct use it is often not feasible; in any case models that would cover all reasonable possibilities would still be incomplete and would tend to make even simple problems complicated with substantial harmful side-effects.

Note, however, that in all the formulations used here some notion of explanations of the data alternative to the null hypothesis is involved by the choice of test statistic; the issue is when this choice is made via an explicit probabilistic formulation. The general principle of evidence **FEV** helps us to see that in specified contexts, the former suffices for carrying out an evidential appraisal (see Section 3.3).

It is, however, sometimes argued that the choice of test statistic can be based on the distribution of the data under the null hypothesis alone, in effect choosing minus the log probability as test statistic, thus summing probabilities over all sample points as or less probable than that observed. While this often leads to sensible results we shall not follow that route here.

3.3. Inductive inferences based on outcomes of tests

How does significance test reasoning underwrite inductive inferences or evidential evaluations in the various cases? The hypothetical operational interpretation of the p-value is clear but what are the deeper implications either of a modest or of a small value of p? These depends strongly both on (i) the type of null hypothesis, and (ii) the nature of the departure or alternative being probed, as well as (iii) whether we are concerned with the interpretation of particular sets of data, as in most detailed statistical work, or whether we are considering a broad model for analysis and interpretation in a field of study. The latter is close to the traditional Neyman-Pearson formulation of fixing a critical level and accepting, in some sense, H_0 if $p > \alpha$ and rejecting H_0 otherwise. We consider some of the familiar shortcomings of a routine or mechanical use of p-values.

3.4. The routine-behavior use of p-values

Imagine one sets $\alpha = 0.05$ and that results lead to a publishable paper if and only for the relevant p, the data yield $p < 0.05$. The rationale is the behavioristic one outlined earlier. Now the great majority of statistical discussion, going back to Yates

[32] and earlier, deplores such an approach, both out of a concern that it encourages mechanical, automatic and unthinking procedures, as well as a desire to emphasize estimation of relevant effects over testing of hypotheses. Indeed a few journals in some fields have in effect banned the use of p-values. In others, such as a number of areas of epidemiology, it is conventional to emphasize 95% confidence intervals, as indeed is in line with much mainstream statistical discussion. Of course, this does not free one from needing to give a proper frequentist account of the use and interpretation of confidence levels, which we do not do here (though see Section 3.6).

Nevertheless the relatively mechanical use of p-values, while open to parody, is not far from practice in some fields; it does serve as a screening device, recognizing the possibility of error, and decreasing the possibility of the publication of misleading results. A somewhat similar role of tests arises in the work of regulatory agents, in particular the FDA. While requiring studies to show p less than some preassigned level by a preordained test may be inflexible, and the choice of critical level arbitrary, nevertheless such procedures have virtues of impartiality and relative independence from unreasonable manipulation. While adhering to a fixed p-value may have the disadvantage of biasing the literature towards positive conclusions, it offers an appealing assurance of some known and desirable long-run properties. They will be seen to be particularly appropriate for Example 3 of Section 4.2.

3.5. The inductive-evidence use of p-values

We now turn to the use of significance tests which, while more common, is at the same time more controversial; namely as one tool to aid the analysis of specific sets of data, and/or base inductive inferences on data. The discussion presupposes that the probability distribution used to assess the p-value is as appropriate as possible to the specific data under analysis.

The general frequentist principle for inductive reasoning, **FEV**, or something like it, provides a guide for the appropriate statement about evidence or inference regarding each type of null hypothesis. Much as one makes inferences about changes in body mass based on performance characteristics of various scales, one may make inferences from significance test results by using error rate properties of tests. They indicate the capacity of the particular test to have revealed inconsistencies and discrepancies in the respects probed, and this in turn allows relating p-values to hypotheses about the process as statistically modelled. It follows that an adequate frequentist account of inference should strive to supply the information to implement **FEV**.

Embedded Nulls. In the case of embedded null hypotheses, it is straightforward to use small p-values as evidence of discrepancy from the null in the direction of the alternative. Suppose, however, that the data are found to accord with the null hypothesis (p not small). One may, if it is of interest, regard this as evidence that any discrepancy from the null is less than δ, using the same logic in significance testing. In such cases concordance with the null may provide evidence of the absence of a discrepancy from the null of various sizes, as stipulated in **FEV**(**ii**).

To infer the absence of a discrepancy from H_0 as large as δ we may examine the probability $\beta(\delta)$ of observing a worse fit with H_0 if $\mu = \mu_0 + \delta$. If that probability is near one then, following **FEV**(**ii**), the data are good evidence that $\mu < \mu_0 + \delta$. Thus $\beta(\delta)$ may be regarded as the stringency or severity with which the test has probed the discrepancy δ; equivalently one might say that $\mu < \mu_0 + \delta$ has passed a severe test (Mayo [17]).

This avoids unwarranted interpretations of consistency with H_0 with insensitive tests. Such an assessment is more relevant to specific data than is the notion of power, which is calculated relative to a predesignated critical value beyond which the test "rejects" the null. That is, power appertains to a prespecified rejection region, not to the specific data under analysis.

Although oversensitivity is usually less likely to be a problem, if a test is so sensitive that a p-value as or even smaller than the one observed, is probable even when $\mu < \mu_0 + \delta$, then a small value of p is not evidence of departure from H_0 in excess of δ.

If there is an explicit family of alternatives, it will be possible to give a set of confidence intervals for the unknown parameter defining H_0 and this would give a more extended basis for conclusions about the defining parameter.

Dividing and absence of structure nulls. In the case of dividing nulls, discordancy with the null (using the two-sided value of p) indicates direction of departure (e.g., which of two treatments is superior); accordance with H_0 indicates that these data do not provide adequate evidence even of the direction of any difference. One often hears criticisms that it is pointless to test a null hypothesis known to be false, but even if we do not expect two means, say, to be equal, the test is informative in order to divide the departures into qualitatively different types. The interpretation is analogous when the null hypothesis is one of absence of structure: a modest value of p indicates that the data are insufficiently sensitive to detect structure. If the data are limited this may be no more than a warning against over-interpretation rather than evidence for thinking that indeed there is no structure present. That is because the test may have had little capacity to have detected any structure present. A small value of p, however, indicates evidence of a genuine effect; that to look for a substantive interpretation of such an effect would not be intrinsically error-prone.

Analogous reasoning applies when assessments about the probativeness or sensitivity of tests are informal. If the data are so extensive that accordance with the null hypothesis implies the absence of an effect of practical importance, and a reasonably high p-value is achieved, then it may be taken as evidence of the absence of an effect of practical importance. Likewise, if the data are of such a limited extent that it can be assumed that data in accord with the null hypothesis are consistent also with departures of scientific importance, then a high p-value does not warrant inferring the absence of scientifically important departures from the null hypothesis.

Nulls of model adequacy. When null hypotheses are assertions of model adequacy, the interpretation of test results will depend on whether one has a relatively focused test statistic designed to be sensitive against special kinds of model inadequacy, or so called omnibus tests. Concordance with the null in the former case gives evidence of absence of the type of departure that the test is sensitive in detecting, whereas, with the omnibus test, it is less informative. In both types of tests, a small p-value is evidence of some departure, but so long as various alternative models could account for the observed violation (i.e., so long as this test had little ability to discriminate between them), these data by themselves may only provide provisional suggestions of alternative models to try.

Substantive nulls. In the preceding cases, accordance with a null could at most provide evidence to rule out discrepancies of specified amounts or types, according to the ability of the test to have revealed the discrepancy. More can be said in the case of substantive nulls. If the null hypothesis represents a prediction from

some theory being contemplated for general applicability, consistency with the null hypothesis may be regarded as some additional evidence for the theory, especially if the test and data are sufficiently sensitive to exclude major departures from the theory. An aspect is encapsulated in Fisher's aphorism (Cochran [3]) that to help make observational studies more nearly bear a causal interpretation, one should make one's theories elaborate, by which he meant one should plan a variety of tests of different consequences of a theory, to obtain a comprehensive check of its implications. The limited result that one set of data accords with the theory adds one piece to the evidence whose weight stems from accumulating an ability to refute alternative explanations.

In the first type of example under this rubric, there may be apparently anomalous results for a theory or hypothesis $\mathcal{T}$, where $\mathcal{T}$ has successfully passed appreciable theoretical and/or empirical scrutiny. Were the apparently anomalous results for $\mathcal{T}$ genuine, it is expected that H_0 will be rejected, so that when it is not, the results are positive evidence against the reality of the anomaly. In a second type of case, one again has a well-tested theory $\mathcal{T}$, and a rival theory $\mathcal{T}^*$ is determined to conflict with $\mathcal{T}$ in a thus far untested domain, with respect to an effect. By identifying the null with the prediction from $\mathcal{T}$, any discrepancies in the direction of $\mathcal{T}^*$ are given a very good chance to be detected, such that, if no significant departure is found, this constitutes evidence for $\mathcal{T}$ in the respect tested.

Although the general theory of relativity, GTR, was not facing anomalies in the 1960s, rivals to the GTR predicted a breakdown of the Weak Equivalence Principle for massive self-gravitating bodies, e.g., the earth-moon system: this effect, called the Nordvedt effect would be 0 for GTR (identified with the null hypothesis) and non-0 for rivals. Measurements of the round trip travel times between the earth and moon (between 1969 and 1975) enabled the existence of such an anomaly for GTR to be probed. Finding no evidence against the null hypothesis set upper bounds to the possible violation of the WEP, and because the tests were sufficiently sensitive, these measurements provided good evidence that the Nordvedt effect is absent, and thus evidence for the null hypothesis (Will [31]). Note that such a negative result does not provide evidence for all of GTR (in all its areas of prediction), but it does provide evidence for its correctness with respect to this effect. The logic is this: theory $\mathcal{T}$ predicts H_0 is at least a very close approximation to the true situation; rival theory $\mathcal{T}^*$ predicts a specified discrepancy from H_0, and the test has high probability of detecting such a discrepancy from $\mathcal{T}$ were $\mathcal{T}^*$ correct. Detecting no discrepancy is thus evidence for its absence.

3.6. Confidence intervals

As noted above in many problems the provision of confidence intervals, in principle at a range of probability levels, gives the most productive frequentist analysis. If so, then confidence interval analysis should also fall under our general frequentist principle. It does. In one sided testing of $\mu = \mu_0$ against $\mu > \mu_0$, a small p-value corresponds to μ_0 being (just) excluded from the corresponding $(1-2p)$ (two-sided) confidence interval (or $1-p$ for the one-sided interval). Were $\mu = \mu_L$, the lower confidence bound, then a less discordant result would occur with high probability $(1-p)$. Thus **FEV** licenses taking this as evidence of inconsistency with $\mu = \mu_L$ (in the positive direction). Moreover, this reasoning shows the advantage of considering several confidence intervals at a range of levels, rather than just reporting whether or not a given parameter value is within the interval at a fixed confidence level.

Neyman developed the theory of confidence intervals *ab initio* i.e. relying only implicitly rather than explicitly on his earlier work with E.S. Pearson on the theory of tests. It is to some extent a matter of presentation whether one regards interval estimation as so different in principle from testing hypotheses that it is best developed separately to preserve the conceptual distinction. On the other hand there are considerable advantages to regarding a confidence limit, interval or region as the set of parameter values consistent with the data at some specified level, as assessed by testing each possible value in turn by some mutually concordant procedures. In particular this approach deals painlessly with confidence intervals that are null or which consist of all possible parameter values, at some specified significance level. Such null or infinite regions simply record that the data are inconsistent with all possible parameter values, or are consistent with all possible values. It is easy to construct examples where these seem entirely appropriate conclusions.

4. Some complications: selection effects

The idealized formulation involved in the initial definition of a significance test in principle starts with a hypothesis and a test statistic, then obtains data, then applies the test and looks at the outcome. The hypothetical procedure involved in the definition of the test then matches reasonably closely what was done; the possible outcomes are the different possible values of the specified test statistic. This permits features of the distribution of the test statistic to be relevant for learning about corresponding features of the mechanism generating the data. There are various reasons why the procedure actually followed may be different and we now consider one broad aspect of that.

It often happens that either the null hypothesis or the test statistic are influenced by preliminary inspection of the data, so that the actual procedure generating the final test result is altered. This in turn may alter the capabilities of the test to detect discrepancies from the null hypotheses reliably, calling for adjustments in its error probabilities.

To the extent that p is viewed as an aspect of the logical or mathematical relation between the data and the probability model such preliminary choices are irrelevant. This will not suffice in order to ensure that the p-values serve their intended purpose for frequentist inference, whether in behavioral or evidential contexts. To the extent that one wants the error-based calculations that give the test its meaning to be applicable to the tasks of frequentist statistics, the preliminary analysis and choice may be highly relevant.

The general point involved has been discussed extensively in both philosophical and statistical literatures, in the former under such headings as requiring novelty or avoiding *ad hoc* hypotheses, under the latter, as rules against peeking at the data or shopping for significance, and thus requiring selection effects to be taken into account. The general issue is whether the evidential bearing of data y on an inference or hypothesis H_0 is altered when H_0 has been either constructed or selected for testing in such a way as to result in a specific observed relation between H_0 and y, whether that is agreement or disagreement. Those who favour logical approaches to confirmation say no (e.g., Mill [20], Keynes [14]), whereas those closer to an error statistical conception say yes (Whewell [30], Pierce [25]). Following the latter philosophy, Popper required that scientists set out in advance what outcomes they would regard as falsifying H_0, a requirement that even he came to reject; the entire issue in philosophy remains unresolved (Mayo [17]).

Error statistical considerations allow going further by providing criteria for when various data dependent selections matter and how to take account of their influence on error probabilities. In particular, if the null hypothesis is chosen for testing because the test statistic is large, the probability of finding some such discordance or other may be high even under the null. Thus, following **FEV(i)**, we would not have genuine evidence of discordance with the null, and unless the p-value is modified appropriately, the inference would be misleading. To the extent that one wants the error-based calculations that give the test its meaning to supply reassurance that apparent inconsistency in the particular case is genuine and not merely due to chance, adjusting the p-value is called for.

Such adjustments often arise in cases involving data dependent selections either in model selection or construction; often the question of adjusting p arises in cases involving multiple hypotheses testing, but it is important not to run cases together simply because there is data dependence or multiple hypothesis testing. We now outline some special cases to bring out the key points in different scenarios. Then we consider whether allowance for selection is called for in each case.

4.1. Examples

Example 1. An investigator has, say, 20 independent sets of data, each reporting on different but closely related effects. The investigator does all 20 tests and reports only the smallest p, which in fact is about 0.05, and its corresponding null hypothesis. The key points are the independence of the tests and the failure to report the results from insignificant tests.

Example 2. A highly idealized version of testing for a DNA match with a given specimen, perhaps of a criminal, is that a search through a data-base of possible matches is done one at a time, checking whether the hypothesis of agreement with the specimen is rejected. Suppose that sensitivity and specificity are both very high. That is, the probabilities of false negatives and false positives are both very small. The first individual, if any, from the data-base for which the hypothesis is rejected is declared to be the true match and the procedure stops there.

Example 3. A microarray study examines several thousand genes for potential expression of say a difference between Type 1 and Type 2 disease status. There are thus several thousand hypotheses under investigation in one step, each with its associated null hypothesis.

Example 4. To study the dependence of a response or outcome variable y on an explanatory variable x it is intended to use a linear regression analysis of y on x. Inspection of the data suggests that it would be better to use the regression of $\log y$ on $\log x$, for example because the relation is more nearly linear or because secondary assumptions, such as constancy of error variance, are more nearly satisfied.

Example 5. To study the dependence of a response or outcome variable y on a considerable number of potential explanatory variables x, a data-dependent procedure of variable selection is used to obtain a representation which is then fitted by standard methods and relevant hypotheses tested.

Example 6. Suppose that preliminary inspection of data suggests some totally unexpected effect or regularity not contemplated at the initial stages. By a formal test the effect is very "highly significant". What is it reasonable to conclude?

4.2. Need for adjustments for selection

There is not space to discuss all these examples in depth. A key issue concerns which of these situations need an adjustment for multiple testing or data dependent selection and what that adjustment should be. How does the general conception of the character of a frequentist theory of analysis and interpretation help to guide the answers?

We propose that it does so in the following manner: Firstly it must be considered whether the context is one where the key concern is the control of error rates in a series of applications (behavioristic goal), or whether it is a context of making a specific inductive inference or evaluating specific evidence (inferential goal). The relevant error probabilities may be altered for the former context and not for the latter. Secondly, the relevant sequence of repetitions on which to base frequencies needs to be identified. The general requirement is that we do not report discordance with a null hypothesis by means a procedure that would report discordancies fairly frequently even though the null hypothesis is true. Ascertainment of the relevant hypothetical series on which this error frequency is to be calculated demands consideration of the nature of the problem or inference. More specifically, one must identify the particular obstacles that need to be avoided for a reliable inference in the particular case, and the capacity of the test, as a measuring instrument, to have revealed the presence of the obstacle.

When the goal is appraising specific evidence, our main interest, **FEV** gives some guidance. More specifically the problem arises when data are used to select a hypothesis to test or alter the specification of an underlying model in such a way that **FEV** is either violated or it cannot be determined whether **FEV** is satisfied (Mayo and Kruse [18]).

Example 1 (Hunting for statistical significance). The test procedure is very different from the case in which the single null found statistically significant was preset as the hypothesis to test, perhaps it is $H_{0,13}$,the 13th null hypothesis out of the 20. In Example 1, the possible results are the possible statistically significant factors that might be found to show a "calculated" statistical significant departure from the null. Hence the type 1 error probability is the probability of finding at least one such significant difference out of 20, even though the global null is true (i.e., all twenty observed differences are due to chance). The probability that this procedure yields an erroneous rejection differs from, and will be much greater than, 0.05 (and is approximately 0.64). There are different, and indeed many more, ways one can err in this example than when one null is prespecified, and this is reflected in the adjusted p-value.

This much is well known, but should this influence the interpretation of the result in a context of inductive inference? According to **FEV** it should. However the concern is not the avoidance of often announcing genuine effects erroneously in a series, the concern is that this test performs poorly as a tool for discriminating genuine from chance effects in this particular case. Because at least one such impressive departure, we know, is common even if all are due to chance, the test has scarcely reassured us that it has done a good job of avoiding such a mistake in this case. Even if there are other grounds for believing the genuineness of the one effect that is found, we deny that this test alone has supplied such evidence.

Frequentist calculations serve to examine the particular case, we have been saying, by characterizing the capability of tests to have uncovered mistakes in inference, and on those grounds, the "hunting procedure" has low capacity to have alerted us

to, in effect, temper our enthusiasm, even where such tempering is warranted. If, on the other hand, one adjusts the p-value to reflect the overall error rate, the test again becomes a tool that serves this purpose.

Example 1 may be contrasted to a standard factorial experiment set up to investigate the effects of several explanatory variables simultaneously. Here there are a number of distinct questions, each with its associated hypothesis and each with its associated p-value. That we address the questions via the same set of data rather than via separate sets of data is in a sense a technical accident. Each p is correctly interpreted in the context of its own question. Difficulties arise for particular inferences only if we in effect throw away many of the questions and concentrate only on one, or more generally a small number, chosen just because they have the smallest p. For then we have altered the capacity of the test to have alerted us, by means of a correctly computed p-value, whether we have evidence for the inference of interest.

Example 2 (Explaining a known effect by eliminative induction). Example 2 is superficially similar to Example 1, finding a DNA match being somewhat akin to finding a statistically significant departure from a null hypothesis: one searches through data and concentrates on the one case where a "match" with the criminal's DNA is found, ignoring the non-matches. If one adjusts for "hunting" in Example 1, shouldn't one do so in broadly the same way in Example 2? No.

In Example 1 the concern is that of inferring a genuine,"reproducible" effect, when in fact no such effect exists; in Example 2, there is a known effect or specific event, the criminal's DNA, and reliable procedures are used to track down the specific cause or source (as conveyed by the low "erroneous-match" rate.) The probability is high that we would not obtain a match with person i, if i were not the criminal; so, by **FEV**, finding the match is, at a qualitative level, good evidence that i is the criminal. Moreover, each non-match found, by the stipulations of the example, virtually excludes that person; thus, the more such negative results the stronger is the evidence when a match is finally found. The more negative results found, the more the inferred "match" is fortified; whereas in Example 1 this is not so.

Because at most one null hypothesis of innocence is false, evidence of innocence on one individual increases, even if only slightly, the chance of guilt of another. An assessment of error rates is certainly possible once the sampling procedure for testing is specified. Details will not be given here.

A broadly analogous situation concerns the anomaly of the orbit of Mercury: the numerous failed attempts to provide a Newtonian interpretation made it all the more impressive when Einstein's theory was found to predict the anomalous results precisely and without any *ad hoc* adjustments.

Example 3 (Micro-array data). In the analysis of micro-array data, a reasonable starting assumption is that a very large number of null hypotheses are being tested and that some fairly small proportion of them are (strictly) false, a global null hypothesis of no real effects at all often being implausible. The problem is then one of selecting the sites where an effect can be regarded as established. Here, the need for an adjustment for multiple testing is warranted mainly by a pragmatic concern to avoid "too much noise in the network". The main interest is in how best to adjust error rates to indicate most effectively the gene hypotheses worth following up. An error-based analysis of the issues is then via the false-discovery rate, i.e. essentially the long run proportion of sites selected as positive in which no effect is present. An alternative formulation is via an empirical Bayes model and the conclusions from this can be linked to the false discovery rate. The latter method may be preferable

because an error rate specific to each selected gene may be found; the evidence in some cases is likely to be much stronger than in others and this distinction is blurred in an overall false-discovery rate. See Shaffer [28] for a systematic review.

Example 4 (Redefining the test). If tests are run with different specifications, and the one giving the more extreme statistical significance is chosen, then adjustment for selection is required, although it may be difficult to ascertain the precise adjustment. By allowing the result to influence the choice of specification, one is altering the procedure giving rise to the p-value, and this may be unacceptable. While the substantive issue and hypothesis remain unchanged the precise specification of the probability model has been guided by preliminary analysis of the data in such a way as to alter the stochastic mechanism actually responsible for the test outcome.

An analogy might be testing a sharpshooter's ability by having him shoot and then drawing a bull's-eye around his results so as to yield the highest number of bull's-eyes, the so-called principle of the Texas marksman. The skill that one is allegedly testing and making inferences about is his ability to shoot when the target is given and fixed, while that is not the skill actually responsible for the resulting high score.

By contrast, if the choice of specification is guided not by considerations of the statistical significance of departure from the null hypothesis, but rather because the data indicates the need to allow for changes to achieve linearity or constancy of error variance, no allowance for selection seems needed. Quite the contrary: choosing the more empirically adequate specification gives reassurance that the calculated p-value is relevant for interpreting the evidence reliably. (Mayo and Spanos [19]). This might be justified more formally by regarding the specification choice as an informal maximum likelihood analysis, maximizing over a parameter orthogonal to those specifying the null hypothesis of interest.

Example 5 (Data mining). This example is analogous to Example 1, although how to make the adjustment for selection may not be clear because the procedure used in variable selection may be tortuous. Here too, the difficulties of selective reporting are bypassed by specifying all those reasonably simple models that are consistent with the data rather than by choosing only one model (Cox and Snell [7]). The difficulties of implementing such a strategy are partly computational rather than conceptual. Examples of this sort are important in much relatively elaborate statistical analysis in that series of very informally specified choices may be made about the model formulation best for analysis and interpretation (Spanos [29]).

Example 6 (The totally unexpected effect). This raises major problems. In laboratory sciences with data obtainable reasonably rapidly, an attempt to obtain independent replication of the conclusions would be virtually obligatory. In other contexts a search for other data bearing on the issue would be needed. High statistical significance on its own would be very difficult to interpret, essentially because selection has taken place and it is typically hard or impossible to specify with any realism the set over which selection has occurred. The considerations discussed in Examples 1-5, however, may give guidance. If, for example, the situation is as in Example 2 (explaining a known effect) the source may be reliably identified in a procedure that fortifies, rather than detracts from, the evidence. In a case akin to Example 1, there is a selection effect, but it is reasonably clear what is the set of possibilities over which this selection has taken place, allowing correction of the p-value. In other examples, there is a selection effect, but it may not be clear how

to make the correction. In short, it would be very unwise to dismiss the possibility of learning from data something new in a totally unanticipated direction, but one must discriminate the contexts in order to gain guidance for what further analysis, if any, might be required.

5. Concluding remarks

We have argued that error probabilities in frequentist tests may be used to evaluate the reliability or capacity with which the test discriminates whether or not the actual process giving rise to data is in accordance with that described in H_0. Knowledge of this probative capacity allows determination of whether there is strong evidence against H_0 based on the frequentist principle we set out **FEV**. What makes the kind of hypothetical reasoning relevant to the case at hand is not the long-run low error rates associated with using the tool (or test) in this manner; it is rather what those error rates reveal about the data generating source or phenomenon. We have not attempted to address the relation between the frequentist and Bayesian analyses of what may appear to be very similar issues. A fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results; we have set out considerations to guide these pieces. Although the complexity of the issues makes it more difficult to set out neatly, as, for example, one could by imagining that a single algorithm encompasses the whole of inductive inference, the payoff is an account that approaches the kind of arguments that scientists build up in order to obtain reliable knowledge and understanding of a field.

References

[1] Birnbaum, A. (1977). The Neyman–Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley–Savage argument for Bayesian theory. *Synthese* **36**, 19–49.
[2] Carnap, R. (1962). *Logical Foundations of Probability*. University of Chicago Press.
[3] Cochran, W. G. (1965). The planning of observational studies in human populations (with discussion). *J.R.Statist. Soc.* A **128**, 234–265.
[4] Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.
[5] Cox, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist.* **4**, 49–70.
[6] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
[7] Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies. *J. R. Statist. Soc.* C **23**, 51–59.
[8] De Finetti, B. (1974). *Theory of Probability*, 2 vols. English translation from Italian. Wiley, New York.
[9] Fisher, R. A. (1935a). *Design of Experiments*. Oliver and Boyd, Edinburgh.
[10] Fisher, R. A. (1935b). The logic of inductive inference. *J. R. Statist. Soc.* **98**, 39–54.
[11] Gibbons, J. D. and Pratt, J. W. (1975). *P*-values: Interpretation and methodology. *American Statistician* **29**, 20–25.

[12] Jeffreys, H. (1961). *Theory of Probability*, Third edition. Oxford University Press.

[13] Kempthorne, O. (1976). Statistics and the philosophers. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* Harper and Hooker (eds.), Vol. 2, 273–314.

[14] Keynes, J. M. [1921] (1952). *A Treatise on Probability*. Reprint. St. Martin's press, New York.

[15] Lehmann, E. L. (1993). The Fisher and Neyman–Pearson theories of testing hypotheses: One theory or two? *J. Amer. Statist. Assoc.* **88**, 1242–1249.

[16] Lehmann, E. L. (1995). Neyman's statistical philosophy. *Probability and Mathematical Statistics* **15**, 29–36.

[17] Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press.

[18] Mayo, D. G. and M. Kruse (2001). Principles of inference and their consequences. In *Foundations of Bayesianism*, D. Cornfield and J. Williamson (eds.). Kluwer Academic Publishers, Netherlands, 381–403.

[19] Mayo, D. G. and Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal of Philosophy of Science* **57**, 323–357.

[20] Mill, J. S. (1988). *A System of Logic*, Eighth edition. Harper and Brother, New York.

[21] Morrison, D. and Henkel, R. (eds.) (1970). *The Significance Test Controversy*. Aldine, Chicago.

[22] Neyman, J. (1955). The problem of inductive inference. *Comm. Pure and Applied Maths* **8**, 13–46.

[23] Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Int. Statist. Rev.* **25**, 7–22.

[24] Pearson, E. S. (1955). Statistical concepts in their relation to reality. *J. R. Statist. Soc.* B **17**, 204–207.

[25] Pierce, C. S. [1931-5]. *Collected Papers*, Vols. 1–6, Hartshorne and Weiss, P. (eds.). Harvard University Press, Cambridge.

[26] Popper, K. (1959). *The Logic of Scientific Discovery*. Basic Books, New York.

[27] Savage, L. J. (1964). The foundations of statistics reconsidered. In *Studies in Subjective Probability*, Kyburg H. E. and H. E. Smokler (eds.). Wiley, New York, 173–188.

[28] Shaffer, J. P. (2005). This volume.

[29] Spanos, A. (2000). Revisiting data mining: 'hunting' with or without a license. *Journal of Economic Methodology* **7**, 231–264.

[30] Whewell, W. [1847] (1967). *The Philosophy of the Inductive Sciences. Founded Upon Their History*, Second edition, Vols. 1 and 2. Reprint. Johnson Reprint, London.

[31] Will, C. (1993). *Theory and Experiment in Gravitational Physics*. Cambridge University Press.

[32] Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J. Amer. Statist. Assoc.* **46**, 19–34.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 98–119
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000419

Where do statistical models come from? Revisiting the problem of specification

Aris Spanos[*1]

Virginia Polytechnic Institute and State University

Abstract: R. A. Fisher founded modern statistical inference in 1922 and identified its fundamental problems to be: *specification, estimation* and *distribution*. Since then the problem of *statistical model specification* has received scant attention in the statistics literature. The paper traces the history of statistical model specification, focusing primarily on pioneers like Fisher, Neyman, and more recently Lehmann and Cox, and attempts a synthesis of their views in the context of the Probabilistic Reduction (PR) approach. As argued by Lehmann [11], a major stumbling block for a general approach to statistical model specification has been the delineation of the appropriate role for substantive subject matter information. The PR approach demarcates the interrelated but complemenatry roles of *substantive* and *statistical information* summarized *ab initio* in the form of a *structural* and a *statistical model*, respectively. In an attempt to preserve the integrity of both sources of information, as well as to ensure the reliability of their fusing, a purely probabilistic construal of statistical models is advocated. This probabilistic construal is then used to shed light on a number of issues relating to specification, including the role of preliminary data analysis, structural vs. statistical models, model specification vs. model selection, statistical vs. substantive adequacy and model validation.

1. Introduction

The current approach to statistics, interpreted broadly as 'probability-based data modeling and inference', has its roots going back to the early 19th century, but it was given its current formulation by R. A. Fisher [5]. He identified the fundamental problems of statistics to be: *specification*, *estimation* and *distribution*. Despite its importance, the question of *specification,* 'where do statistical models come from?' received only scant attention in the statistics literature; see Lehmann [11].

The cornerstone of modern statistics is the notion of a *statistical model* whose meaning and role have changed and evolved along with that of statistical modeling itself over the last two centuries. Adopting a retrospective view, a statistical model is defined to be an internally consistent set of probabilistic assumptions aiming to provide an 'idealized' probabilistic description of the stochastic mechanism that gave rise to the observed data $\mathbf{x} := (x_1, x_2, \ldots, x_n)$. The quintessential statistical model is the *simple Normal model*, comprising a statistical Generating Mechanism (GM):

$$X_k = \mu + u_k,\ k \in \mathbb{N} := \{1, 2, \ldots n, \ldots\} \tag{1.1}$$

[*]I'm most grateful to Erich Lehmann, Deborah G. Mayo, Javier Rojo and an anonymous referee for valuable suggestions and comments on an earlier draft of the paper.

[1]Department of Economics, Virginia Polytechnic Institute, and State University, Blacksburg, VA 24061, e-mail: aris@vt.edu

AMS 2000 subject classifications: 62N-03, 62A01, 62J20, 60J65.

Keywords and phrases: specification, statistical induction, misspecification testing, respecification, statistical adequacy, model validation, substantive vs. statistical information, structural vs. statistical models.

together with the probabilistic assumptions:

$$X_k \sim \mathsf{NIID}(\mu, \sigma^2),\ k \in \mathbb{N}, \tag{1.2}$$

where $X_k \sim$NIID stands for *Normal, Independent and Identically Distributed*. The nature of a statistical model will be discussed in section 3, but as a prelude to that, it is important to emphasize that it is specified exclusively in terms of probabilistic concepts that can be related directly to the *joint distribution* of the *observable* stochastic process $\{X_k,\ k \in \mathbb{N}\}$. This is in contrast to other forms of models that play a role in statistics, such as *structural* (explanatory, substantive), which are based on substantive subject matter information and are specified in terms of theory concepts.

The motivation for such a purely probabilistic construal of statistical models arises from an attempt to circumvent some of the difficulties for a general approach to statistical modeling. These difficulties were raised by early pioneers like Fisher [5]–[7] and Neyman [17]–[26], and discussed extensively by Lehmann [11] and Cox [1]. The main difficulty, as articulated by Lehmann [11], concerns the role of *substantive subject matter* information. His discussion suggests that if statistical model specification requires such information at the outset, then any attempt to provide a general approach to statistical modeling is unattainable. His main conclusion is that, despite the untenability of a general approach, statistical theory has a contribution to make in model specification by extending and improving: (a) the reservoir of models, (b) the model selection procedures, and (c) the different types of models.

In this paper it is argued that Lehmann's case concerning (a)–(c) can be strengthened and extended by adopting a purely probabilistic construal of statistical models and placing statistical modeling in a broader framework which allows for fusing statistical and substantive information in a way which does not compromise the integrity of either. Substantive subject matter information emanating from the theory, and statistical information reflecting the probabilistic structure of the data, need to be viewed as bringing to the table different but complementary information. The *Probabilistic Reduction* (PR) approach offers such a modeling framework by integrating several innovations in Neyman's writings into Fisher's initial framework with a view to address a number of modeling problems, including the role of preliminary data analysis, structural vs. statistical models, model specification vs. model selection, statistical vs. substantive adequacy and model validation. Due to space limitations the picture painted in this paper will be dominated by broad brush strokes with very few details; see Spanos [31]–[42] for further discussion.

1.1. Substantive vs. statistical information

Empirical modeling in the social and physical sciences involves an intricate blending of *substantive subject matter* and *statistical information*. Many aspects of empirical modeling implicate both sources of information in a variety of functions, and others involve one or the other, more or less separately. For instance, the development of *structural* (explanatory) *models* is primarily based on *substantive information* and it is concerned with the mathematization of theories to give rise to theory models, which are amenable to empirical analysis; that activity, by its very nature, cannot be separated from the disciplines in question. On the other hand, certain aspects of empirical modeling, which focus on *statistical information* and are concerned with the nature and use of statistical models, can form a body of knowledge which is shared by all fields that use data in their modeling. This is the body of knowledge

that *statistics* can claim as its subject matter and develop it with only one eye on new problems/issues raised by empirical modeling in other disciplines. This ensures that statistics is not subordinated to the other applied fields, but remains a separate discipline which provides, maintains and extends/develops the common foundation and overarching framework for empirical modeling.

To be more specific, *statistical model specification* refers to the choice of a model (parameterization) arising from the probabilistic structure of a stochastic process $\{X_k,\ k\in\mathbb{N}\}$ that would render the data in question $\mathbf{x}:=(x_1,x_2,\ldots,x_n)$ a *truly typical realization* thereof. This perspective on the data is referred to as the Fisher–Neyman probabilistic perspective for reasons that will become apparent in section 2. When one specifies the simple Normal model (1.1), the only thing that matters from the statistical specification perspective is whether the data $\mathbf{x}$ can be realistically viewed as a truly typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$ assumed to be NIID, devoid of any substantive information. A model is said to be *statistical adequate* when the assumptions constituting the statistical model in question, NIID, are valid for the data $\mathbf{x}$ in question. Statistical adequacy can be assessed qualitatively using analogical reasoning in conjunction with data graphs (t-plots, P-P plots etc.), as well as quantitatively by testing the assumptions constituting the statistical model using probative Mis-Specification (M-S) tests; see Spanos [36].

It is argued that certain aspects of statistical modeling, such as statistical model specification, the use of graphical techniques, M-S testing and respecification, together with optimal inference procedures (estimation, testing and prediction), can be developed *generically* by viewing data $\mathbf{x}$ as a realization of a (nameless) stochastic process $\{X_k,\ t\in\mathbb{N}\}$. All these aspects of empirical modeling revolve around a central axis we call a *statistical model*. Such models can be viewed as *canonical models,* in the sense used by Mayo [12], which are developed without any reference to substantive subject matter information, and can be used equally in physics, biology, economics and psychology. Such canonical models and the associated statistical modeling and inference belong to the *realm of statistics*. Such a view will broaden the scope of modern statistics by integrating preliminary data analysis, statistical model specification, M-S testing and respecification into the current textbook discourses; see Cox and Hinkley [2], Lehmann [10].

On the other hand the question of *substantive adequacy*, i.e. whether a structural model adequately captures the main features of the *actual Data Generating Mechanism* (DGM) giving rise to data $\mathbf{x}$, cannot be addressed in a generic way because it concerns the bridge between the particular model and the phenomenon of interest. Even in this case, however, assessing substantive adequacy will take the form of applying statistical procedures within an embedding statistical model. Moreover, for the error probing to be *reliable* one needs to ensure that the embedding model is statistically adequate; it captures all the statistical systematic information (Spanos, [41]). In this sense, substantive subject matter information (which might range from vary vague to highly informative) constitutes important supplementary information which, under *statistical* and *substantive adequacy*, enhances the explanatory and predictive power of statistical models.

In the spirit of Lehmann [11], models in this paper are classified into:

(a) *statistical* (empirical, descriptive, interpolatory formulae, data models), and
(b) *structural* (explanatory, theoretical, mechanistic, substantive).

The harmonious integration of these two sources of information gives rise to an

(c) *empirical model*; the term is not equivalent to that in Lehmann [11].

In Section 2, the paper traces the development of ideas, issues and problems surrounding statistical model specification from Karl Pearson [27] to Lehmann [11], with particular emphasis on the perspectives of Fisher and Neyman. Some of the ideas and modeling suggestions of these pioneers are synthesized in Section 3 in the form of the PR modeling framework. Kepler's first law of planetary motion is used to illustrate some of the concepts and ideas. The PR perspective is then used to shed light on certain issues raised by Lehmann [11] and Cox [1].

2. 20th century statistics

2.1. Early debates: description vs. induction

Before Fisher, the notion of a statistical model was both vague and implicit in data modeling, with its role primarily confined to the *description* of the distributional properties of *the data in hand* using the histogram and the first few sample moments. A crucial problem with the application of *descriptive statistics* in the late 19th century was that statisticians would often claim generality beyond the data in hand for their inferences. This is well-articulated by Mills [16]:

"In approaching this subject [statistics] we must first make clear the distinction between *statistical description* and *statistical induction*. By employing the methods of statistics it is possible, as we have seen, to describe succinctly a mass of quantitative data." ... "In so far as the results are confined to the cases actually studied, these various statistical measures are merely devices for describing certain features of a distribution, or certain relationships. Within these limits the measures may be used to perfect confidence, as accurate descriptions of the given characteristics. But when we seek to extend these results, to generalize the conclusions, to apply them to cases not included in the original study, a quite new set of problems is faced." (p. 548-9)

Mills [16] went on to discuss the 'inherent assumptions' necessary for the validity of *statistical induction:*

"... in the larger population to which this result is to be applied, there exists a uniformity with respect to the characteristic or relation we have measured" ..., and "... the sample from which our first results were derived is thoroughly representative of the entire population to which the results are to be applied." (pp. 550-2).

The fine line between statistical description and statistical induction was nebulous until the 1920s, for several reasons. *First*, "No distinction was drawn between a sample and the population, and what was calculated from the sample was attributed to the population." (Rao, [29], p. 35). *Second*, it was thought that the inherent assumptions for the validity of statistical induction are *not empirically verifiable*; see Mills [16], p. 551). *Third*, there was a widespread belief, exemplified in the first quotation from Mills, that statistical description does not require any assumptions. It is well-known today that there is no such thing as a meaningful summary of the data that does not involve any implicit assumptions; see Neyman [21]. For instance, the arithmetic average of a trending time series represents no meaningful feature of the underlying 'population'.

2.2. Karl Pearson

Karl Pearson was able to take descriptive statistics to a higher level of sophistication by proposing the 'graduation (smoothing) of histograms' into 'frequency curves'; see Pearson [27]. This, however, introduced additional fuzziness into the distinction between statistical description vs. induction because the frequency curves were the precursors to the density functions; one of the crucial components of a statistical model introduced by Fisher [5] providing the foundation of statistical induction. The statistical modeling procedure advocated by Pearson, however, was very different from that introduced by Fisher.

For Karl Pearson statistical modeling would begin with data $\mathbf{x} := (x_1, x_2, \ldots, x_n)$ in search of a descriptive model which would be in the form of a frequency curve $f(x)$, chosen from the Pearson family $f(x;\theta)$, $\theta := (a, b_0, b_1, b_2)$, after applying the method of moments to obtain $\widehat{\theta}$ (see Pearson, [27]). Viewed from today's perspective, the solution $\widehat{\theta}$, would deal with two different statistical problems simultaneously, (a) specification (the choice a descriptive model $f(x;\widehat{\theta})$) and (b) *estimation* of θ using $\widehat{\theta}$. $f(x;\widehat{\theta})$ can subsequently be used to draw inferences beyond the original data $\mathbf{x}$.

Pearson's view of statistical induction, as late as 1920, was that of *induction by enumeration* which relies on both *prior* distributions and the stability of relative frequencies; see Pearson [28], p. 1.

2.3. R. A. Fisher

One of Fisher's most remarkable but least appreciated achievements, was to initiate the recasting of the form of *statistical induction* into its modern variant. Instead of starting with data $\mathbf{x}$ in search of a descriptive model, he would interpret the data as a truly *representative sample* from a pre-specified 'hypothetical infinite population'. This might seem like a trivial re-arrangement of Pearson's procedure, but in fact it constitutes a complete recasting of the problem of statistical induction, with the notion of a *parameteric statistical model* delimiting its premises.

Fisher's first clear statement of this major change from the then prevailing modeling process is given in his classic 1922 paper:

"... the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data. This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion." ([5], p. 311)

Fisher goes on to elaborate on the modeling process itself: "The problems which arise in reduction of data may be conveniently divided into three types: (1) **Problems of Specification**. These arise in the choice of the mathematical form of the population. (2) **Problems of Estimation**. (3) **Problems of Distribution**.

It will be clear that when we know (1) what parameters are required to specify the population from which the sample is drawn, (2) how best to calculate from

the sample estimates of these parameters, and (3) the exact form of the distribution, in different samples, of our derived statistics, then the theoretical aspect of the treatment of any particular body of data has been completely elucidated." (p. 313-4)

One can summarize Fisher's view of the statistical modeling process as follows. The process begins with a *prespecified parametric statistical model* $\mathcal{M}$ ('a hypothetical infinite population'), chosen so as to ensure that the observed data $\mathbf{x}$ are viewed as a *truly representative sample* from that 'population':

"The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician." ([5], p. 313)

Fisher was fully aware of the fact that the specification of a statistical model premises all forms of statistical inference. Once $\mathcal{M}$ was specified, the original uncertainty relating to the 'population' was reduced to uncertainty concerning the unknown parameter(s) θ, associated with $\mathcal{M}$. In Fisher's set up, the parameter(s) θ, are *unknown constants* and become the focus of inference. The problems of 'estimation' and 'distribution' revolve around θ.

Fisher went on to elaborate further on the 'problems of specification': "As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and *the adequacy of our choice may be tested a posteriori*. We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed. More or less elaborate form will be suitable according to the volume of the data." (p. 314) [emphasis added]

Based primarily on the above quoted passage, Lehmann's [11] assessment of Fisher's view on specification is summarized as follows: "Fisher's statement implies that in his view there can be no theory of modeling, no general modeling strategies, but that instead each problem must be considered entirely on its own merits. He does not appear to have revised his opinion later... Actually, following this uncompromisingly negative statement, Fisher unbends slightly and offers two general suggestions concerning model building: (a) "We must confine ourselves to those forms which we know how to handle," and (b) "More or less elaborate forms will be suitable according to the volume of the data."" (p. 160-1).

Lehmann's interpretation is clearly warranted, but Fisher's view of specification has some additional dimensions that need to be brought out. The original choice of a statistical model may be guided by simplicity and experience, but as Fisher emphasizes "the adequacy of our choice may be tested *a posteriori*." What comes after the above quotation is particularly interesting to be quoted in full: "Evidently these are considerations the nature of which may change greatly during the work of a single generation. We may instance the development by Pearson of a very extensive system of skew curves, the elaboration of a method of calculating their parameters, and the preparation of the necessary tables, a body of work which has enormously extended the power of modern statistical practice, and which has been, by pertinacity and inspiration alike, practically the work of a single man. Nor is the introduction of the Pearsonian system of frequency curves the only contribution which their author has made to the solution of problems of specification: of even greater importance is the introduction of an objective criterion of goodness of fit. For empirical as the specification of the hypothetical population may be, *this empiricism*

is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts. Once a statistic suitable for applying such a test, has been chosen, the exact form of its distribution in random samples must be investigated, in order that we may evaluate the probability that a worse fit should be obtained from a random sample of a population of the type considered. The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, *since therein lies our justification for the free use which is made of empirical frequency formulae.* Problems of distribution of great mathematical difficulty have to be faced in this direction." (p. 314) [emphasis (in italic) added]

In this quotation Fisher emphasizes the empirical dimension of the specification problem, and elaborates on testing the assumptions of the model, lavishing Karl Pearson with more praise for developing the *goodness of fit test* than for his family of densities. He clearly views this test as a primary tool for assessing the validity of the original specification (misspecification testing). He even warns the reader of the potentially complicated sampling theory required for such form of testing. Indeed, most of the tests he discusses in chapters 3 and 4 of his 1925 book [6] are *misspecification tests*: tests of departures from Normality, Independence and Homogeneity. Fisher emphasizes the fact that the reliability of every form of inference depend crucially on the validity of the statistical model postulated. The premises of statistical induction in Fisher's sense no longer rely on prior assumptions of 'ignorance', but on testable probabilistic assumptions which concern the observed data; this was a major departure from Pearson's form of enumerative induction relying on prior distributions.

A more complete version of the three problems of the 'reduction of data' is repeated in Fisher's 1925 book [6], which is worth quoting in full with the major additions indicated in italic: "The problems which arise in the reduction of data may thus conveniently be divided into three types:

(i) Problems of Specification, which arise in the choice of the mathematical form of the population. *This is not arbitrary, but requires an understanding of the way in which the data are supposed to, or did in fact, originate. Its further discussion depends on such fields as the theory of Sample Survey, or that of Experimental Design.*

(ii) *When the specification has been obtained, problems of Estimation arise.* These involve the choice among the methods of calculating, from our sample, statistics fit to estimate the unknown parameters of the population.

(iii) Problems of *Distribution* include the mathematical deduction of the exact nature of the distributions in random samples of our estimates of the parameters, *and of the other statistics designed to test the validity of our specification (tests of Goodness of Fit).*" (see ibid. p. 8)

In (i) Fisher makes a clear reference to the *actual Data Generating Mechanism* (DGM), which might often involve specialized knowledge beyond statistics. His view of specification, however, is narrowed down by his focus on data from 'sample surveys' and 'experimental design', where the gap between the actual DGM and the statistical model is not sizeable. This might explain his claim that: "... for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type." In his 1935 book, Fisher states that: "Statistical procedure and experimental design are only two aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to

natural knowledge by experimentation" (p. 3)

In (iii) Fisher adds the derivation of the sampling distributions of misspecification tests as part of the 'problems of distribution'.

In summary, Fisher's view of specification, as a facet of modeling providing the foundation and the overarching framework for *statistical induction*, was a radical departure from Karl Pearson's view of the problem. By interpreting the observed data as 'truly representative' of a prespecified statistical model, Fisher initiated the recasting of *statistical induction* and rendered its premises testable. By ascertaining *statistical adequacy*, using *misspecification tests*, the modeler can ensure the reliability of inductive inference. In addition, his pivotal contributions to the 'problems of Estimation and Distribution', in the form of finite sampling distributions for estimators and test statistics, shifted the emphasis in statistical induction, from enumerative induction and its reliance on asymptotic arguments, to 'reliable procedures' based on finite sample 'ascertainable error probabilities':

"In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." (Fisher [7], p. 14)

This constitutes a clear description of *inductive inference* based on ascertainable error probabilities, under the 'control' of the experimenter, used to assess the 'optimality' of inference procedures. Fisher was the first to realize that for precise (finite sample) 'error probabilities', to be used for calibrating statistical induction, one needs a complete model specification including a distribution assumption. Fisher's most enduring contribution is his devising a general way to 'operationalize' the errors for statistical induction by embedding the *material experiment* into a *statistical model* and define the frequentist error probabilities in the context of the latter. These statistical error probabilities provide a measure of the 'trustworthiness' of the inference procedure: how often it will give rise to true inferences concerning the underlying DGM. That is, the inference is reached by an inductive procedure which, with high probability, will reach true conclusions from true (or approximately true) premises (statistical model). This is in contrast to *induction by enumeration* where the focus is on observed 'events' and not on the 'process' generating the data.

In relation to this, C. S. Peirce put forward a similar view of quantitative induction, almost half a century earlier. This view of statistical induction, was called the error statistical approach by Mayo [12], who has formalized and extended it to include a post-data evaluation of inference in the form of severe testing. Severe testing can be used to address chronic problems associated with Neyman-Pearson testing, including the classic fallacies of acceptance and rejection; see Mayo and Spanos [14].

2.4. Neyman

According to Lehmann [11], Neyman's views on the theory of statistical modeling had three distinct features:

"1. Models of complex phenomena are constructed by combining simple building blocks which, "partly through experience and partly through imagination, appear to us familiar, and therefore, simple." ...

2. An important contribution to the theory of modeling is Neyman's distinction between two types of models: "interpolatory formulae" on the one hand and "explanatory models" on the other. The latter try to provide an explanation of the mechanism underlying the observed phenomena; Mendelian inheritance was Neyman's favorite example. On the other hand an interpolatory formula is based on a convenient and flexible family of distributions or models given a priori, for example the Pearson curves, one of which is selected as providing the best fit to the data. ...

3. The last comment of Neyman's we mention here is that to develop a "genuine explanatory theory" requires substantial knowledge of the scientific background of the problem." (p. 161)

Lehmann's first hand knowledge of Neyman's views on modeling is particularly enlightening. It is clear that Neyman adopted, adapted and extended Fisher's view of statistical modeling. What is especially important for our purposes is to bring out both the similarities as well as the subtle differences with Fisher's view.

Neyman and Pearson [26] built their hypothesis testing procedure in the context of Fisher's approach to statistical modeling and inference, with the notion of a prespecified parametric statistical model providing the cornerstone of the whole inferential edifice. Due primarily to Neyman's experience with empirical modeling in a number of applied fields, including genetics, agriculture, epidemiology and astronomy, his view of statistical models, evolved beyond Fisher's 'infinite populations' in the 1930s into *frequentist* 'chance mechanisms' in the 1950s:

"(ii) Guessing and then verifying the 'chance mechanism', the repeated operations of which produces the observed frequencies. This is a problem of 'frequentist probability theory'. Occasionally, this step is labelled 'model building'. Naturally, the guessed chance mechanism is hypothetical." (Neyman [25], p. 99)

In this quotation we can see a clear statement concerning the nature of specification. Neyman [18] describes statistical modeling as follows: "The application of the theory involves the following steps:

(i) If we wish to treat certain phenomena by means of the theory of probability we must find some element of these phenomena that could be considered as random, following the law of large numbers. This involves a construction of a mathematical model of the phenomena involving one or more probability sets.

(ii) The mathematical model is found satisfactory, or not. This must be checked by observation.

(iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future." (ibid., p. 27)

In this quotation Neyman in (i) demarcates the domain of statistical modeling to *stochastic phenomena*: observed phenomena which exhibit *chance regularity patterns*, and considers statistical (mathematical) models as probabilistic constructs. He also emphasizes the reliance of frequentist inductive inference on the long-run stability of relative frequencies. Like Fisher, he emphasizes in (ii) the testing of the assumptions comprising the statistical model in order to ensure its adequacy. In (iii) he clearly indicates that statistical adequacy is a necessary condition for any inductive inference. This is because the 'error probabilities', in terms of which the optimality of inference is defined, depend crucially on the validity of the model:

"... any statement regarding the performance of a statistical test depends upon the postulate that the observable random variables are random variables and posses

the properties specified in the definition of the set Ω of the admissible simple hypotheses." (Neyman [17], p. 289)

A crucial implication of this is that when the statistical model is misspecified, the actual error probabilities, in terms of which 'optimal' inference procedures are chosen, are likely to be very different from the nominal ones, leading to unreliable inferences; see Spanos [40].

Neyman's experience with modeling *observational data* led him to take statistical modeling a step further and consider the question of *respecifying* the original model whenever it turns out to be inappropriate (statistically inadequate): "Broadly, the methods of bringing about an agreement between the predictions of statistical theory and observations may be classified under two headings:(a) Adaptation of the statistical theory to the enforced circumstances of observation. (b) Adaptation of the experimental technique to the postulates of the theory. The situations referred to in (a) are those in which the observable random variables are largely outside the control of the experimenter or observer." ([17], p. 291)

Neyman goes on to give an example of (a) from his own applied research on the effectiveness of insecticides where the Poisson model was found to be inappropriate: "Therefore, if the statistical tests based on the hypothesis that the variables follow the Poisson Law are not applicable, the only way out of the difficulty is to modify or adapt the theory to the enforced circumstances of experimentation." (ibid., p. 292)

In relation to (b) Neyman continues (ibid., p. 292): "In many cases, particularly in laboratory experimentation, the nature of the observable random variables is much under the control of the experimenter, and here it is usual to adapt the experimental techniques so that it agrees with the assumptions of the theory."

He goes on to give due credit to Fisher for introducing the crucially important technique of *randomization* and discuss its application to the 'lady tasting tea' experiment. Arguably, Neyman's most important extension of Fisher's specification facet of statistical modeling, was his underscoring of *the gap between a statistical model and the phenomena of interest*:

"...it is my strong opinion that no mathematical theory refers exactly to happenings in the outside world and that any application requires a solid bridge over an abyss. The construction of such a bridge consists first, in explaining in what sense the mathematical model provided by the theory is expected to "correspond" to certain actual happenings and second, in checking empirically whether or not the correspondence is satisfactory." ([18], p. 42)

He emphasizes the bridging of the gap between a statistical model and the observable phenomenon of interest, arguing that, beyond *statistical adequacy*, one needs to ensure *substantive adequacy*: the accord between the statistical model and 'reality' must also be adequate: "Since in many instances, the phenomena rather than their models are the subject of scientific interest, the transfer to the phenomena of an inductive inference reached within the model must be something like this: granting that the model M of phenomena P is adequate (or valid, of satisfactory, etc.) the conclusion reached within M applies to P." (Neyman [19], p. 17)

In a purposeful attempt to bridge this gap, Neyman distinguished between a *statistical model* (interpolatory formula) and a *structural model* (see especially Neyman [24], p. 3360), and raised the important issue of *identification* in Neyman [23]: "This particular finding by Polya demonstrated a phenomenon which was unanticipated – two radically different stochastic mechanisms can produce identical distributions

of the same variable X! Thus, the study of this distribution cannot answer the question which of the two mechanisms is actually operating. " ([23], p. 158)

In summary, Neyman's views on statistical modeling elucidated and extended that of Fisher's in several important respects: (a) Viewing statistical models primarily as 'chance mechanisms'. (b) Articulating fully the role of 'error probabilities' in assessing the optimality of inference methods. (c) Elaborating on the issue of *respecification* in the case of statistically *inadequate* models. (d) Emphasizing the gap between a *statistical model* and the *phenomenon of interest*. (e) Distinguishing between *structural* and *statistical models*. (f) Recognizing the problem of *Identification*.

2.5. Lehmann

Lehmann [11] considers the question of 'what contribution statistical theory can potentially make to model specification and construction'. He summarizes the views of both Fisher and Neyman on model specification and discusses the meagre subsequent literature on this issue. His primary conclusion is rather pessimistic: apart from some vague guiding principles, such as simplicity, imagination and the use of past experience, no general theory of modeling seems attainable: "This requirement [to develop a "genuine explanatory theory" requires substantial knowledge of the scientific background of the problem] is agreed on by all serious statisticians but it constitutes of course an obstacle to any general theory of modeling, and is likely a principal reason for Fisher's negative feeling concerning the possibility of such a theory." (Lehmann [11], p. 161)

Hence, Lehmann's source of pessimism stems from the fact that 'explanatory' models place a major component of model specification beyond the subject matter of the statistician: "An explanatory model, as is clear from the very nature of such models, requires detailed knowledge and understanding of the substantive situation that the model is to represent. On the other hand, an empirical model may be obtained from a family of models selected largely for convenience, on the basis solely of the data without much input from the underlying situation." (p. 164)

In his attempt to demarcate the potential role of statistics in a general theory of modeling, Lehmann [11], p. 163, discusses the difference in the basic objectives of the two types of models, arguing that: "Empirical models are used as a guide to action, often based on forecasts ... In contrast, explanatory models embody the search for the basic mechanism underlying the process being studied; they constitute an effort to achieve *understanding*."

In view of these, he goes on to pose a crucial question (Lehmann [11], p. 161-2): "Is applied statistics, and more particularly model building, an art, with each new case having to be treated from scratch, ..., completely on its own merits, or does theory have a contribution to make to this process?"

Lehmann suggests that one (indirect) way a statistician can contribute to the theory of modeling is via: "... the existence of a reservoir of models which are well understood and whose properties we know. Probability theory and statistics have provided us with a rich collection of such models." (p. 161)

Assuming the existence of a sizeable reservoir of models, the problem still remains 'how does one make a choice among these models?' Lehmann's view is that the current methods on model selection *do not* address this question:

"Procedures for choosing a model not from the vast storehouse mentioned in (2.1 Reservoir of Models) but from a much more narrowly defined class of models

are discussed in the theory of model selection. A typical example is the choice of a regression model, for example of the best dimension in a nested class of such models. ... However, this view of model selection ignores a preliminary step: the specification of the class of models from which the selection is to be made." (p. 162)

This is a most insightful comment because a closer look at model selection procedures suggests that the problem of model specification is largely assumed away by commencing the procedure by assuming that the prespecified family of models includes the true model; see Spanos [42].

In addition to differences in their nature and basic objectives, Lehmann [11] argues that explanatory and empirical models pose very different problems for model validation: "The difference in the aims and nature of the two types of models [empirical and explanatory] implies very different attitudes toward checking their validity. Techniques such as goodness of fit test or cross validation serve the needs of checking an empirical model by determining whether the model provides an adequate fit for the data. Many different models could pass such a test, which reflects the fact that there is not a unique correct empirical model. On the other hand, ideally there is only one model which at the given level of abstraction and generality describes the mechanism or process in question. To check its accuracy requires identification of the details of the model and their functions and interrelations with the corresponding details of the real situation." (ibid. pp. 164-5)

Lehmann [11] concludes the paper on a more optimistic note by observing that statistical theory has an important role to play in model specification by extending and enhancing: (1) the reservoir of models, (2) the model selection procedures, as well as (3) utilizing different classifications of models. In particular, in addition to the subject matter, every model also has a 'chance regularity' dimension and probability theory can play a crucial role in 'capturing' this. This echoes Neyman [21], who recognized the problem posed by explanatory (stochastic) models, but suggested that probability theory does have a crucial role to play: "The problem of stochastic models is of prime interest but is taken over partly by the relevant substantive disciplines, such as astronomy, physics, biology, economics, etc., and partly by the theory of probability. In fact, the primary subject of the modern theory of probability may be described as the study of properties of particular chance mechanisms." (p. 447)

Lehmann's discussion of model specification suggests that *the* major stumbling block in the development of a general modeling procedure is the substantive knowledge, beyond the scope of statistics, called for by explanatory models; see also Cox and Wermuth [3]. To be fair, both Fisher and Neyman in their writings seemed to suggest that statistical model specification is based on an amalgam of substantive and statistical information.

Lehmann [11] provides a key to circumventing this stumbling block: "Examination of some of the classical examples of revolutionary science shows that the eventual explanatory model is often reached in stages, and that in the earlier efforts one may find models that are descriptive rather than fully explanatory. ... This is, for example, true of Kepler whose descriptive model (laws) of planetary motion precede Newton's explanatory model." (p. 166).

In this quotation, Lehmann acknowledges that a descriptive (statistical) model can have 'a life of its own', separate from substantive subject matter information. However, the question that arises is: 'what is such model a description of?' As argued in the next section, in the context of the Probabilistic Reduction (PR) framework, such a model provides a description of the systematic *statistical infor-*

mation exhibited by data $\mathbf{Z}:=(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$. This raises another question 'how does the substantive information, when available, enter statistical modeling?' Usually substantive information enters the empirical modeling as restrictions on a statistical model, when the structural model, carrying the substantive information, is embedded into a statistical model. As argued next, when these restrictions are data-acceptable, assessed in the context of a statistically adequate model, they give rise to an *empirical model* (see Spanos, [31]), which is both statistically as well as substantively meaningful.

3. The Probabilistic Reduction (PR) Approach

The foundations and overarching framework of the *PR approach* (Spanos, [31]–[42]) has been greatly influenced by Fisher's recasting of statistical induction based on the notion of a statistical model, and calibrated in terms of frequentist error probabilities, Neyman's extensions of Fisher's paradigm to the modeling of observational data, and Kolmogorov's crucial contributions to the theory of stochastic processes. The emphasis is placed on learning from data about observable phenomena, and on actively encouraging thorough probing of the different ways an inference might be in error, by localizing the error probing in the context of different models; see Mayo [12]. Although the broader problem of bridging the gap between theory and data using a sequence of interrelated models (see Spanos, [31], p. 21) is beyond the scope of this paper, it is important to discuss how the separation of substantive and statistical information can be achieved in order to make a case for treating statistical models as *canonical models* which can be used in conjunction with substantive information from any applied field.

It is widely recognized that stochastic phenomena amenable to empirical modeling have two *interrelated* sources of information, the *substantive subject matter* and the *statistical information* (chance regularity). What is not so apparent is how these sources of information are integrated in the context of empirical modeling. The PR approach treats the statistical and substantive information as complementary and, *ab initio,* are described separately in the form of a statistical and a structural model, respectively. The key for this *ab initio* separation is provided by viewing a statistical model generically as a particular parameterization of a stochastic processes $\{\mathbf{Z}_t, t\in\mathbb{T}\}$ underlying the data $\mathbf{Z}$, which, under certain conditions, can nest (parametrically) the structural model(s) in question. This gives rise to a framework for integrating the various facets of modeling encountered in the discussion of the early contributions by Fisher and Neyman: *specification, misspecification testing, respecification, statistical adequacy, statistical (inductive) inference, and identification.*

3.1. Structural vs. statistical models

It is widely recognized that most stochastic phenomena (the ones exhibiting chance regularity patterns) are commonly influenced by a very large number of contributing factors, and that explains why theories are often dominated by *ceteris paribus* clauses. The idea behind a *theory* is that in explaining the behavior of a variable, say y_k, one demarcates the segment of reality to be modeled by selecting the primary influencing factors $\mathbf{x}_k$, cognizant of the fact that there might be numerous other potentially relevant factors ξ_k (observable and unobservable) that jointly determine the behavior of y_k via a *theory model:*

$$y_k = h^*(\mathbf{x}_k, \xi_k), \quad k \in \mathbb{N}, \tag{3.1}$$

where $h^*(.)$ represents the *true* behavioral relationship for y_k. The guiding principle in selecting the variables in $\mathbf{x}_k$ is to ensure that they collectively account for the *systematic behavior* of y_k, and the *unaccounted factors* $\boldsymbol{\xi}_k$ represent non-essential disturbing influences which have only a *non-systematic* effect on y_k. This reasoning transforms (3.1) into a *structural model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\phi}) + \epsilon(\mathbf{x}_k \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \tag{3.2}$$

where $h(.)$ denotes the postulated functional form, $\boldsymbol{\phi}$ stands for the *structural parameters* of interest, and $\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k)$ represents the *structural error term,* viewed as a function of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$. By definition the error term process is:

$$\{\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k \boldsymbol{\phi}),\ k \in \mathbb{N}\}, \tag{3.3}$$

and represents all unmodeled influences, *intended* to be a *white-noise* (non-systematic) process, i.e. for all possible values $(\mathbf{x}_k \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}}$:

[i] $E[\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k)] = 0$, [ii] $E[\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k)^2] = \sigma^2_\epsilon$, [iii] $E[\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k) \cdot \epsilon(\mathbf{x}_j \boldsymbol{\xi}_j)] = 0$, for $k \neq j$.

In addition, (3.2) represents a 'nearly isolated' generating mechanism in the sense that its error should be uncorrelated with the modeled influences (systematic component $h(\mathbf{x}_k \boldsymbol{\phi})$), i.e. [iv] $E[\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k) \cdot h(\mathbf{x}_k \boldsymbol{\phi})] = 0$; the term 'nearly' refers to the non-deterministic nature of the isolation - see Spanos ([31], [35]).

In summary, a *structural model* provides an 'idealized' substantive description of the phenomenon of interest, in the form of a 'nearly isolated' mathematical system (3.2). The specification of a structural model comprises several choices: (a) the demarcation of the segment of the phenomenon of interest to be captured, (b) the important aspects of the phenomenon to be measured, and (c) the extent to which the inferences based on the structural model are germane to the phenomenon of interest. The kind of *errors* one can probe for in the context of a structural model concern the choices (a)–(c), including the form of $h(\mathbf{x}_k; \boldsymbol{\phi})$ and the circumstances that render the error term potentially systematic, such as the presence of relevant factors, say $\mathbf{w}_k$, in $\boldsymbol{\xi}_k$ that might have a systematic effect on the behavior of y_t; see Spanos [41].

It is important to emphasize that (3.2) depicts a 'factual' Generating Mechanism (GM), which aims to approximate the *actual data GM*. However, the assumptions [i]–[iv] of the structural error are *non-testable* because their assessment would involve verification for all possible values $(\mathbf{x}_k \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}}$. To render them testable one needs to embed this structural into a statistical model; a crucial move that often goes unnoticed. Not surprisingly, the nature of the embedding itself depends crucially on whether the data $\mathbf{Z} := (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ are the result of an experiment or they are non-experimental (observational) in nature.

3.2. Statistical models and experimental data

In the case where one can perform experiments, 'experimental design' techniques, might allow one to operationalize the 'near isolation' condition (see Spanos, [35]), including the *ceteris paribus* clauses, and ensure that the *error term* is no longer a function of $(\mathbf{x}_k \boldsymbol{\xi}_k)$, but takes the generic form:

$$\epsilon(\mathbf{x}_k \boldsymbol{\xi}_k) = \varepsilon_k \sim \mathsf{IID}(0, \sigma^2), \quad k = 1, 2, \ldots, n. \tag{3.4}$$

For instance, *randomization* and *blocking* are often used to 'neutralize' the phenomenon from the potential effects of $\boldsymbol{\xi}_k$ by ensuring that these uncontrolled factors

cancel each other out; see Fisher [7]. As a direct result of the experimental 'control' via (3.4) the structural model (3.2) is essentially transformed into a *statistical model:*

$$y_k = h(\mathbf{x}_k;\boldsymbol{\theta}) + \varepsilon_k, \quad \varepsilon_k \sim \mathsf{IID}(0,\sigma^2),\ k=1,2,\ldots,n. \tag{3.5}$$

The statistical error terms in (3.5) are qualitatively very different from the structural errors in (3.2) because they no longer depend on $(\mathbf{x}_k\boldsymbol{\xi}_k)$; the clause 'for all $(\mathbf{x}_k\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}$' has been rendered irrelevant. The most important aspect of embedding the structural (3.2) into the statistical model (3.5) is that, in contrast to [i]–[iv] for $\{\epsilon(\mathbf{x}_k\boldsymbol{\xi}_k),\ k\in\mathbb{N}\}$, the probabilistic assumptions $\mathsf{IID}(0,\sigma^2)$ concerning the *statistical error term* are rendered *testable*. That is, by operationalizing the 'near isolation' condition via (3.4), the error term has been tamed. For more precise inferences one needs to be more specific about the probabilistic assumptions defining the statistical model, including the functional form $h(.)$. This is because the more finical the probabilistic assumptions (the more constricting the statistical premises) the more precise the inferences; see Spanos [40].

The ontological status of the statistical model (3.5) is different from that of the structural model (3.2) in so far as (3.4) has operationalized the 'near isolation' condition. The statistical model has been 'created' as a result of the experimental design and control. As a consequence of (3.4) the *informational universe of discourse* for the statistical model (3.5) has been delimited to the probabilistic information relating to the observables $\mathbf{Z}_k$. This probabilistic structure, according to Kolmogorov's consistency theorem, can be fully described, under certain mild regularity conditions, in terms of the joint distribution $D(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_n;\boldsymbol{\phi})$; see Doob [4]. It turns out that a statistical model can be viewed as a *parameterization* of the presumed probabilistic structure of the process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$; see Spanos ([31], [35]).

In summary, a *statistical model* constitutes an 'idealized' *probabilistic description* of a stochastic process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$, giving rise to data $\mathbf{Z}$, in the form of an internally consistent set of probabilistic assumptions, chosen to ensure that this data constitute a 'truly typical realization' of $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$.

In contrast to a *structural model*, once $\mathbf{Z}_k$ is chosen, a *statistical model* relies exclusively on the statistical information in $D(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_n;\boldsymbol{\phi})$, that 'reflects' the chance regularity patterns exhibited by the data. Hence, a statistical model acquires 'a life of its own' in the sense that it constitutes a self-contained GM defined exclusively in terms of probabilistic assumptions pertaining to the observables $\mathbf{Z}_k{:=}(y_k,\mathbf{X}_k)$. For example, in the case where $h(\mathbf{x}_k;\boldsymbol{\phi}){=}\beta_0{+}\boldsymbol{\beta}_1^{\top}\mathbf{x}_k$, and $\varepsilon_k \backsim \mathsf{N}(.,.)$, (3.5) becomes the *Gauss Linear model,* comprising the statistical GM:

$$y_k = \beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{x}_k + u_k,\ k\in\mathbb{N}, \tag{3.6}$$

together with the probabilistic assumptions (Spanos [31]):

$$y_k \backsim \mathsf{NI}(\beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{x}_k,\ \sigma^2),\ k\in\mathbb{N}, \tag{3.7}$$

where $\boldsymbol{\theta} := (\beta_0,\boldsymbol{\beta}_1,\sigma^2)$ is assumed to be k-invariant, and 'NI' stands for 'Normal, Independent'.

3.3. Statistical models and observational data

This is the case where the observed data on $(y_t,\mathbf{x}_t)$ are the result of an ongoing actual data generating process, undisturbed by any experimental control or intervention. In this case the route followed in (3.4) in order to render the *statistical*

error term (a) free of $(\mathbf{x}_t,\xi_t)$, and (b) non-systematic in a statistical sense, is no longer feasible. It turns out that sequential *conditioning* supplies the primary tool in modeling observational data because it provides an alternative way to ensure the non-systematic nature of the statistical error term without controls and intervention.

It is well-known that sequential conditioning provides a general way to transform an arbitrary stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ into a *Martingale Difference (MD) process* relative to an increasing sequence of sigma-fields $\{\mathfrak{D}_t,\ t\in\mathbb{T}\}$; a modern form of a non-systematic error process (Doob, [4]). This provides the key to an alternative approach to specifying statistical models in the case of non-experimental data by replacing the 'controls' and 'interventions' with the choice of the *relevant conditioning information set* $\mathfrak{D}_t$ that would render the error term a MD; see Spanos [31].

As in the case of experimental data the universe of discourse for a statistical model is fully described by the joint distribution $D(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_T;\boldsymbol{\phi})$, $\mathbf{Z}_t{:=}(y_t,\mathbf{X}_t^\top)^\top$. Assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ has bounded moments up to order two, one can choose the conditioning information set to be:

$$\mathfrak{D}_{t-1}=\sigma\left(y_{t-1},y_{t-2},\ldots,y_1,\mathbf{X}_t,\mathbf{X}_{t-1},\ldots,\mathbf{X}_1\right). \tag{3.8}$$

This renders the error process $\{u_t,\ t\in\mathbb{T}\}$, defined by:

$$u_t=y_t-E(y_t|\mathfrak{D}_{t-1}), \tag{3.9}$$

a MD process relative to $\mathfrak{D}_{t-1}$, irrespective of the probabilistic structure of $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$; see Spanos [36]. This error process is based on $D(y_t|\,\mathbf{X}_t,\mathbf{Z}_{t-1}^0;\psi_{1t})$, where $\mathbf{Z}_{t-1}^0{:=}(\mathbf{Z}_{t-1},\ldots,\mathbf{Z}_1)$, which is directly related to $D(\mathbf{Z}_1,\ldots,\mathbf{Z}_T;\phi)$ via:

$$\begin{aligned}
&D(\mathbf{Z}_1,\ldots,\mathbf{Z}_T;\boldsymbol{\phi})\\
&\quad=D(\mathbf{Z}_1;\psi_1)\textstyle\prod_{t=2}^{T}D_t(\mathbf{Z}_t|\,\mathbf{Z}_{t-1}^0;\psi_t)\\
&\quad=D(\mathbf{Z}_1;\psi_1)\textstyle\prod_{t=2}^{T}D_t(y_t|\,\mathbf{X}_t,\mathbf{Z}_{t-1}^0;\psi_{1t})\cdot D_t(\mathbf{X}_t|\,\mathbf{Z}_{t-1}^0;\psi_{2t}).
\end{aligned} \tag{3.10}$$

The Greek letters ϕ and ψ are used to denote the unknown parameters of the distribution in question. This sequential conditioning gives rise to a statistical GM of the form:

$$y_t=E(y_t\mid\mathfrak{D}_{t-1})+u_t,\ t\in\mathbb{T}, \tag{3.11}$$

which is *non-operational* as it stands because without further restrictions on the process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$, the systematic component $E(y_t\mid\mathfrak{D}_{t-1})$ cannot be specified explicitly. For operational models one needs to postulate some probabilistic structure for $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ that would render the data $\mathbf{Z}$ a 'truly typical' realization thereof. These assumptions come from a menu of three broad categories: (D) **Distribution,** (M) **Dependence,** (H) **Heterogeneity**; see Spanos ([34]–[38]).

Example. *The Normal/Linear Regression model* results from the reduction (3.10) by assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ is a NIID vector process. These assumptions ensure that the relevant information set that would render the error process a MD is reduced from $\mathfrak{D}_{t-1}$ to $\mathfrak{D}_t^x{=}\{\mathbf{X}_t{=}\,\mathbf{x}_t\}$, ensuring that:

$$(u_t|\,\mathbf{X}_t{=}\,\mathbf{x}_t)\sim\mathsf{NIID}(0,\sigma^2),\ k{=}1,2,\ldots,T. \tag{3.12}$$

This is analogous to (3.4) in the case of experimental data, but now the error term has been operationalized by a judicious choice of $\mathfrak{D}_t^x$. The Linear Regression model comprises the statistical GM:

$$y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + u_t,\ t \in \mathbb{T}, \tag{3.13}$$

$$(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t) \sim \mathsf{NI}(\beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t,\ \sigma^2),\ t \in \mathbb{T}, \tag{3.14}$$

where $\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ is assumed to be t-invariant; see Spanos [35].

The probabilistic perspective gives a statistical model 'a life of its own' in the sense that the probabilistic assumptions in (3.14) bring to the table *statistical information* which supplements, and can be used to assess the appropriateness of, the *substantive subject matter information*. For instance, in the context of the structural model $h(\mathbf{x}_t;\boldsymbol{\phi})$ is determined by the theory. In contrast, in the context of a statistical model it is determined by the probabilistic structure of the process $\{\mathbf{Z}_t,\ t{\in}\mathbb{T}\}$ via $h(\mathbf{x}_t;\boldsymbol{\theta}){=}E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t)$, which, in turn, is determined by the joint distribution $D(y_t,\mathbf{X}_t;\psi)$; see Spanos [36].

An important aspect of embedding a structural into a statistical model is to ensure (whenever possible) that the former can be viewed as a *reparameterization/restriction* of the latter. The structural model is then tested against the benchmark provided by a statistically adequate model. *Identification* refers to being able to define $\boldsymbol{\phi}$ uniquely in terms of $\boldsymbol{\theta}$. Often $\boldsymbol{\theta}$ has more parameters than $\boldsymbol{\phi}$ and the embedding enables one to test the validity of the additional restrictions, known as over-identifying restrictions; see Spanos [33].

3.4. Kepler's first law of planetary motion revisited

In an attempt to illustrate some of the concepts and procedures introduced in the PR framework, we revisit Lehmann's [11] example of Kepler's statistical model predating, by more than 60 years, the eventual structural model proposed by Newton.

Kepler's law of planetary motion was originally just an *empirical regularity* that he 'deduced' from Brahe's data, stating that the motion of any planet around the sun is elliptical. That is, the loci of the motion in *polar coordinates* takes the form $(1/r){=}\alpha_0+\alpha_1\cos\vartheta$, where r denotes *the distance of the planet from the sun*, and ϑ denotes *the angle between the line joining the sun and the planet and the principal axis of the ellipse*. Defining the observable variables by $y := (1/r)$ and $x := \cos\vartheta$, Kepler's empirical regularity amounted to an estimated linear regression model:

$$y_t = \underset{(.000002)}{0.662062} + \underset{(.000003)}{.061333}x_t + \widehat{u}_t,\ R^2 = .999,\ s = .0000111479; \tag{3.15}$$

these estimates are based on Kepler's original 1609 data on Mars with $n = 28$. Formal misspecification tests of the model assumptions in (3.14) (Section 3.3), indicate that the estimated model is statistically adequate; see Spanos [39] for the details.

Substantive interpretation was bestowed on (3.15) by *Newton's law of universal gravitation*: $F{=}\frac{G(m\cdot M)}{r^2}$, where F is the force of attraction between two bodies of *mass* m (planet) and M (sun), G is a *constant of gravitational attraction,* and r is the *distance* between the two bodies, in the form of a *structural model:*

$$Y_k = \alpha_0 + \alpha_1 X_k + \epsilon(x_k,\xi_k),\ k{\in}\mathbb{N}, \tag{3.16}$$

where the parameters (α_0, α_1) are given a *structural interpretation:* $\alpha_0 = \frac{MG}{4\kappa^2}$, where κ denotes Kepler's constant, $\alpha_1 = (\frac{1}{d} - \alpha_0)$, d denotes the shortest distance between the planet and the sun. The error term $\epsilon(x_k, \xi_k)$ also enjoys a structural interpretation in the form of *unmodeled effects*; its assumptions [i]–[iv] (Section 3.1) will be *inappropriate* in cases where (a) the data suffer from 'systematic' observation errors, and there are significant (b) *third body* and/or (c) *general relativity* effects.

3.5. Revisiting certain issues in empirical modeling

In what follows we indicate very briefly how the PR approach can be used to shed light on certain crucial issues raised by Lehmann [11] and Cox [1].

Specification: a 'Fountain' of statistical models. The PR approach broadens Lehmann's *reservoir of models* idea to the set of all possible statistical models $\mathcal{P}$ that could (potentially) have given rise to data $\mathbf{Z}$. The statistical models in $\mathcal{P}$ are *characterized* by their reduction assumptions from three broad categories: *Distribution, Dependence,* and *Heterogeneity.* This way of viewing statistical models provides (i) a systematic way to *characterize statistical models*, (different from Lehmann's) and, at the same time it offers (ii) a general procedure to generate new statistical models.

The capacity of the PR approach to generate new statistical models is demonstrated in Spanos [36], ch. 7, were several bivariate distributions are used to derive different regression models via (3.10); this gives rise to several non-linear and/or heteroskedastic regression models, most of which remain unexplored. In the same vein, the reduction assumptions of (D) *Normality,* (M) *Markov dependence,* and (H) *Stationarity*, give rise to Autoregressive models; see Spanos ([36], [38]).

Spanos [34] derives a new family of Linear/heteroskedastic regression models by replacing the Normal in (3.10) with the Student's t distribution. When the IID assumptions were also replaced by Markov dependence and Stationarity, a surprising family of models emerges that extends the ARCH formulation; see McGuirk et al [15], Heracleous and Spanos [8].

Model validation: statistical vs. structural adequacy. The PR approach also addresses Lehmann's concern that structural and statistical models 'pose very different problems for model validation'; see Spanos [41]. The purely probabilistic construal of statistical models renders *statistical adequacy* the only relevant criterion for model validity is *statistical adequacy.* This is achieved by thorough *misspecification testing* and *respecification*; see Mayo and Spanos [13].

MisSpecification (M-S) testing is different from Neyman and Pearson (N–P) testing in one important respect. N–P testing assumes that the prespecified statistical model class $\mathcal{M}$ includes the true model, say $f_0(\mathbf{z})$, and probes *within the boundaries* of this model using the hypotheses:

$$H_0\colon f_0(\mathbf{z}) \in \mathcal{M}_0 \text{ vs. } H_1\colon f_0(\mathbf{z}) \in \mathcal{M}_1,$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ form a partition of $\mathcal{M}$. In contrast, M-S testing probes *outside the boundaries* of the prespecified model:

$$H_0\colon f_0(\mathbf{z}) \in \mathcal{M} \text{ vs. } \overline{H}_0\colon f_0(\mathbf{z}) \in [\mathcal{P} - \mathcal{M}],$$

where $\mathcal{P}$ denotes the set of all possible statistical models, rendering them Fisherian type *significance tests.* The problem is how one can operationalize $\mathcal{P} - \mathcal{M}$ in order to

probe thoroughly for possible departures; see Spanos [36]. Detection of departures from the null in the direction of, say $\mathcal{P}_1 \subset [\mathcal{P}-\mathcal{M}]$, is sufficient to deduce that the null is false but not to deduce that $\mathcal{P}_1$ is true; see Spanos [37]. More formally, $\mathcal{P}_1$ *has not passed a severe test*, since its own statistical adequacy has not been established; see Mayo and Spanos ([13], [14]).

On the other hand, validity for a structural model refers to *substantive adequacy*: a combination of *data-acceptability* on the basis of a statistically adequate model, and *external validity* - how well the structural model 'approximates' the reality it aims to explain. Statistical adequacy is a precondition for the assessment of substantive adequacy because without it no reliable inference procedures can be used to assess substantive adequacy; see Spanos [41].

Model specification vs. model selection. The PR approach can shed light on Lehmann's concern about *model specification* vs. *model selection*, by underscoring the fact that the primary criterion for model specification within $\mathcal{P}$ is statistical adequacy, not goodness of fit. As pointed out by Lehmann [11], the current *model selection* procedures (see Rao and Wu, [30], for a recent survey) do not address the original statistical model specification problem. One can make a strong case that Akaike-type model selection procedures assume the statistical model specification problem solved. Moreover, when the statistical adequacy issue is addressed, these model selection procedure becomes superfluous; see Spanos [42].

Statistical Generating Mechanism (GM). It is well-known that a statistical model can be specified fully in terms of the joint distribution of the observable random variables involved. However, if the statistical model is to be related to any structural models, it is imperative to be able to specify a statistical GM which will provide the bridge between the two models. This is succinctly articulated by Cox [1]:

"The essential idea is that if the investigator cannot use the model directly to simulate artificial data, how can "Nature" have used anything like that method to generate real data?" (p. 172)

The PR specification of statistical models brings the statistical GM based on the orthogonal decomposition $y_t = E(y_t|\mathfrak{D}_{t-1})+u_t$ in (3.11) to the forefront. The onus is on the modeler to choose (i) an appropriate probabilistic structure for $\{y_t,\ t \in \mathbb{T}\}$, and (ii) the associated information set $\mathfrak{D}_{t-1}$, relative to which the error term is rendered a martingale difference (MD) process; see Spanos [36].

The role of exploratory data analysis. An important feature of the PR approach is to render the use of graphical techniques and exploratory data analysis (EDA), more generally, an integral part of statistical modeling. EDA plays a crucial role in the specification, M-S testing and respecification facets of modeling. This addresses a concern raised by Cox [1] that:

"... the separation of 'exploratory data analysis' from 'statistics' are counter-productive." (ibid., p. 169)

4. Conclusion

Lehmann [11] raised the question whether the presence of substantive information subordinates statistical modeling to other disciplines, precluding statistics from having its own intended scope. This paper argues that, despite the uniqueness of

every modeling endeavor arising from the *substantive subject matter information*, all forms of statistical modeling share certain generic aspects which revolve around the notion of *statistical information*. The key to upholding the integrity of both sources of information, as well as ensuring the reliability of their fusing, is a purely probabilistic construal of statistical models in the spirit of Fisher and Neyman. The PR approach adopts this view of specification and accommodates the related facets of modeling: *misspecification testing* and *respecification*.

The PR modeling framework gives the statistician a *pivotal role* and extends the intended scope of statistics, without relegating the role of substantive information in empiridal modeling. The judicious use of probability theory, in conjunction with graphical techniques, can transform the specification of statistical models into purpose-built conjecturing which can be assessed subsequently. In addition, thorough misspecification testing can be used to assess the appropriateness of a statistical model, in order to ensure the reliability of inductive inferences based upon it. Statistically adequate models *have a life of their own* in so far as they can be (sometimes) the ultimate objective of modeling or they can be used to establish *empirical regularities* for which substantive explanations need to account; see Cox [1]. By embedding a structural into a statistically adequate model and securing substantive adequacy, confers upon the former statistical meaning and upon the latter substantive meaning, rendering learning from data, using statistical induction, a reliable process.

References

[1] Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science*, **5**, 169–174.

[2] Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman & Hall, London.

[3] Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. CRC Press, London.

[4] Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.

[5] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* **222**, 309–368.

[6] Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[7] Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

[8] Heracleous, M. and A. Spanos (2006). The Student's t dynamic linear regression: re-examining volatility modeling. *Advances in Econometrics*. **20**, 289–319.

[9] Lahiri, P. (2001). *Model Selection*. Institute of Mathematical Statistics, Ohio.

[10] Lehmann, E. L. (1986). *Testing statistical hypotheses,* 2nd edition. Wiley, New York.

[11] Lehmann, E. L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* **5**, 160–168.

[12] Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago.

[13] Mayo, D. G. and A. Spanos (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science* **71**, 1007–1025.

[14] Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a

Neyman–Pearson philosophy of induction. *The British Journal of the Philosophy of Science* **57**, 321–356.

[15] McGuirk, A., J. Robertson and A. Spanos (1993). Modeling exchange rate dynamics: non-linear dependence and thick tails. *Econometric Reviews* **12**, 33–63.

[16] Mills, F. C. (1924). *Statistical Methods*. Henry Holt and Co., New York.

[17] Neyman, J. (1950). *First Course in Probability and Statistics*, Henry Holt, New York.

[18] Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd edition. U.S. Department of Agriculture, Washington.

[19] Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics* **VIII**, 13–46.

[20] Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Revue Inst. Int. De Stat.* **25**, 7–22.

[21] Neyman, J. (1969). Behavioristic points of view on mathematical statistics. In *On Political Economy and Econometrics: Essays in Honour of Oskar Lange*. Pergamon, Oxford, 445–462.

[22] Neyman, J. (1971). Foundations of behavioristic statistics. In *Foundations of Statistical Inference*, Godambe, V. and Sprott, D., eds. Holt, Rinehart and Winston of Canada, Toronto, 1–13.

[23] Neyman, J. (1976a). The emergence of mathematical statistics. In *On the History of Statistics and Probability*, Owen, D. B., ed. Dekker, New York, ch. 7.

[24] Neyman, J. (1976b). A structural model of radiation effects in living cells. *Proceedings of the National Academy of Sciences*. **10**, 3360–3363.

[25] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131.

[26] Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. of the Royal Society A* **231**, 289–337.

[27] Pearson, K. (1895). Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London Series A* **186**, 343–414.

[28] Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika* **XIII**, 1–16.

[29] Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science* **7**, 34–48.

[30] Rao, C. R. and Y. Wu (2001). On Model Selection. In P. Lahiri (2001), 1–64.

[31] Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*. Cambridge University Press, Cambridge.

[32] Spanos, A. (1989). On re-reading Haavelmo: a retrospective view of econometric modeling. *Econometric Theory*. **5**, 405–429.

[33] Spanos, A. (1990). The simultaneous equations model revisited: statistical adequacy and identification. *Journal of Econometrics* **44**, 87–108.

[34] Spanos, A. (1994). On modeling heteroskedasticity: the Student's t and elliptical regression models. *Econometric Theory* **10**, 286–315.

[35] Spanos, A. (1995). On theory testing in Econometrics: modeling with nonexperimental data. *Journal of Econometrics* **67**, 189–226.

[36] Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge University Press, Cambridge.

[37] Spanos, A. (2000). Revisiting data mining: 'hunting' with or without a license. *The Journal of Economic Methodology* **7**, 231–264.

[38] Spanos, A. (2001). Time series and dynamic models. *A Companion to Theoretical Econometrics*, edited by B. Baltagi. Blackwell Publishers, Oxford, 585–609, chapter 28.

[39] Spanos, A. (2005). Structural vs. statistical models: Revisiting Kepler's law of planetary motion. Working paper, Virginia Tech.

[40] Spanos, A. (2006a). Econometrics in retrospect and prospect. In *New Palgrave Handbook of Econometrics*, vol. 1, Mills, T.C. and K. Patterson, eds. MacMillan, London. 3–58.

[41] Spanos, A. (2006b). Revisiting the omitted variables argument: Substantive vs. statistical reliability of inference. *Journal of Economic Methodology* **13**, 174–218.

[42] Spanos, A. (2006c). The curve-fitting problem, Akaike-type model selection, and the error statistical approach. Working paper, Virginia Tech.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 120–130
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000428

Modeling inequality and spread in multiple regression*

Rolf Aaberge[1], Steinar Bjerve[2] and Kjell Doksum[3]

Statistics Norway, University of Oslo and University of Wisconsin, Madison

Abstract: We consider concepts and models for measuring inequality in the distribution of resources with a focus on how inequality varies as a function of covariates. Lorenz introduced a device for measuring inequality in the distribution of income that indicates how much the incomes below the u^{th} quantile fall short of the egalitarian situation where everyone has the same income. Gini introduced a summary measure of inequality that is the average over u of the difference between the Lorenz curve and its values in the egalitarian case. More generally, measures of inequality are useful for other response variables in addition to income, e.g. wealth, sales, dividends, taxes, market share and test scores. In this paper we show that a generalized van Zwet type dispersion ordering for distributions of positive random variables induces an ordering on the Lorenz curve, the Gini coefficient and other measures of inequality. We use this result and distributional orderings based on transformations of distributions to motivate parametric and semiparametric models whose regression coefficients measure effects of covariates on inequality. In particular, we extend a parametric Pareto regression model to a flexible semiparametric regression model and give partial likelihood estimates of the regression coefficients and a baseline distribution that can be used to construct estimates of the various conditional measures of inequality.

1. Introduction

Measures of inequality provide quantifications of how much the distribution of a resource Y deviates from the egalitarian situation where everyone has the same amount of the resource. The coefficients in location or location-scale regression models are not particularly informative when attention is turned to the influence of covariates on inequality. In this paper we consider regression models that are not location-scale regression models and whose coefficients are associated with the effect of covariates on inequality in the distribution of the response Y.

We start in Section 2.1 by discussing some familiar and some new measures of inequality. Then in Section 2.2 we relate the properties of these measures to a statistical ordering of distributions based on transformations of random variables that

*We would like to thank Anne Skoglund for typing and editing the paper and Javier Rojo and an anonymous referee for helpful comments. Rolf Aaberge gratefully acknowledges ICER in Torino for financial support and excellent working conditions and Steinar Bjerve for the support of The Wessmann Society during the course of this work. Kjell Doksum was supported in part by NSF grants DMS-9971301 and DMS-0505651.

[1]Research Department, Statistics Norway, P.O. Box 813, Dep., N-0033, Oslo, Norway, e-mail: Rolf.Aaberge@ssb.no

[2]Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, 0316, Oslo, Norway, e-mail: steinar@math.uio.no

[3]Department of Statistics, University of Wisconsin, 1300 University Ave, Madison, WI 53706, USA, e-mail: doksum@stat.wisc.edu

AMS 2000 subject classifications: primary 62F99, 62G99, 61J99; secondary 91B02, 91C99.

Keywords and phrases: Lorenz curve, Gini index, Bonferroni index, Lehmann model, Cox regression, Pareto model.

is equivalent to defining the distribution H of the response Z to have more resource inequality than the distribution F of Y if Z has the same distribution as $q(Y)Y$ for some positive nondecreasing function $q(\cdot)$. Then we show that this ordering implies the corresponding ordering of each measure of inequality. We also consider orderings of distributions based on transformations of distribution functions and relate them to inequality. These notions and results assist in the construction of regression models with coefficients that relate to the concept of inequality.

Section 3 shows that scaled power transformation models with the power parameter depending on covariates provide regression models where the coefficients relate to the concept of resource inequality. Two interesting particular cases are the Pareto and the log normal transformation regression models. For these models the Lorenz curve for the conditional distribution of Y given covariate values takes a particularly simple and intuitive form. We discuss likelihood methods for the statistical analysis of these models.

Finally, in Section 4 we consider semiparametric Lehmann and Cox type models that are based on power transformations of a baseline distribution F_0, or of $1 - F_0$, where the power parameter is a function of the covariates. In particular, we consider a power transformation model of the form

$$F(y) = 1 - (1 - F_0(y))^{\alpha(\boldsymbol{x})}, \tag{1.1}$$

where $\alpha(\boldsymbol{x})$ is a parametric function depending on a vector $\boldsymbol{\beta}$ of regression coefficients and an observed vector of covariates $\boldsymbol{x}$. This is an extension of the Pareto regression model to a flexible semiparametric model. For this model we present theoretical and empirical formulas for inequality measures and point out that computations can be based on available software.

2. Measures of inequality and spread

2.1. Defining curves and measures of inequality

The Lorenz curve (LC) is defined (Lorenz [19]) to be the proportion of the total amount of wealth that is owned by the "poorest" $100 \times u$ percent of the population. More precisely, let the random income $Y > 0$ have the distribution function $F(y)$, let $F^{-1}(y) = \inf\{y : F(y) \geq u\}$ denote the left inverse, and assume that $0 < \mu < \infty$, where $\mu = E(Y)$. Then the LC (see e.g. Gastwirth [14]) is defined by

$$L(u) = L_F(u) = \mu^{-1} \int_0^u F^{-1}(s)ds, \quad 0 \leq u \leq 1. \tag{2.1}$$

Let $I_{\{A\}}$ denote the indicator of the event A. For F continuous we can write

$$L(u) = \mu^{-1} E\{Y I_{\{Y \leq F^{-1}(u)\}}\}. \tag{2.2}$$

When the population consists of incomes of people, the LC measures deviation from the *egalitarian* case $L(u) = u$ corresponding to where everyone has the same income $a > 0$ and the distribution of Y is degenerate at a. The other extreme occurs when one person has all the income which corresponds to $L(u) = 0,\ 0 \leq u \leq 1$. The intermediate case where Y is uniform on $[0, b], b > 0$, corresponds to $L(u) = u^2$. In general $L(u)$ is non-decreasing, convex, below the line $L(u) = u,\ 0 \leq u \leq 1$, and the greater the "distance" from u, the greater is the inequality in the population. If the population consists of companies providing a certain service or product, the LC

measures to what extent a few companies dominate the market with the extreme case corresponding to monopoly.

A closely related curve is the Bonferroni curve (BC) $B(u)$ which is defined (Aaberge [1], [2]), Giorgi and Mondani [15], Csörgö, Gastwirth and Zitikis [11]) as

$$B(u) = B_F(u) = u^{-1}L(u),\ 0 \le u \le 1. \tag{2.3}$$

When F is continuous the BC is the LC except that truncation is replaced by conditioning

$$B(u) = \mu^{-1}E\{Y|Y \le F^{-1}(u)\}. \tag{2.4}$$

The BC possesses several attractive properties. First, it provides a convenient alternative interpretation of the information content of the Lorenz curve. For a fixed u, $B(u)$ is the ratio of the mean income of the poorest $100 \times u$ percent of the population to the overall mean. Thus, the BC may also yield essential information on poverty provided that we know the poverty rate. Second, the BC of a uniform $(0,a)$ distribution proves to be the diagonal line joining the points (0,0) and (1,1) and thus represents a useful reference line, in addition to the two well-known standard reference lines. The egalitarian reference line coincides with the horizontal line joining the points (0,1) and (1,1). At the other extreme, when one person holds all income, the BC coincides with the horizontal axis except for $u = 1$.

In the next subsection we will consider ordering concepts from the statistics literature. Those concepts motivate the introduction of the following measures of concentration

$$C(u) = C_F(u) = \int_0^u \Big[\frac{F^{-1}(s)}{F^{-1}(u)}\Big]ds = \mu_F \frac{L_F(u)}{F^{-1}(u)},\ 0 < u < 1 \tag{2.5}$$

and

$$D(u) = D_F(u) = \frac{1}{u}\int_0^u \Big[\frac{F^{-1}(s)}{F^{-1}(u)}\Big]ds = \mu_F \frac{B_F(u)}{F^{-1}(u)},\ 0 < u < 1. \tag{2.6}$$

Accordingly, *D(u)* emerges by replacing the overall mean μ in the dominator of *B(u)* by the u^{th} quantile $y_u = F^{-1}(u)$ and is equal to the ratio between the mean income of those with lower income than the u^{th} quantile and the u-quantile income. Thus, *C(u)* and $D(u)$ measure inequality in income below the u^{th} quantile. They satisfy $C(u) \le u$, $D(u) \le 1$, $0 < u < 1$, and *C(u)* equals u and 0 while $D(u)$ equals 1 and 0 in the egalitarian and extreme non-egalitarian cases, respectively, and they equal $u/2$ and $1/2$ in the uniform case.

To summarize the information content of the inequality curves we recall the following inequality indices

$$G = 2\int_0^1 \{u - L(u)\}du \text{ (Gini)},\ B = \int_0^1 \{1 - B(u)\}du \text{ (Bonferroni)}, \tag{2.7}$$

$$C = 2\int_0^1 \{u - C(u)\}du,\ D = \int_0^1 \{1 - D(u)\}du. \tag{2.8}$$

These indices measure distances from the curves to their values in the egalitarian case, take values between 0 and 1 and are increasing with increasing inequality. If

all units have the same income then $G = B = C = D = 0$, and in the extreme non-egalitarian case where one unit has all the income and the others zero, $G = B = C = D = 1$. When F is uniform on $[0, b]$, $B = C = D = 1/2$ and $G = 1/3$. The inequality curves $L(u), B(u), C(u), D(u)$, and *the inequality measures* G, B, C and D are scale invariant; that is, they remain the same if Y is replaced by $aY, a > 0$.

2.2. Ordering inequality by transforming variables

When we are interested in how covariates influence inequality we may ask whether larger values of a covariate lead to more or less inequality. For instance, is there less inequality among the higher educated? To answer such questions we consider orderings of distributions on the basis of inequality, see e.g. Atkinson [5], Shorrocks and Foster [26], Dardanoni and Lambert [12], Muliere and Scarsini [20], Yitzhaki and Olkin [29], Zoli [30], and Aaberge [3]. In statistics and reliability engineering, orderings are plentiful, e.g. Lehmann [18], van Zwet [27], Barlow and Prochan [6], Birnbaum, Esary and Marshall [9], Doksum [13], Yanagimoto and Sibuya [28], Bickel and Lehmann [7], [8], Rojo and He [21], Rojo [22] and Shaked and Shanthikumar [25]. In statistics, similar orderings are often discussed in terms of spread or dispersion. Thus, for non-negative random variables, we could define Y to have a distribution which is more spread out to the right than that of Y_0 if Y can be written as $Y = h(Y_0)$ for some non-negative, nondecreasing convex function h (using van Zwet [27]). It turns out to be more general and more convenient to replace "convex" with "starshaped" (convex functions h are starshaped and concave functions g are anti-starshaped provided $g(0) = h(0) = 0$).

Recall that a nondecreasing function g defined on the interval $I \subset [0, \infty)$, is *starshaped* on I if $g(\lambda x) \leq \lambda g(x)$ whenever $x \in I, \lambda x \in I$ and $0 \leq \lambda \leq 1$. Thus if $I = (0, \infty)$, for any straight line through the origin, then the graph of g initially lies on or below it, and then lies on or above it. If $g(\lambda x) \geq \lambda g(x)$, g is *anti-starshaped*. On the class $\mathcal{F}$ of continuous and strictly increasing distributions F with $F(0) = 0$, Doksum [13] introduced the following partial ordering $F <_* H$ (F is *starshaped* with respect to H) if $H^{-1}F$ is starshaped on $\{x : 0 < F(x) < 1\}$. This ordering was also considered by Yanagimoto and Sibuya [28] and Bickel and Lehmann [7,8]. Note that if $F <_* H$ and X has distribution F, then $Z = H^{-1}[F(X)]$ has distribution H and is a starshaped transformation of X. Moreover,

Proposition 2.1. *Suppose that X and Z have distributions F and H, where $F, H \in \mathcal{F}$. Then $F <_* H$ if and only if there exists a positive nondecreasing function $q(\cdot)$ on $\{x : 0 < F(x) < 1\}$ such that Z has the same distribution as $q(X)X$.*

Proof. Suppose that $F <_* H$; then $q(x) = H^{-1}(F(x))/x$ will do because the starshaped condition $g(\lambda x) \leq \lambda x$ is equivalent to $[g(\lambda x)/\lambda x] \leq g(x)/x$. That is,

$$F <_* H \Leftrightarrow H^{-1}(F(x))/x \text{ is nondecreasing.} \tag{2.9}$$

Next, suppose that $q(\cdot)$ is positive and nondecreasing and that $Z = q(X)X$. Set $h(x) = q(x)x$. Then $h(x)$ is increasing and $P(X \leq x) = P(h(X) \leq h(x))$. It follows that $F(x) = H(q(x)x)$. That is, $q(x) = H^{-1}(F(x))/x$. □

Because of this proposition we say that if $F <_* H$ then F is a *more egalitarian* distribution of resources than H.

We next show that the preceding definition of inequality leads to the corresponding ordering of the inequality curves $C_F(\cdot)$ and $D_F(\cdot)$ as well as of the indices C and D.

Proposition 2.2. *Suppose that* $F, H \in \mathcal{F}$ *and* $F <_* H$. *Then* $C_F(u) \geq C_H(u)$ *and* $D_F(u) \geq D_H(u),\ 0 < u < 1$. *Therefore* $C_F \leq C_H$ *and* $D_F \leq D_H$.

Proof. It follows from (2.9), setting $u = F(x), v = F(x'),\ x < x'$, that $H^{-1}(u)/H^{-1}(v) \leq F^{-1}(u)/F^{-1}(v)$, for $0 < u < v < 1$. If we integrate this inequality over $u \in (0, v)$, we obtain $C_F(v) \geq C_H(v),\ 0 < v < 1$. The other inequalities follow from this. □

The same order preservation as stated in Proposition 2.2 holds for $L(u)$ and $B(u)$.

Theorem 2.1. *Suppose that* $F, H \in \mathcal{F}$ *and* $F <_* H$. *Then* $L_F(u) \geq L_H(u)$ *and* $B_F(u) \geq B_H(u),\ 0 < u < 1$. *Moreover,* $B_F \leq B_H$ *and* $G_F \leq G_H$.

Proof. Let

$$a = \int_0^1 H^{-1}(v)dv \Big/ \int_0^1 F^{-1}(v)dv, \tag{2.10}$$

and consider the line $y = ax$ through the origin. Then $H^{-1}(F(x))$ initially lies on or below this line, say up to the point $x = b$. Thus

$$\begin{aligned} \int_0^u H^{-1}(v)dv &= \int_0^{F^{-1}(u)} H^{-1}(F(x))dF(x) \leq \int_0^{F^{-1}(u)} axdF(x) \\ &= a \int_0^u F^{-1}(v)dv \end{aligned} \tag{2.11}$$

for all $u \leq F^{-1}(b)$ which establishes $L_F(u) \geq L_H(u)$ for $u \leq F(b)$. On the other hand, for $x \geq b$, $y = H^{-1}(F(x))$ is above $y = ax$. Thus, for $u > F(b)$

$$s(u) = \int_b^{F^{-1}(u)} \big[ax - H^{-1}(F(x))\big]dF(x)$$

is a negative, decreasing function of u. We can write, for $u > F(b)$,

$$\begin{aligned} a\int_0^u F^{-1}(v)dv - \int_0^u H^{-1}(v)dv &= a\int_0^{F(b)} F^{-1}(v)dv - \int_0^{F(b)} H^{-1}(v)dv + s(u) \\ &\equiv c + s(u) \end{aligned}$$

where c is nonnegative by (2.11). It follows that $c + s(u)$ is a decreasing function that equals 0 when $u = 1$ by the definition of a. Thus, $c + s(u) \geq 0$ which establishes $L_F(u) \geq L_H(u)$ again by the definition of a. The other inequalities follow from this. □

2.3. Ordering inequality by transforming distributions

A partial ordering on $\mathcal{F}$ based on transforming distributions rather than random variables is the following: F represents *more equality than* H ($F >_e H$) if

$$H(z) = g(F(z))$$

for some nonnegative increasing concave function g on $[0, 1]$ with $g(0) = 0$ and $g(1) = 1$. In other words, $F <_e H$ if $F(H^{-1})$ is convex. If F is uniform, the

orderings $F >_e H$ implies $F <_* H$ and in this case the results of Propositions 2.1, 2.2 and Theorem 2.1 hold. Note that when F and H have densities f and h, then $h(z) = g'(F(z))f(z)$ where $g'F$ is decreasing. That is, $F >_e H$ means that F has relatively more probability mass on the right than H.

A similar ordering involves $\bar{F}(x) = 1 - F(x)$ and $\bar{H}(z) = 1 - H(z)$. In this case we say that F represents a more equal distribution of resources than H ($F >_r H$) if

$$\bar{H}(x) = g(\bar{F}(x))$$

for some nonegative increasing convex transformation g on $[0, 1]$ with $g(0) = 0$ and $g(1) = 1$. In this case, if densities exist, they satisfy $h(z) = g'(\bar{F}(z))f(z)$, where $g'\bar{F}$ is decreasing. That is, relative to F, H has mass shifted to the left.

Remark. Orderings of inequality based on transforming distributions can be restated in terms of orderings based on transforming random variables. Thus $F >_e H$ is equivalent to the distribution function of $V = F(Z)$ being convex when $X \sim F$ and $Z \sim H$.

3. Regression inequality models

3.1. Notation and introduction

Next consider the case where the distribution of Y depends on covariates such as education, work experience, status of parents, sex, etc. Let $X_1, \ldots, X_d$ denote the covariates. We include an intercept term in the regression models, which makes it convenient to write $\boldsymbol{X} = (1, X_1, \ldots, X_d)^T$. Let $F(y|\boldsymbol{x})$ denote the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ and define the quantile regression function as the left inverse of this distribution function. The key quantity is

$$\mu(u|\boldsymbol{x}) \equiv \int_0^u F^{-1}(v|\boldsymbol{x})dv.$$

With this notation we can write the regression versions of the Lorenz curve, for $0 < u < 1$ as

$$L(u|\boldsymbol{x}) = \mu(u|\boldsymbol{x})/\mu(1|\boldsymbol{x}),\ B(u|\boldsymbol{x}) = L(u|\boldsymbol{x})/u.$$

Similarly, $C(u|\boldsymbol{x})$, $D(u|\boldsymbol{x})$ and the summary coefficients $G(\boldsymbol{x})$, $B(\boldsymbol{x})$, $C(\boldsymbol{x})$ and $D(\boldsymbol{x})$ are defined by replacing $F(y)$ by $F(y|\boldsymbol{x})$. Note that estimates of $F(y|\boldsymbol{x})$ and $\mu(y|\boldsymbol{x})$ provide estimates of the regression versions of the curves and measures of inequality. Thus, the rest of the paper discusses regression models for $F(y|\boldsymbol{x})$ and $\mu(y|\boldsymbol{x})$. Using the results of Section 2, these models are constructed so that the regression coefficients reflect relationships between covariates and measures of inequality.

3.2. Transformation regression models

Let Y_0 with distribution F_0 denote a baseline variable which corresponds to the case where the covariate vector $\boldsymbol{x}$ has no effect on the distribution of income. We assume that Y has a conditional distribution $F(y|\boldsymbol{x})$ which depends on $\boldsymbol{x}$ through some real valued function $\Delta(\boldsymbol{x}) = g(\boldsymbol{x}, \boldsymbol{\beta})$ which is known up to a vector $\boldsymbol{\beta}$ of unknown parameters. Let $Y \sim Z$ denote "Y is distributed as Z". As we have seen in

Section 2.2, if large values of $\Delta(\boldsymbol{x})$ correspond to a more egalitarian distribution of income than F_0, then it is reasonable to model this as

$$Y \sim h(Y_0),$$

for some increasing anti-starshaped function h depending on $\Delta(\boldsymbol{x})$. On the other hand, an increasing starshaped h would correspond to income being less egalitarian.

A convenient parametric form of h is

$$Y \sim \tau Y_0^{\Delta}, \tag{3.1}$$

where $\Delta = \Delta(\boldsymbol{x}) > 0$, and $\tau > 0$ does not depend on $\boldsymbol{x}$. Since $h(y) = y^{\Delta(\boldsymbol{x})}$ is concave for $0 < \Delta(\boldsymbol{x}) \leq 1$, while convex for $\Delta(\boldsymbol{x}) > 1$, the model (3.1) with $0 < \Delta(\boldsymbol{x}) \leq 1$ corresponds to covariates that lead to a less unequal distribution of income for Y than for Y_0, while $\Delta(\boldsymbol{x}) \geq 1$ is the opposite case. Thus it follows from the results of Section 2.2 that if we use the parametrization $\Delta(\boldsymbol{x}) = exp(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta})$, then the coefficient β_j in $\boldsymbol{\beta}$ measures how the covariate x_j relates to inequality in the distribution of resources Y.

Example 3.1. Suppose that $Y_0 \sim F_0$ where F_0 is the Pareto distribution $F_0(y) = 1 - (c/y)^a$, with $a > 1, c > 0, y \geq c$. Then $Y = \tau Y_0^{\Delta}$ has the Pareto distribution

$$F(y|\boldsymbol{x}) = F_0\big((\frac{y}{\tau})^{\frac{1}{\Delta}}\big) = 1 - \big(\frac{\lambda}{y}\big)^{\alpha(\boldsymbol{x})}, \;\; y \geq \lambda, \tag{3.2}$$

where $\lambda = c\tau$ and $\alpha(\boldsymbol{x}) = a/\Delta(\boldsymbol{x})$. In this case $\mu(u|\boldsymbol{x})$ and the regression summary measures have simple expressions, in particular

$$L(u|\mathbf{x}) = 1 - (1-u)^{1-\Delta(\mathbf{x})}.$$

When $\Delta(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$ then $\log Y$ already has a scale parameter and we set $\alpha = 1$ without loss of generality. One strategy for estimating $\boldsymbol{\beta}$ is to temporarily assume that λ is known and to use the maximum likelihood estimate $\hat{\boldsymbol{\beta}}(\lambda)$ based on the distribution of $\log Y_1, \ldots, \log Y_n$. Next, in the case where $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d., we can use $\hat{\lambda} = n \min\{Y_i\}/(n+1)$ to estimate λ. Because $\hat{\lambda}$ converges to λ at a faster than $\sqrt{n}$ rate, $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ is consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}}(\hat{\lambda}) - \boldsymbol{\beta})$ is asymptotically normal with the covariance matrix being the inverse of the λ-known information matrix.

Example 3.2. Another interesting case is obtained by setting F_0 equal to the log normal distribution $\Phi\big((\log(y) - \mu_0)/\sigma_0\big)$, $y > 0$. For the scaled log normal transformation model we get by straightforward calculation the following explicit form for the *conditional Lorenz curve*:

$$L(u|\boldsymbol{x}) = \Phi\big(\Phi^{-1}(u) - \sigma_0\Delta(\boldsymbol{x})\big). \tag{3.3}$$

In this case when we choose the parametrization $\Delta(\boldsymbol{x}) = \exp\big(x^T\boldsymbol{\beta}\big)$, the model already includes the scale parameter $\exp(-\beta_0)$ for $\log Y$. Thus we set $\mu_0 = 1$. To estimate $\boldsymbol{\beta}$ for this model we set $Z_i = \log Y_i$. Then Z_i has a $N\big(\alpha + \Delta(\boldsymbol{x}_i), \sigma_0^2\Delta^2(\boldsymbol{x}_i)\big)$ distribution, where $\alpha = \log\tau$ and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id})^T$. Because σ_0 and α are unknown there are $d+3$ parameters. When $Y_1, \ldots, Y_n$ are independent, this gives the log likelihood function (leaving out the constant term)

$$l(\alpha, \beta, \sigma_0^2) = -n\log(\sigma_0) - \sum_{i=1}^{n} \boldsymbol{x}_i^T\boldsymbol{\beta} - \frac{1}{2}\sigma_0^{-2}\sum_{i=1}^{n} \exp(-2\boldsymbol{x}_i^T\beta)\{Z_i - \alpha - \exp(\boldsymbol{x}_i^T\beta)\}^2$$

Likelihood methods will provide estimates, confidence intervals, tests and their properties. Software that only require the programming of the likelihood is available, e.g. Mathematica 5.2 and Stata 9.0.

4. Lehmann–Cox type semiparametric models. Partial likelihood

4.1. The distribution transformation model

Let $Y_0 \sim F_0$ be a baseline income distribution and let $Y \sim F(y|\boldsymbol{x})$ denote the distribution of income for given covariate vector $\boldsymbol{x}$. In Section 2.3 it was found that one way to express that $F(y|\boldsymbol{x})$ corresponds to more equality than $F_0(y)$ is to use the model

$$F(y|\boldsymbol{x}) = h(F_0(y))$$

for some nonnegative increasing concave transformation h depending on $\boldsymbol{x}$ with $h(0) = 0$ and $h(1) = 1$. Similarly, h convex corresponds to a more egalitarian income. A model of the form $F_2(y) = h(F_1(y))$ was considered for the two-sample case by Lehmann [17] who noted that $F_2(y) = F_1^{\Delta}(y)$ for $\Delta > 0$ was a convenient choice of h. For regression experiments, we consider a regression version of this Lehmann model which we define as

$$F(y|\boldsymbol{x}) = F_0^{\Delta}(y) \tag{4.1}$$

where $\Delta = \Delta(\boldsymbol{x}) = g(\boldsymbol{x},\boldsymbol{\beta})$ is a real valued parametric function and where $\Delta < 1$ or $\Delta > 1$ corresponds to $F(y|\boldsymbol{x})$ representing a more or less egalitarian distribution of resources than $F_0(y)$, respectively.

To find estimates of $\boldsymbol{\beta}$, note that if we set $U_i = 1 - F_0(Y_i)$, then U_i has the distribution

$$H(u) = 1 - (1-u)^{\Delta(\boldsymbol{x})},\ 0 < u < 1$$

which is the distribution of $F_0(Y_i)$ in the next subsection. Since the rank R_i of Y_i equals $N + 1 - S_i$, where S_i is the rank of $1 - F_0(Y_i)$, we can use rank methods, or Cox partial likelihood methods, to estimate $\boldsymbol{\beta}$ without knowing F_0. In fact, because the Cox partial likelihood is a rank likelihood and $rank[1 - F_0(Y_i)]$=$rank(-Y_i)$, we can apply the likelihood in the next subsection to estimate the parameters in the current model provided we reverse the ordering of the Y's.

4.2. The semiparametric generalized Pareto model

In this section we show how the Pareto parametric regression model for income can be extended to a semiparametric model where the shape of the income distribution is completely general. This model coincides with the Cox proportional hazard model for which a wealth of theory and methods are available.

We defined a regression version of the Pareto model in Example 3.1 as

$$F(y|\boldsymbol{x}) = 1 - \left(\tfrac{c}{y}\right)^{\alpha_i},\ \ y \geq c; \alpha_i > 0,$$

where $\alpha_i = \Delta_i^{-1}, \Delta_i = \exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}$. This model satisfies

$$1 - F(y|\boldsymbol{x}) = (1 - F_0(y))^{\alpha_i}, \tag{4.2}$$

where $F_0(y) = 1 - c/y, y \geq c$. When F_0 is an arbitrary continuous distribution on $[0, \infty)$, the model (4.2) for the two sample case was called the Lehmann alternative by Savage [23], [24] because if V satisfies model (4.1), then $Y = -V$ satisfies model (4.2). Cox [10] introduced proportional hazard models for regression experiments in survival analysis which also satisfy (4.2) and introduced partial likelihood methods that can be used to analyse such models even in the presence of censoring and time dependent covariates (in our case, wage dependent covariates).

Cox introduced the model equivalent to (4.2) as a generalization of the exponential model where $F_0(y) = 1 - \exp(-y)$ and $F(y|\boldsymbol{x_i})=F_0(\Delta_i^{-1}y)$. That is, (4.2) is in the Cox case a semiparametric generalization of a scale model with scale parameter Δ_i. However, in our case we regard (4.2) as a semiparametric shape model which generalizes the Pareto model, and Δ_i represents the degree of inequality for a given covariate vector $\boldsymbol{x}_i$. The inequality measures correct for this confounding of shape and scale by being scale invariant.

Note from Section 2.3 that $\Delta_i < 1$ corresponds to $F(y|\boldsymbol{x})$ more egalitarian than $F_0(y)$ while $\Delta_i > 1$ corresponds to F_0 more egalitarian.

The Cox [10] partial likelihood to estimate $\boldsymbol{\beta}$ for (4.2) is (see also Kalbfleisch and Prentice [16], page 102),

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big\{ \exp(-\boldsymbol{x}_{(i)}^T\boldsymbol{\beta}) \sum_{k \in R(Y_{(i)})} \exp(-\boldsymbol{x}_{(k)}^T\boldsymbol{\beta}) \Big\}$$

where $Y_{(i)}$ is the i-th order statistic, $\boldsymbol{x}_{(i)}$ is the covariate vector for the subject with response $Y_{(i)}$, and $R(Y_{(i)}) = \{k : Y_{(k)} \geq Y_{(i)}\}$. Here $\hat{\boldsymbol{\beta}}$=arg max $L(\boldsymbol{\beta})$ can be found in many statistical packages, such as S-Plus, SAS, and STATA 9.0. These packages also give the standard errors of the $\hat{\beta}_j$. Note that $L(\boldsymbol{\beta})$ does not involve F_0.

Many estimates are available for F_0 in model (4.2) in the same packages. If we maximize the likelihood keeping $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ fixed, we find (e.g., Kalbfleisch and Prentice [16], p. 116, Andersen et al. [4], p. 483) $\hat{F}_0(Y_{(i)}) = 1 - \prod_{j=1}^{n} \hat{\alpha}_j$, where $\hat{\alpha}_j$ is the Breslow-Nelson-Aalen estimate,

$$\hat{\alpha}_j = \left\{ 1 - \frac{\exp(-\boldsymbol{x}_{(i)}^T\hat{\beta})}{\sum_{k \in R(Y_{(i)})} \exp(-\boldsymbol{x}_{(i)}^T\hat{\beta})} \right\}^{\exp(\boldsymbol{x}_{(i)}^T\beta)}$$

Andersen et al. [4] among others give the asymptotic properties of $\hat{F}_0$.

We can now give theoretical and empirical expressions for the conditional *inequality curves* and measures. Using (4.2), we find

$$F^{-1}(u|\boldsymbol{x}_i) = F_0^{-1}(1 - (1-u)^{\Delta_i}) \tag{4.3}$$

and

$$\mu(u|\boldsymbol{x}_i) = \int_0^u F^{-1}(t|\boldsymbol{x}_i)dt = \int_0^u F_0^{-1}(1 - (1-v)^{\Delta_i})dv. \tag{4.4}$$

We set $t = F_0^{-1}(1 - (1-v)^{\Delta_i})$ and obtain

$$\mu(u|\boldsymbol{x}_i) = \Delta_i^{-1} \int_0^{\delta(u)} t(1 - F_0(t))^{\Delta_i^{-1}-1} dF_0(t),$$

where $\delta_i(u) = F_0^{-1}(1-(1-u)^{\Delta_i})$. To estimate $\mu(u|\boldsymbol{x_i})$, we let

$$b_i = \hat{F}_0(Y_{(i)}) - \hat{F}_0(Y_{(i-1)}) = \prod_{j=1}^{i-1} \hat{\alpha}_j = (1-\hat{\alpha}_i) \prod_{j=1}^{i-1} \hat{\alpha}_j$$

be the jumps of $\hat{F}_0(\cdot)$; then

$$\hat{\mu}(u|\boldsymbol{x}_i) = \hat{\Delta}_i^{-1} \sum_j b_j Y_{(j)} (1-\hat{F}_0(Y_{(j)}))^{\hat{\Delta}_i^{-1}-1}$$

where the sum is over j with $\hat{F}_0(Y_{(j)}) \leq 1-(1-u)^{\hat{\Delta}_i}$. Finally,

$$\hat{L}(u|\boldsymbol{x}) = \hat{\mu}(u|\boldsymbol{x})/\hat{\mu}(1|\boldsymbol{x}), \hat{B}(u|\boldsymbol{x}) = \hat{L}(u|\boldsymbol{x})/u,$$

and

$$\hat{C}(u|\boldsymbol{x}) = \hat{\mu}(u|\boldsymbol{x})/\hat{F}^{-1}(u|\boldsymbol{x}), \hat{D}(u|\boldsymbol{x}) = \hat{C}(u|\boldsymbol{x})/u,$$

where $\hat{F}^{-1}(u|\boldsymbol{x})$ is the estimate of the conditional quantile function obtained from (4.3) by replacing Δ_i with $\hat{\Delta}_i$ and F_0 with $\hat{F}_0$.

Remark. The methods outlined here for the Cox proportional hazard model have been extended to the case of ties among the responses Y_i, to censored data, and to time dependent covariates (see e.g. Cox [10], Andersen et al. [4] and Kalbfleisch and Prentice [16]). These extensions can be used in the analysis of the semiparametric generalized Pareto model with tied wages, censored wages, and dependent covariates.

References

[1] Aaberge, R. (1982). On the problem of measuring inequality (in Norwegian). Rapporter 82/9, Statistics Norway.

[2] Aaberge, R. (2000a). Characterizations of Lorenz curves and income distributions. *Social Choice and Welfare* **17**, 639–653.

[3] Aaberge, R. (2000b). Ranking intersecting Lorenz Curves. Discussion Paper No. 412, Statistics Norway.

[4] Andersen, P. K., Borgan, Ø. Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[5] Atkinson, A. B. (1970). On the measurement of inequality, *J. Econ. Theory* **2**, 244–263.

[6] Barlow, R. E. and Proschan, F. (1965). *Mathematical Theory of Reliability*. Wiley, New York.

[7] Bickel, P. J. and Lehmann, E. L. (1976). Descriptive statistics for nonparametric models. III. Dispersion. *Ann. Statist.* **4**, 1139–1158.

[8] Bickel, P. J. and Lehmann, E. L. (1979). Descriptive measures for nonparametric models IV, Spread. In *Contributions to Statistics, Hajek Memorial Volume*, J. Juneckova (ed.). Reidel, London, 33–40.

[9] Birnbaum, S. W., Esary, J. D. and Marshall, A. W. (1966). A stochastic characterization of wear-out for components and systems. *Ann. Math. Statist.* **37**, 816–826.

[10] Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Stat. Soc. B* **34**, 187–220.

[11] Csörgö, M., Gastwirth, J. L. and Zitikis, R. (1998). Asymptotic confidence bands for the Lorenz and Bonferroni curves based on the empirical Lorenz curve. *Journal of Statistical Planning and Inference* **74**, 65–91.

[12] Dardanoni, V. and Lambert, P. J. (1988). Welfare rankings of income distributions: A role for the variance and some insights for tax reforms. *Soc. Choice Welfare* **5**, 1–17.

[13] Doksum, K. A. (1969). Starshaped transformations and the power of rank tests. *Ann. Math. Statist.* **40**, 1167–1176.

[14] Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica* **39**, 1037–1039.

[15] Giorgi, G. M. and Mondani, R. (1995). Sampling distribution of the Bonferroni inequality index from exponential population. *Sankyā* **57**, 10–18.

[16] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Wiley, New York.

[17] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24**, 23–43.

[18] Lehmann, E. L. (1955). Ordered families of distributions. *Ann. Math. Statist.* **37**, 1137–1153.

[19] Lorenz, M. C. (1905). Methods of measuring the concentration of wealth. *J. Amer. Statist.* **9**, 209–219.

[20] Muliere, P. and Scarsini, M. (1989). A Note on Stochastic Dominance and Inequality Measures. *Journal of Economic Theory* **49**, 314–323.

[21] Rojo, J. and He, G. Z. (1991). New properties and characterizations of the dispersive orderings. *Statistics and Probability Letters* **11**, 365–372.

[22] Rojo, J. (1992). A pure-tail ordering based on the ratio of the quantile functions. *Ann. Statist.* **20**, 570–579.

[23] Savage, I. R. (1956). Contributions to the theory of rank order statistics – the two-sample case. *Ann. Math. Statist.* **27**, 590–615.

[24] Savage, I. R. (1980). *Lehmann Alternatives.* Colloquia Mathematica Societatis János Bolyai, Nonparametric Statistical Inference, Proceedings, Budapest, Hungary.

[25] Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications.* Academic Press, San Diego.

[26] Shorrocks, A. F. and Foster, J. E. (1987). Transfer sensitive inequality measures. *Rev. Econ. Stud.* **14**, 485–497.

[27] van Zwet, W. R. (1964). *Convex Transformations of Random Variables.* Math. Centre, Amsterdam.

[28] Yanagimoto, T. and Sibuya, M. (1976). Isotonic tests for spread and tail. *Annals of Statist. Math.* **28**, 329–342.

[29] Yitzhaki, S. and Olkin, I. (1991). Concentration indices and concentration curves. Stochastic Order and Decision under Risk. *IMS Lecture Notes–Monograph Series.*

[30] Zoli, C. (1999). Intersecting generalized Lorenz curves and the Gini index. *Soc. Choice Welfare* **16**, 183–196.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 131–169
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000437

Estimation in a class of semiparametric transformation models

Dorota M. Dabrowska[1,*]

University of California, Los Angeles

Abstract: We consider estimation in a class of semiparametric transformation models for right–censored data. These models gained much attention in survival analysis; however, most authors consider only regression models derived from frailty distributions whose hazards are decreasing. This paper considers estimation in a more flexible class of models and proposes conditional rank M-estimators for estimation of the Euclidean component of the model.

1. Introduction

Semiparametric transformation models provide a common tool for regression analysis. We consider estimation in a class of such models designed for analysis of failure time data with time independent covariates. Let μ be the marginal distribution of a covariate vector Z and let $H(t|z)$ be the cumulative hazard function of the conditional distribution of failure time T given Z. We assume that for μ–almost all z (μ a.e. z) this function is of the form

$$H(t|z) = A(\Gamma(t), \theta|z) \tag{1.1}$$

where Γ is an unknown continuous increasing function mapping the support of the failure time T onto the positive half-line. For μ a.e. z, $A(x, \theta|z)$ is a conditional cumulative hazard function dependent on a Euclidean parameter θ and having hazard rate $\alpha(x, \theta|z)$ strictly positive at $x = 0$ and supported on the whole positive half-line. Special cases include

(i) the proportional hazards model with constant hazard rate $\alpha(x, \theta|z) = \exp(\theta^T z)$ (Lehmann [23], Cox [12]);
(ii) transformations to distributions with monotone hazards such as the proportional odds and frailty models or linear hazard rate regression model (Bennett [2], Nielsen *et al.* [28], Kosorok *et al.* [22], Bogdanovicius and Nikulin [9]);
(iii) scale regression models induced by half-symmetric distributions (section 3).

The proportional hazards model remains the most commonly used transformation model in survival analysis. Transformation to exponential distribution entails that for any two covariate levels z_1 and z_2, the ratio of hazards is constant in x and equal to $\alpha(x, \theta|z_1)/\alpha(x, \theta|z_2) = \exp(\theta^T[z_1 - z_2])$. Invariance of the model with respect to monotone transformations enstails that this constancy of hazard ratios is preserved by the transformation model. However, in many practical circumstances

*Research supported in part by NSF grant DMS 9972525 and NCI grant 2R01 95 CA 65595-01.
[1]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, e-mail: dorota@ucla.edu
AMS 2000 subject classifications: primary 62G08; secondary 62G20.
Keywords and phrases: transformation models, M-estimation, Fredholm and Volterra equations.

this may fail to hold. For example, a new treatment ($z_1 = 1$) may be initially beneficial as compared to a standard treatment ($z_2 = 0$), but the effects may decay over time, $\alpha(x, \theta|z_1 = 1)/\alpha(x, \theta|z_2 = 0) \downarrow 1$ as $x \uparrow \infty$. In such cases the choice of the proportional odds model or a transformation model derived from frailty distributions may be more appropriate. On the other hand, transformation to distributions with increasing or non-monotone hazards allows for modeling treatment effects which have divergent long-term effects or crossing hazards. Transformation models have also found application in regression analyses of multivariate failure time data, where models are often defined by means of copula functions and marginals are specified using models (1.1).

We consider parameter estimation in the presence of right censoring. In the case of uncensored data, the model is invariant with respect to the group of increasing transformations mapping the positive half-line onto itself so that estimates of the parameter θ are often sought within the class of conditional rank statistics. Except for the proportional hazards model, the conditional rank likelihood does not have a simple tractable form and estimation of the parameter θ requires joint estimation of the pair (θ, Γ). An extensive study of this estimation problem was given by Bickel [4], Klaassen [21] and Bickel and Ritov [5]. In particular, Bickel [4] considered the two sample testing problem, $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H} : \theta > \theta_0$, in one-parameter transformation models. He used projection methods to show that a nonlinear rank statistic provides an efficient test, and applied Sturm-Liouville theory to obtain the form of its score function. Bickel and Ritov [5] and Klaassen [21] extended this result to show that under regularity conditions, the rank likelihood in regression transformation models forms a locally asymptoticaly normal family and estimation of the parameter θ can be based on a one-step MLE procedure, once a preliminary $\sqrt{n}$ consistent estimate of θ is given. Examples of such estimators, specialized to linear transformation models, can be found in [6, 13, 15], among others.

In the case of censored data, the estimation problem is not as well understood. Because of the popularity of the proportional hazards model, the most commonly studied choice of (1.1) corresponds to transformation models derived from frailty distributions. Murphy *et al.* [27] and Scharfstein *et al.* [31] proposed a profile likelihood method of analysis for the generalized proportional odds ratio models. The approach taken was similar to the classical proportional hazards model. The model (1.1) was extended to include all monotone functions Γ. With fixed parameter θ, an approximate likelihood function for the pair (θ, Γ) was maximized with respect to Γ to obtain an estimate $\Gamma_{n\theta}$ of the unknown transformation. The estimate $\Gamma_{n\theta}$ was shown to be a step function placing mass at each uncensored observation, and the parameter θ was estimated by maximizing the resulting profile likelihood. Under certain regularity conditions on the censoring distribution, the authors showed that the estimates are consistent, asymptotically Gaussian at rate $\sqrt{n}$, and asymptotically efficient for estimation of both components of the model. The profile likelihood method discussed in these papers originates from the counting process proportional hazards frailty intensity models of Nielsen et al. [28]. Murphy [26] and Parner [30] developed properties of the profile likelihood method in multi-jump counting process models. Kosorok *et al* [22] extended the results to one-jump frailty intensity models with time dependent covariates, including the gamma, the lognormal and the generalized inverse Gaussian frailty intensity models. Slud and Vonta [33] provided a separate study of consistency properties of the nonparametric maximum profile likelihood estimator in transformation models assuming that the cumulative hazard function (1.1) is of the form $H(t|z) = A(\exp[\theta^T z]\Gamma(t))$ where A is a known concave function.

Several authors proposed also ad hoc estimates of good practical performance. In particular, Cheng *et al.* [11] considered estimation in the linear transformation model in the presence of censoring independent of covariates. They showed that estimation of the parameter θ can be accomplished without estimation of the transformation function by means of U-statistics estimating equations. The approach requires estimation of the unknown censoring distribution, and does not extend easily to models with censoring dependent on covariates. Further, Yang and Prentice [34] proposed minimum distance estimation in the proportional odds ratio model and showed that the unknown odds ratio function can be estimated based on a sample analogue of a linear Volterra equation. Bogdanovicius *et al.* [9, 10] considered estimation in a class of generalized proportional hazards intensity models that includes the transformation model (1.1) as a special case and proposed a modified partial likelihood for estimation of the parameter θ. As opposed to the profile likelihood method, the unknown transformation was profiled out from the likelihood using a martingale-based estimate of the unknown transformation obtained by solving recurrently a Volterra equation.

In this paper we consider an extension of estimators studied by Cuzick [13] and Bogdanovicius *et al.* [9, 10] to a class of M-estimators of the parameter θ. In Section 2 we shall apply a general method for construction of M-estimates in semiparametric models outlined in Chapter 7 of Bickel *et al.* [6]. In particular, the approach requires that the nuisance parameter and a consistent estimate of it be defined in a larger model $\mathcal{P}$ than the stipulated semiparametric model. Denoting by (X, δ, Z), the triple corresponding to a nonnegative time variable X, a binary indicator δ and a covariate Z, in this paper we take $\mathcal{P}$ as the class of all probability measures such that the covariate Z is bounded and the marginal distribution of the withdrawal times is either continuous or has a finite number of atoms. Under some regularity conditions on the core model $\{A(x, \theta|z) : \theta \in \Theta, x > 0\}$, we define a parameter $\Gamma_{P,\theta}$ as a mapping of $\mathcal{P}\times\Theta$ into a convex set of monotone functions. The parameter represents a transformation function that is defined as a solution to a nonlinear Volterra equation. We show that its "plug-in" estimate $\Gamma_{P_n,\theta}$ is consistent and asymptotically linear at rate $\sqrt{n}$. Here P_n is the empirical measure of the data corresponding to an iid sample of the (X, δ, Z) observations. Further, we propose a class of M-estimators for the parameter θ. The estimate will be obtained by solving a score equation $U_n(\theta) = 0$ or $U_n(\theta) = o_P(n^{-1/2})$ for θ. Similarly to the case of the estimator $\Gamma_{n\theta}$, the score function $U_n(\theta)$ is well defined (as a statistic) for any $P \in \mathcal{P}$. It forms, however, an approximate V-process so that its asymptotic properties cannot be determined unless the "true" distribution $P \in \mathcal{P}$ is defined in sufficient detail (Serfling [32]). The properties of the score process will be developed under the added assumption that at true $P \in \mathcal{P}$, the observation $(X, \delta, Z) \sim \mathcal{P}$ has the same distribution as $(T \wedge \tilde{T}, 1(T \leq \tilde{T}), Z)$, where T and $\tilde{T}$ represent failure and censoring times conditionally independent given the covariate Z, and the conditional distribution of the failure time T given Z follows the transformation model (1.1).

Under some regularity conditions, we show that the M-estimates converge at rate $\sqrt{n}$ to a normal limit with a simple variance function. By solving a Fredholm equation of second kind, we also show that with an appropriate choice of the score process, the proposed class of censored data rank statistics includes estimators of the parameter θ whose asymptotic variance is equal to the inverse of the asymptotic variance of the M-estimating score function $\sqrt{n}U_n(\theta_0)$. We give a derivation of the resolvent and solution of the equation based on Fredholm determinant formula. We also show that this is a Sturm-Liouville equation, though of a different form than in [4, 5] and [21].

The class of transformation models considered in this paper is different than in the literature on nonparametric maximum likelihood estimation (NMPLE); in particular, hazard rates of core models need not be decreasing. In section 2, the core models are assumed to have hazards $\alpha(x, \theta, z)$ uniformly bounded between finite positive constants. With this aid we show that the mapping $\Gamma_{P,\theta}$ of $\mathcal{P}\times\Theta$ into the class of monotone functions is well defined on the entire support of the withdrawal time distribution, and without any special conditions on the probability distribution P. Under the assumption that the upper support point τ_0 of the withdrawal time distribution is a discontinuity point, the function $\Gamma_{P,\theta}$ is shown to be bounded. If τ_0 is a continuity point of this distribution, the function $\Gamma_{P,\theta}(t)$ is shown to grow to infinity as $t \uparrow \tau_0$. In the absence of censoring, the model (1.1) assumes that the unknown transformation is an unbounded function, so we require $\Gamma_{P,\theta}$ to have this property as well. In section 3, we use invariance properties of the model to show that the results can also be applied to hazards $\alpha(x, \theta, z)$ which are positive at the origin, but only locally bounded and locally bounded away from 0. All examples in this section refer to models whose conditional hazards are hyperbolic, i.e can be bounded (in a neighbourhood of the true parameter) between a linear function $a+bx$ and a hyperbola $(c+dx)^{-1}$, for some $a > 0, c > 0$ and $b \geq 0, d \geq 0$. As an example, we discuss the linear hazard rate transformation model, whose conditional hazard function is increasing, but its conditional density is decreasing or non-monotone, and the gamma frailty model with fixed frailty parameter or frailty parameters dependent on covariates.

We also examine in some detail scale regression models whose core models have cumulative hazards of the form $A_0(x \exp[\beta^T z])$. Here A_0 is a known cumulative hazard function of a half-symmetric distribution with density α_0. Our results apply to such models if for some fixed $\xi \in [-1, 1]$ and $\eta \geq 0$, the ratio α_0/g, $g(x) = [1+\eta x]^{\xi}$ is a function locally bounded and locally bounded away from zero. We show that this choice includes half-logistic, half-normal and half-t scale regression models, whose conditional hazards are increasing or non-monotone while densities are decreasing. We also give examples of models (with coefficient $\xi \notin [-1, 1]$) to which the results derived here cannot be applied.

Finally, this paper considers only the gamma frailty model with the frailty parameter fixed or dependent on covariates. We show, however, that in the case that the known transformation is the identity map, the gamma frailty regression model (frailty parameter independent of covariates) is not regular in its entire parameter range. When the transformation is unknown, and the parameter set restricted to $\eta \geq 0$, we show that the frailty parameter controls the shape of the transformation. We do not know at the present time, if there exists a class of conditional rank statistics which allows to estimate the parameter η, without any additional regularity conditions on the unknown transformation.

In Section 4 we summarize the findings of this paper and outline some open problems. The proofs are given in the remaining 5 sections.

2. Main results

We shall first give regularity conditions on the model (Section 2.1). The asymptotic properties of the estimate of the unknown transformation are discussed in Section 2.2. Section 2.3 introduces some additional notation. Section 2.4 considers estimation of the Euclidean component of the model and gives examples of M-estimators of this parameter.

2.1. The model

Throughout the paper we assume that (X, δ, Z) is defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and represents a nonnegative withdrawal time (X), a binary indicator (δ) and a vector of covariates (Z). Set $N(t) = 1(X \le t, \delta = 1)$, $Y(t) = 1(X \ge t)$ and let $\tau_0 = \tau_0(P) = \sup\{t : E_P Y(t) > 0\}$. We shall make the following assumption about the "true" probability distribution P.

Condition 2.0. $P \in \mathcal{P}$ where $\mathcal{P}$ is the class of all probability distributions such that

(i) The covariate Z has a nondegenerate marginal distribution μ and is bounded: $\mu(|Z| \le C) = 1$ for some constant C.
(ii) The function $E_P Y(t)$ has at most a finite number of discontinuity points, and $E_P N(t)$ is either continuous or discrete.
(iii) The point $\tau > 0$ satisfies $\inf\{t : E_P[N(t)|Z = z] > 0\} < \tau$ for μ a.e. z. In addition, $\tau = \tau_0$, if τ_0 is an discontinuity point of $E_P Y(t)$, and $\tau < \tau_0$, if τ_0 is a continuity point of $E_P Y(t)$.

For given τ satisfying Condition 2.0(iii), we denote by $\|\cdot\|_\infty$ the supremum norm in $\ell^\infty([0, \tau])$. The second set of conditions refers to the core model $\{A(\cdot, \theta|z) : \theta \in \Theta\}$.

Condition 2.1. (i) The parameter set $\Theta \subset R^d$ is open, and θ is identifiable in the core model: $\theta \ne \theta'$ iff $A(\cdot, \theta|z) \not\equiv A(\cdot, \theta'|z)$ μ a.e. z.
(ii) For μ almost all z, the function $A(\cdot, \theta|z)$ has a hazard rate $\alpha(\cdot, \theta|z)$. There exist constants $0 < m_1 < m_2 < \infty$ such that $m_1 \le \alpha(x, \theta|z) \le m_2$ for μ a.e. z and all $\theta \in \Theta$.
(iii) The function $\ell(x, \theta, z) = \log \alpha(x, \theta, z)$ is twice continuously differentiable with respect to both x and θ. The derivatives with respect to x (denoted by primes) and with respect to θ (denoted by dots) satisfy

$$\begin{aligned} |\ell'(x, \theta, z)| &\le \psi(x), \qquad |\ell''(x, \theta, z)| \le \psi(x), \\ |\dot\ell(x, \theta, z)| &\le \psi_1(x), \qquad |\ddot\ell(x, \theta, z)| \le \psi_2(x), \\ |g(x, \theta, z) - g(x', \theta', z)| &\le \max(\psi_3(x), \psi_3(x'))[|x - x'| + |\theta - \theta'|], \end{aligned}$$

where $g = \ddot\ell, \dot\ell'$ and ℓ''. Here ψ is a constant or a continuous bounded decreasing function. The functions $\dot\ell, \ddot\ell$ and $\dot\ell'$ are locally bounded and $\psi_p, p = 1, 2, 3$ are continuous, bounded or strictly increasing and such that $\psi_p(0) < \infty$,

$$\int_0^\infty e^{-x}\psi_1^2(x)dx < \infty, \quad \int_0^\infty e^{-x}\psi_p(x)dx < \infty, \quad p = 2, 3.$$

To evaluate the score process for estimation of the parameter θ, we shall use the following added assumption.

Condition 2.2. The true distribution $P, P \in \mathcal{P}$, is the same as that of $(X, \delta, Z) \sim (T \wedge \tilde T, 1(T \le \tilde T), Z)$, where T and $\tilde T$ represent failure and censoring times. The variables T and $\tilde T$ are conditionally independent given Z. In addition

(i) The conditional cumulative hazard function of T given Z is of the form $H(t|z) = A(\Gamma_0(t), \theta_0|z)$ μ a.e. z, where Γ_0 is a continuous increasing function, and $A(x, \theta_0|z) = \int_0^x \alpha(u, \theta_0|z)du, \theta_0 \in \Theta$, is a cumulative hazard function with hazard rate $\alpha(u, \theta_0|z)$ satisfying Conditions 2.1.

(ii) If τ_0 is a discontinuity point of the survival function $E_P Y(t)$, then $\tau_0 = \sup\{t : P(\tilde{T} \geq t) > 0\} < \sup\{t : P(T \geq t) > 0\}$. If τ_0 is a continuity point of this survival function, then $\tau_0 = \sup\{t : P(T \geq t) > 0\} \leq \sup\{t : P(\tilde{T} \geq t) > 0\}$.

For $P \in \mathcal{P}$, let $A(t) = A_P(t)$ be given by

$$A(t) = \int_0^t \frac{EN_P(du)}{E_P Y(u)}. \tag{2.1}$$

If the censoring time $\tilde{T}$ is independent of covariates, then $A(t)$ reduces to the marginal cumulative hazard function of the failure time T, restricted to the interval $[0, \tau_0]$. Under Assumption 2.2 this parameter forms in general a function of the marginal distribution of covariates, and conditional distributions of both failure and censoring times. Nevertheless, we shall find it, and the associated Aalen–Nelson estimator, quite useful in the sequel. In particular, under Assumption 2.2, the conditional cumulative hazard function $H(t|z)$ of T given Z is uniformly dominated by $A(t)$. We have

$$A(t) = \int_0^t E[\alpha(\Gamma_0(u-), \theta_0, Z)|X \geq u]\Gamma_0(du)$$

and

$$\frac{H(dt|z)}{A(dt)} = \frac{\alpha(\Gamma_0(t-), \theta_0, z)}{E\alpha(\Gamma_0(t-), \theta_0, Z)|X \geq t)},$$

for $t \leq \tau(z) = \sup\{t : EY(t)|Z = z > 0\}$ and μ a.e. z. These identities suggest to define a parameter $\Gamma_{P,\theta}$ as solution to the nonlinear Volterra equation

$$\begin{aligned} \Gamma_{P,\theta}(t) &= \int_0^t \frac{E_P N(du)}{E_P Y(u)\alpha(\Gamma_\theta(u-), \theta, Z)} \\ &= \int_0^t \frac{A_P(du)}{E_P \alpha(\Gamma_\theta(u-), \theta, Z)|X \geq u)}, \end{aligned} \tag{2.2}$$

with boundary condition $\Gamma_{P,\theta}(0-) = 0$. Because Conditions 2.2 are not needed to solve this equation, we shall view Γ as a map of the set $\mathcal{P} \times \Theta$ into $\mathcal{X} = \cup\{\mathcal{X}(P) : P \in \mathcal{P}\}$, where

$$\mathcal{X}(P) = \{g : g \text{ increasing}, e^{-g} \in D(\mathcal{T}), g \ll E_P N, m_2^{-1} A_P \leq g \leq m_1^{-1} A_P\}$$

and m_1, m_2 are constants of Condition 2.1(iii). Here $D(\mathcal{T})$ denotes the space of right-continuous functions with left-hand limits, and we choose $\mathcal{T} = [0, \tau_0]$, if τ_0 is a discontinuity point of the survival function $E_P Y(t)$, and $\mathcal{T} = [0, \tau_0)$, if it is a continuity point. The assumption $g \ll E_P N$ means that the functions g in $\mathcal{X}(P)$ are absolutely continuous with respect to the sub-distribution function $E_P N(t)$. The monotonicity condition implies that they admit integral representation $g(t) = \int_0^t h(u) dE_P N(u)$ and $h \geq 0$, $E_P N$-almost everywhere.

2.2. Estimation of the transformation

Let (N_i, Y_i, Z_i), $i = 1, \ldots, n$ be an iid sample of the (N, Y, Z) processes. Set $S(x, \theta, t) = n^{-1} \sum_{i=1}^n Y_i(t)\alpha(x, \theta, Z_i)$ and denote by $\dot{S}, S'$ the derivatives of these

processes with respect to θ (dots) and x (primes) and let $s, \dot{s}, s'$ be the corresponding expectations. Suppressing dependence of the parameter $\Gamma_{P,\theta}$ on P, set

$$C_\theta(t) = \int_0^t \frac{EN(du)}{s^2(\Gamma_\theta(u-), \theta, u)}.$$

For $u \le t$, define also

$$\begin{aligned} \mathcal{P}_\theta(u,t) &= \prod\nolimits_{(u,t]}(1 - s'(\Gamma_\theta(w-), \theta, w)C_\theta(dw)), \\ &= \exp[-\int_u^t s'(\Gamma_\theta(w-), \theta, w)C_\theta(dw)] \quad \text{if } EN(t) \text{ is continuous}, \\ &= \prod_{u<w\le t} [1 - s'(\Gamma_\theta(w-), \theta, w)C_\theta(dw)] \quad \text{if } EN(t) \text{ is discrete}. \end{aligned} \tag{2.3}$$

Finally, we follow Bogdanovicius and Nikulin [9], and use

$$\Gamma_{n\theta}(t) = \int_0^t \frac{N_\cdot(du)}{S(\Gamma_{n\theta}(u-), \theta, u)}, \quad \Gamma_{n\theta}(0-) = 0, \theta \in \Theta,$$

to estimate the unknown transformation. Here $N_\cdot = n^{-1}\Sigma_{i=1}^n N_i$.

Proposition 2.1. *Let $P \in \mathcal{P}$ be a distribution satisfying Conditions* 2.0(i), (ii) *and let (X_i, δ_i, Z_i), $i = 1, \dots, n$ be an iid sample from this distribution. Suppose that Conditions* 2.1 *are fulfilled by the family $\{A(\cdot, \theta|z) : \theta \in \Theta\}$, and let τ be an arbitrary point such that Condition* 2.0(iii) *holds.*

(i) *Equation* (2.2) *has a unique locally bounded solution satisfying $0 < \Gamma_\theta(\tau_0) < \infty$ if $\tau_0 = \tau_0(P)$ is a discontinuity point of $EY_P(t)$ and $\lim_{t\uparrow\infty}\Gamma_\theta(t) = \infty$ if τ_0 is a continuity point of this survival function. For any point τ, the plug-in estimate $\{\Gamma_{n\theta}(t) : t \le \tau, \theta \in \Theta\}$ satisfies $\sup_{\theta\in\Theta}\|\Gamma_{n\theta} - \Gamma_\theta\|_\infty \to 0$ a.s. In addition, if τ_0 is a continuity point of $E_P Y(t)$, then $\sup\{|\exp(-\Gamma_\theta) - \exp(-\Gamma_{n\theta})|(t) : \theta \in \Theta, t \in \mathcal{T}\} = o_P(1)$.*

(ii) *The function $\Theta \ni \theta \to \{\Gamma_\theta(t) : t \in [0,\tau]\} \in \ell^\infty([0,\tau])$ is Frèchet differentiable with respect to θ and the derivative satisfies*

$$\dot{\Gamma}_\theta(t) = -\int_0^t \dot{s}(\Gamma_\theta(u-), \theta, u)C_\theta(du)\mathcal{P}_\theta(u,t).$$

The estimate $\{\dot{\Gamma}_{n\theta}(t) : t \le \tau, \theta \in \Theta\}$ satisfies $\sup_{\theta\in\Theta}\|\dot{\Gamma}_{n\theta} - \dot{\Gamma}_\theta\|_\infty \to 0$ a.s.

(iii) *The process $\{\hat{W}(t,\theta) = \sqrt{n}[\Gamma_{n\theta} - \Gamma_\theta](t) : t \le \tau, \theta \in \Theta\}$ converges weakly in $\ell^\infty([0,\tau] \times \Theta)$ to*

$$W(t,\theta) = R(t,\theta) - \int_{[0,t]} R(u-,\theta)C_\theta(du)V_\theta(u,t),$$

where $V_\theta(u,t) = 1(u < t)s'(\Gamma_\theta(u-), \theta, u)\mathcal{P}_\theta(u,t)$ and $R(t,\theta)$ is a mean zero Gaussian process. Its covariance function is given in Section 3.

(iv) *Let $E_P N(t)$ be continuous, and let θ_0 be an arbitrary point in Θ. If $\hat{\theta}$ is a $\sqrt{n}$-consistent estimate of it, then the process $\hat{W}_0 = \{\hat{W}_0(t) : t \le \tau\}$, $\hat{W}_0 = \sqrt{n}[\Gamma_{n\hat{\theta}} - \Gamma_{\theta_0} - (\hat{\theta} - \theta_0)\dot{\Gamma}_{\hat{\theta}}]$ converges weakly in $\ell^\infty([0,\tau])$ to $W_0 = W(\cdot, \theta_0)$.*

The proof of this proposition can be found in Section 6.

2.3. Some auxiliary notation

From now on we assume that the function $EN(t)$ is continuous. We shall need some auxiliary notation. Define

$$e[f](u,\theta) = \frac{E\{Y(u)[f\alpha](\Gamma_\theta(u),\theta,Z)\}}{E\{Y_i(u)\alpha(\Gamma_\theta(u),\theta,Z)\}},$$

where $f(x,\theta,Z)$, is a function of covariates. Likewise, for any two such functions, f_1 and f_2, let $\mathrm{cov}[f_1,f_2](u,\theta) = e[f_1 f_2^T](u,\theta) - (e[f_1]e[f_2]^T)(u,\theta)$ and $\mathrm{var}[f](u,\theta) = \mathrm{cov}[f,f](u,\theta)$. We shall write

$$\begin{aligned} e(u,\theta) &= e[\ell'](u,\theta), \qquad \bar{e}(u,\theta) = e[\dot{\ell}](u,\theta), \\ v(u,\theta) &= \mathrm{var}[\ell'](u,\theta), \quad \bar{v}(u,\theta) = \mathrm{var}[\dot{\ell}](u,\theta), \quad \rho(u,\theta) = \mathrm{cov}[\dot{\ell},\ell'](u,\theta), \end{aligned}$$

for short. Further, let

$$\begin{aligned} K_\theta(t,t') &= \int_0^{t\wedge t'} C_\theta(du)\mathcal{P}_\theta(u,t)\mathcal{P}_\theta(u,t'), \\ B_\theta(t) &= \int_0^t v(u,\theta)EN(du) \end{aligned} \tag{2.4}$$

and define

$$\kappa_\theta(\tau) = \int\int_{0<u\le t\le\tau} C_\theta(du)\mathcal{P}_\theta(u,t)^2 B_\theta(dt). \tag{2.5}$$

This constant is finite for any point τ satisfying the condition 2.0 (iii), but is in general infinite, if τ_0 is a continuity point of the survival function $EY(t)$. Finally, we set

$$\begin{aligned} v_\varphi(t,\theta) &= \bar{v}(t,\theta) + v(t,\theta)\varphi_\theta^{\otimes 2}(t) - \rho(t,\theta)\varphi_\theta^T(t) - \varphi_\theta(t)\rho(t,\theta)^T, \\ \rho_\varphi(t,\theta) &= \rho(t,\theta) - v(t,\theta)\varphi_\theta(t), \end{aligned}$$

for any function φ_θ square integrable with respect to B_θ. Under the added condition 2.2, we have

$$\begin{aligned} e(u,\theta_0) &= E[\ell'(\Gamma_{\theta_0}(X),\theta_0,Z)|X=u,\delta=1], \\ v(u,\theta_0) &= \mathrm{var}\,[\ell'(\Gamma_{\theta_0}(X),\theta_0),Z|X=u,\delta=1], \\ \bar{e}(u,\theta_0) &= E[\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z)|X=u,\delta=1], \\ \bar{v}(u,\theta_0) &= \mathrm{var}\,[\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z)|X=u,\delta=1], \\ \rho(u,\theta_0) &= \mathrm{cov}[\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z),\ell'(\Gamma_{\theta_0}(X),\theta_0,Z)|X=u,\delta=1]. \end{aligned}$$

Similarly,

$$\begin{aligned} v_\varphi(u,\theta_0) &= \mathrm{var}[\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z) - \ell'(\Gamma_{\theta_0}(X),\theta_0,Z)\varphi_{\theta_0}(X)|X=u,\delta=1], \\ \rho_\varphi(u,\theta_0) &= \mathrm{cov}[\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z) \\ &\qquad - \ell'(\Gamma_{\theta_0}(X)\theta_0,Z)\varphi_{\theta_0}(X),\ell'(\Gamma_{\theta_0}(X),\theta_0,Z)|X=u,\delta=1]. \end{aligned}$$

However, $e[f], \mathrm{var}[f]$ and $\mathrm{cov}[f,g]$ form conditional expectation and variance–covariance operators even when this assumption fails. This observation, the Cauchy-Schwarz inequality, and the monotone convergence theorem can be used to verify the next lemma.

Lemma 2.1. *Suppose that Conditions* 2.0 *and* 2.1 *are satisfied. Let* $EN(t)$ *be continuous, and let* $v(u,\theta) \not\equiv 0$ *a.e.* EN.

(i) *If* $\kappa_\theta(\tau_0) < \infty$ *then the kernel* K_θ *is square integrable with respect to* B_θ. *In addition, if* $e[(\dot\ell)^{\otimes 2}](u,\theta) \in L_1(EN)$ *then* $\dot\Gamma_\theta \in L_2(B_\theta)$.

(ii) *Suppose that the integrability conditions of part (i) are satisfied. For any vector valued function* $\varphi_\theta(t) = \int_0^t g_\theta d\Gamma_\theta \in L_2(B_\theta)$ *the matrices*

$$\begin{aligned}
\Sigma_{0,\varphi}(\theta,\tau) &= \int_0^\tau v_\varphi(t,\theta) EN(dt),\\
\Sigma_{1,\varphi}(\theta,\tau) &= \Sigma_{0,\varphi}(\theta,\tau) + \int_0^\tau \rho_\varphi(t,\theta)[\dot\Gamma_\theta(t) + \varphi_\theta(t)]^T EN(dt),\\
\Sigma_{2,\varphi}(\theta,\tau) &= \Sigma_{0,\varphi}(\theta,\tau)\\
&\quad + \int_0^\tau \int_0^\tau K_\theta(t,u)\rho_\varphi(t,\theta)\rho_\varphi(u,\theta)^T EN(du)EN(dt)
\end{aligned}$$

have finite components for any point $\tau \le \tau_0$.

Here $L_1(EN)$ is the space of functions integrable with respect to EN and $L_2(B_\theta)$ is the space of functions square integrable with respect to B_θ.

Remark 2.1. For $\varphi_\theta = -\dot\Gamma_\theta$, we have $\Sigma_{1,\varphi}(\theta,\tau) = \Sigma_{2,\varphi}(\theta,\tau)$ if $\rho_{-\dot\Gamma}(u,\theta) \equiv 0$ and $v(u,\theta) \not\equiv 0$ a.e. EN. If $v(u,\theta) \equiv 0$, then for the sake of completeness, we define $\Sigma_{1,\varphi}(\theta,\tau) = \Sigma_{2,\varphi}(\theta,\tau) = \int_0^\tau \bar v(u,\theta) EN(du)$. In this case $\ell_i'(x,\theta,z)$ is a function not depending on covariates. In particular, in the proportional hazards model, we have, $\ell_i'(x,\theta,z) \equiv 0$ for all θ. In scale regression models with hazards $\alpha(x,\theta,z) = \exp(\theta^T z)\alpha_0(x\exp(\theta^T z))$, where α_0 is a known function, we have $\ell_i'(x,\theta,z) = \alpha_0'(x)/\alpha_0(x)$ for $\theta = 0$ (independence).

We shall assume now that the point τ satisfies the condition 2.0 (iii). With this choice, any function $\varphi_\theta(t) = \int_0^t g_\theta d\Gamma_\theta$ of bounded variation on $[0,\tau]$ is square integrable with respect to the measure B_θ, restricted to the interval $[0,\tau]$. However, in Proposition 2.3, we shall allow also for $\tau = \tau_0$ to be a continuity point of the survival function $EY_P(t)$ and assume integrability conditions of Lemma 2.1.

2.4. Estimation of the Euclidean component of the model

To estimate the parameter θ, we use a solution to the score equation $U_n(\theta) = U_{n\varphi_n}(\theta) = 0$, where

$$\begin{aligned}
U_{n\varphi_n}(\theta) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau [b_{1i}(\Gamma_{n\theta}(t),t,\theta) - b_{2i}(\Gamma_{n\theta}(t),t,\theta)\varphi_{n\theta}(t)]N_i(dt), \qquad (2.6)\\
b_{1i}(x,t,\theta) &= \dot\ell(x,\theta,Z_i) - [\dot S/S](x,\theta,t),\\
b_{2i}(x,t,\theta) &= \ell'(x,\theta,Z_i) - [S'/S](x,\theta,t)
\end{aligned}$$

and $\varphi_{n\theta}(t)$ is an estimate of a function $\varphi_\theta(t) = \int_0^t g_\theta d\Gamma_\theta$. We shall make the following regularity assumption.

Condition 2.3. Suppose that Conditions 2.0–2.2 hold, and let $\|\cdot\|_v$ be the variation norm on the interval $[0,\tau]$. Let $B(\theta_0,\varepsilon_n) = \{\theta : |\theta - \theta_0| \le \varepsilon_n\}$ for some sequence $\varepsilon_n \downarrow 0, \sqrt{n}\varepsilon_n \to \infty$. In addition

(i) The matrix $\Sigma_{0,\varphi}(\theta_0,\tau)$ is positive definite.
(ii) The matrix $\Sigma_{1,\varphi}(\theta_0,\tau)$ is non-singular.
(iii) The function $\varphi_{\theta_0}(t) = \int_0^t g_{\theta_0} d\Gamma_{\theta_0}$ satisfies $\|\varphi_{\theta_0}\|_v = O(1)$,
(iv) $\|\varphi_{n\theta_0} - \varphi_{\theta_0}\|_\infty \to_P 0$ and $\limsup_n \|\varphi_{n\theta_0}\|_v = O_P(1)$.
(v) We have either

(v.1) $\varphi_{n\theta} - \varphi_{n\theta'} = (\theta - \theta')\psi_{n\theta,\theta'}$, where $\limsup_n \sup\{\|\psi_{n\theta,\theta'}\|_v : \theta, \theta' \in B(\theta_0, \varepsilon_n)\} = O_P(1)$ or

(v.2) $\limsup_n \sup\{\|\varphi_{n\theta}\|_v : \theta \in B(\theta_0, \varepsilon_n)\} = O_P(1)$ and $\sup\{\|\varphi_{n\theta} - \varphi_{\theta_0}\|_\infty : \theta \in B(\theta_0, \varepsilon_n)\} = o_P(1)$.

Proposition 2.2. *Suppose that Conditions* 2.3(i)–(iv) *hold.*

(i) *For any $\sqrt{n}$ consistent estimate $\hat\theta$ of the parameter θ_0, $\hat W_0 = \sqrt{n}[\Gamma_{n\hat\theta} - \Gamma_{\theta_0} - (\hat\theta - \theta_0)\dot\Gamma_{\hat\theta}]$ converges weakly in $\ell^\infty([0,\tau])$ to a mean zero Gaussian process W_0 with covariance function $\mathrm{cov}(W_0(t), W_0(t')) = K_{\theta_0}(t,t')$.*

(ii) *Suppose that Condition* 2.3(v.1) *is satisfied. Then, with probability tending to 1, the score equation $U_{n\varphi_n}(\theta) = 0$ has a unique solution $\hat\theta$ in $B(\theta_0, \varepsilon_n)$. Under Condition* 2.3(v.2)*, the score equation $U_{n\varphi_n}(\theta) = o_P(n^{-1/2})$ has a solution, with probability tending to 1.*

(iii) *Define $[\hat T, \hat W_0], \hat T = \sqrt{n}(\hat\theta - \theta_0)$, $\hat W_0 = \sqrt{n}[\Gamma_{n\hat\theta} - \Gamma_{\theta_0} - (\hat\theta - \theta_0)\dot\Gamma_{\hat\theta}]$, where $\hat\theta$ are the estimates of part* (ii)*. Then $[\hat T, \hat W_0]$ converges weakly in $R^p \times \ell^\infty([0,\tau])$ to a mean zero Gaussian process $[T, W_0]$ with covariance $\mathrm{cov}T = \Sigma_1^{-1}(\theta_0,\tau)\Sigma_2(\theta_0,\tau)[\Sigma_1^{-1}(\theta_0,\tau)]^T$ and*

$$\mathrm{cov}(T, W_0(t)) = -\Sigma_1^{-1}(\theta_0,\tau)\int_0^\tau K_{\theta_0}(t,u)\rho_\varphi(u,\theta_0)EN(du).$$

Here the matrices $\Sigma_{q,\varphi}, q = 1, 2$ are defined as in Lemma 2.2*.*

(iv) *Let $\tilde\theta_0$ be any $\sqrt{n}$ estimate, and let $\hat\varphi_n = \varphi_{n\tilde\theta_0}$ be an estimator of the function φ_{θ_0} such that $\|\hat\varphi_n - \varphi_{\theta_0}\|_\infty = o_P(1)$ and $\limsup_n \|\hat\varphi_n\|_v = O_P(1)$. Define a one-step M-estimator $\hat\theta = \tilde\theta_0 + \Sigma_{1\hat\varphi_n}(\tilde\theta_0,\tau)^{-1}U_{n\hat\varphi_n}(\tilde\theta_0)$, where $\Sigma_{1,\hat\varphi_n}$ is the plug-in analogue of the matrix $\Sigma_{1,\varphi}(\theta_0,\tau)$. Then part* (iii) *holds for the one-step estimator $\hat\theta$.*

The proof of this proposition is postponed to Section 7.

Example 2.1. A simple choice of the φ_θ function is provided by $\varphi_\theta \equiv 0 = \varphi_{n\theta}$. The resulting score equation, is approximately equal to

$$\hat U_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left[N_i(\tau)\dot\ell(\Gamma_{n\theta}(X_i), \theta, Z_i) - \dot A(\Gamma_{n\theta}(X_i \wedge \tau), \theta, Z_i)\right],$$

and this score process may be easier to compute in some circumstances. If the transformation Γ had been known, the right-hand side would have represented the MLE score function for estimation of the parameter θ. Using results of section 5, we can show that solving equation $\hat U_n(\theta) = 0$ or $\hat U_n(\theta) = o_P(n^{-1/2})$ for θ leads to an M estimator asymptotically equivalent to the one in Proposition 2.2. However, this equivalence holds only at rate $\sqrt{n}$. In particular, at the true θ_0, the two score processes satisfy $\sqrt{n}|\hat U_n(\theta_0) - U_n(\theta_0)| = o_P(1)$, but they have a different higher order expansions.

Example 2.2. The second possible choice corresponds to $\varphi_\theta = -\dot\Gamma_\theta$. The score function $U_n(\theta)$ is in this case approximately equal to the derivative of the pseudo-profile likelihood criterion function considered by Bogdanovicius and Nikulin [9] in

the case of generalized proportional hazards intensity models. Using results of section 6, we can show that the sample analogue of the function $\dot{\Gamma}_\theta$ satisfies Conditions 2.3(iv) and 2.3(v).

Example 2.3. The logarithmic derivatives of $\ell(x,\theta,Z) = \log\alpha(x,\theta,Z)$ may be difficult to compute in some models, so we can try to replace them by different functions. In particular, suppose that $h(x,\theta,Z)$ is a differentiable function with respect to both arguments and the derivatives satisfy a similar Lipschitz continuity assumption as in condition 2.1. Consider the score process (2.6) with function $\varphi_\theta = 0$ and weights $b_{1i}(x,t,\theta) = h(x,\theta,Z_i) - [S_h/S](x,\theta,t)$ where $S_h(x,\theta,t) = \sum_{i=1}^n Y_i(u)[h\alpha](x,\theta,Z_i)$, and $\varphi_{n\theta} \equiv 0$. For $p=0$ and $p=2$, define matrices $\Sigma^h_{p\varphi}$ by replacing the functions v_φ and ρ_φ appearing in matrices $\Sigma_{0\varphi}$ and $\Sigma_{2\varphi}$ with

$$\begin{aligned} v^h_\varphi(t,\theta_0) &= \operatorname{var}[h(\Gamma_{\theta_0}(X),\theta_0,Z)|X=t,\delta=1], \\ \rho^h_\varphi(t,\theta_0) &= \operatorname{cov}[h(\Gamma_{\theta_0}(X_i),\theta_0,Z_i),\ell'(\Gamma_{\theta_0}(X_i),\theta_0,Z_i)|X=t,\delta=1]. \end{aligned}$$

The matrix $\Sigma_{1\varphi}(\theta_0,\tau)$ is changed to $\Sigma^h_{1\varphi}(\theta_0,\tau) = \int \bar{\rho}^h_\varphi(t,\theta_0)EN(du)$, where the integrand is equal to

$$\operatorname{cov}[h(\Gamma_{\theta_0}(X),\theta_0,Z),\dot{\ell}(\Gamma_{\theta_0}(X),\theta_0,Z) + \ell'(\Gamma_{\theta_0}(X),\theta_0,Z)\dot{\Gamma}_{\theta_0}(X)|X=t,\delta=1].$$

The statement of Proposition 2.2 remains valid with matrices $\Sigma_{p\varphi}$ replaced by $\Sigma^h_{p\varphi}$, $p=1,2$, provided in Condition 2.3 we assume that the matrix $\Sigma^h_{0\varphi}$ is positive definite and the matrix $\Sigma^h_{1\varphi}$ is non-singular. The resulting estimates have a structure analogous to that of the M-estimates considered in the case of uncensored data by Bickel *et al.* [6] and Cuzick [13]. Alternatively, instead of functions $\dot{\ell}_i(x,\theta,z)$ and $\ell'(x,\theta,z)$, the weight functions b_{1i} and b_{2i} can use logarithmic derivatives of a different distribution, with the same parameter θ. The asymptotic variance is of similar form as above. In both cases, the derivations are similar to Section 7, so we do not consider analysis of these score processes in any detail.

Example 2.4. Our final example shows that we can choose the φ_θ function so that the asymptotic variance of the estimate $\hat{\theta}$ is equal to the inverse of the asymptotic variance of the normalized score process, $\sqrt{n}U_n(\theta_0)$. Remark 2.1 implies that if $\rho_{-\dot{\Gamma}}(u,\theta_0) \equiv 0$ but $v(u,\theta_0) \not\equiv 0$ a.e. EN, then for $\varphi_\theta = -\dot{\Gamma}_\theta$ the matrices $\Sigma_{q,\varphi}, q = 1,2$ are equal. This also holds for $v(u,\theta_0) \equiv 0$. We shall consider now the case $v(u,\theta_0) \not\equiv 0$ and $\rho_{-\dot{\Gamma}}(u,\theta_0) \not\equiv 0$ a.e. EN, and without loss of generality, we shall assume that the parameter θ is one dimensional.

We shall show below that the equation

$$\begin{aligned} \varphi_\theta(t) + \int_0^\tau K_\theta(t,u)v(u,\theta)\varphi_\theta(u)EN(du) \\ = -\dot{\Gamma}_\theta(t) + \int_0^\tau K_\theta(t,u)\rho(u,\theta)EN(du) \end{aligned} \tag{2.7}$$

has a unique solution φ_θ square integrable with respect to the measure (2.4). For $\theta = \theta_0$, the corresponding matrices $\Sigma_{1,\varphi}(\theta_0,\tau)$ and $\Sigma_{2,\varphi}(\theta_0,\tau)$ are finite. Substitution of the conditional correlation function $\rho_\varphi(t,\theta_0) = \rho(t,\theta_0) - \varphi_{\theta_0}(t)v(t,\theta_0)$ into the matrix $\Sigma_{2,\varphi}(\theta_0,\tau)$ shows that they are also equal. (In the multiparameter case, the equation (2.7) is solved for each component of the θ).

Equation (2.7) simplifies if we replace the function φ_θ by $\psi_\theta = \varphi_\theta + \dot{\Gamma}_\theta$. We get

$$\psi_\theta(t) - \lambda\int_0^\tau K_\theta(t,u)\psi_\theta(u)B_\theta(du) = \eta_\theta(t), \tag{2.8}$$

where $\lambda = -1$,

$$\eta_\theta(t) = \int_0^\tau K_\theta(t,u)\rho_{-\dot\Gamma}(u,\theta)EN(du),$$

$\rho_{-\dot\Gamma}(u,\theta) = v(u,\theta)\dot\Gamma_\theta(u)+\rho(u,\theta)$ and B_θ is given by (2.4). For fixed θ, the kernel K_θ is symmetric, positive definite and square integrable with respect to B_θ. Therefore it can have only positive eigenvalues. For $\lambda = -1$, the equation has a unique solution given by

$$\psi_\theta(t) = \eta_\theta(u) - \int_0^\tau \Delta_\theta(t,u,-1)\eta_\theta(u)B_\theta(du), \tag{2.9}$$

where $\Delta_\theta(t,u,\lambda)$ is the resolvent corresponding to the kernel K_θ. By definition, the resolvent satisfies a pair of integral equations

$$\begin{aligned} K_\theta(t,u) &= \Delta_\theta(t,u,\lambda) - \lambda\int_0^\tau \Delta_\theta(t,w,\lambda)B_\theta(dw)K_\theta(w,u) \\ &= \Delta_\theta(t,u,\lambda) - \lambda\int_0^\tau K_\theta(t,w)B_\theta(dw)\Delta_\theta(w,u,\lambda), \end{aligned}$$

where integration is with respect to different variables in the two equations. For $\lambda = -1$ the solution to the equation is given by

$$\begin{aligned} \psi_\theta(t) &= \int_0^\tau K_\theta(t,u)\rho_{-\dot\Gamma}(u,\theta)EN(du) \\ &\quad - \int_0^\tau \Delta_\theta(t,w,-1)B_\theta(dw)\int_0^\tau K_\theta(w,u)\rho_{-\dot\Gamma}(u,\theta)EN(du) \end{aligned}$$

and the resolvent equations imply that the right-hand side is equal to

$$\psi_\theta(t) = \int_0^\tau \Delta_\theta(t,u,-1)\rho_{-\dot\Gamma}(u,\theta)EN(du). \tag{2.10}$$

For $\theta = \theta_0$, substitution of this expression into the formula for the matrices $\Sigma_{1,\varphi}(\theta_0,\tau)$ and $\Sigma_{2,\varphi}(\theta_0,\tau)$ and application of the resolvent equations yields also

$$\begin{aligned} \Sigma_{1,\varphi}(\theta_0,\tau) &= \Sigma_{2,\varphi}(\theta_0,\tau) \\ &= \int_0^\tau v_{-\dot\Gamma}(u,\theta_0)EN(du) \\ &\quad - \int_0^\tau\int_0^\tau \Delta_{\theta_0}(t,u,-1)\rho_{-\dot\Gamma}(u,\theta_0)\rho_{-\dot\Gamma}(t,\theta_0)^T EN(du)EN(dt). \end{aligned}$$

It remains to find the resolvent Δ_θ. We shall consider first the case of $\theta = \theta_0$. To simplify algebra, we multiply both sides of the equation (2.8) by $\mathcal{P}_{\theta_0}(0,t)^{-1} = \exp\int_0^t s'(\theta_0,\Gamma_{\theta_0}(u),u)C_{\theta_0}(du)$. For this purpose set

$$\begin{aligned} \tilde\psi(t) &= \mathcal{P}_{\theta_0}(0,t)^{-1}\psi(t), & \dot G(t) &= \mathcal{P}_{\theta_0}(0,t)^{-1}\dot\Gamma_{\theta_0}(t), \\ \tilde v(t,\theta_0) &= v(t,\theta_0)\mathcal{P}_{\theta_0}(0,t)^2, & \tilde\rho_{-\dot G}(t,\theta_0) &= \mathcal{P}_{\theta_0}(0,t)\rho_{-\dot\Gamma}(t,\theta_0), \\ b(t) &= \int_0^t \tilde v(u,\theta_0)dEN(u), & c(t) &= \int_0^t \mathcal{P}_{\theta_0}(0,u)^{-2}dC_{\theta_0}(u). \end{aligned}$$

Multiplication of (2.8) by $\mathcal{P}_{\theta_0}(0,t)^{-1}$ yields

$$\tilde\psi(t) + \int_0^\tau k(t,u)\tilde\psi(u)b(du) = \int_0^\tau k(t,u)\tilde\rho_{-\dot G}(u,\theta_0)EN(du), \tag{2.11}$$

where the kernel k is given by $k(t,u) = c(t \wedge u)$. Since this is the covariance function of a time transformed Brownian motion, we obtain a simpler equation. The solution to this Fredholm equation is

$$\tilde{\psi}(t) = \int_0^\tau \tilde{\Delta}(t,u)\tilde{\rho}_{-\dot{G}}(u,\theta_0)EN(du), \tag{2.12}$$

where $\tilde{\Delta}(t,u) = \tilde{\Delta}(t,u,-1)$, and $\tilde{\Delta}(t,u,\lambda)$ is the resolvent corresponding to the kernel k. More generally, we consider the equation

$$\tilde{\psi}(t) + \int_0^\tau k(t,u)\tilde{\psi}(u)b(du) = \tilde{\eta}(t). \tag{2.13}$$

Its solution is of the form

$$\tilde{\psi}(t) = \tilde{\eta}(t) - \int_0^\tau \tilde{\Delta}(t,u)b(du)\tilde{\eta}(u).$$

To give the form of the $\tilde{\Delta}$ function, note that the constant $\kappa_{\theta_0}(\tau)$ defined in (2.5) satisfies

$$\kappa(\tau) = \kappa_{\theta_0}(\tau) = \int_0^\tau c(u)b(du).$$

Proposition 2.3. *Suppose that Assumptions* 2.0(i) *and* (ii) *are satisfied and* $v(u,\theta_0) \not\equiv 0$*, For* $j = 0,1,2,3$*,* $n \geq 1$ *and* $s < t$ *define interval functions* $\Psi_j(s,t) = \sum_{m=0}^\infty \Psi_{jm}(s,t)$ *as follows:*

$$\begin{aligned}
\Psi_{00}(s,t) &= 1, \quad \Psi_{20}(s,t) = 1,\\
\Psi_{0n}(s,t) &= \int\int_{s<u_1\leq u_2\leq t} \Psi_{0,n-1}(s,u_1-)c(du_1)b(du_2) \quad n \geq 1,\\
\Psi_{1n}(s,t) &= \int_{(s,t]} \Psi_{0n}(s,u-)c(du) \quad n \geq 0,\\
\Psi_{2n}(s,t) &= \int\int_{s\leq u_1<u_2<t} b(du_1)c(du_2)\Psi_{2,n-1}(u,t)\\
&= \int_{[s,t)} b(du_1)\Psi_{1,n-1}(u,t-) \quad n \geq 1,\\
\Psi_{3n}(s,t) &= \int_{[s,t)} \Psi_{2n}(s,u)b(dw) \quad n \geq 0.
\end{aligned}$$

For $j = 2,3$*, define* $\Psi_{jn}(s,t+)$ *by replacing the intervals* $[s,t)$ *with* $[s,t]$ *in the last two lines, and similarly, define* $\Psi_{jn}(s,t-)$ *by replacing intervals* $(s,t]$ *with* (s,t) *in the first two definitions. For* $s > t$*, set* $\Psi_j(s,t) = 0, j = 0,1,2,3$ *and let* $\Psi_{j0}(t,t) = 1$ *for* $j = 0,2$*,* $\Psi_{10}(t,t) = c(\Delta t)$*,* $\Psi_{30}(t,t) = b(\Delta t)$*, and* $\Psi_{jn}(t,t) = 0$ *for* $n \geq 1$*,* $j = 0,1,2,3$*.*

(i) *We have*

$$\begin{aligned}
\Psi_0(s,t) &= 1 + \int_{(s,t]} \Psi_1(s,u)b(du) = 1 + \int_{(s,t]} c(du)\Psi_3(u,t+),\\
\Psi_1(s,t) &= \int_{(s,t]} \Psi_0(s,u-)c(du) = \int_{(s,t]} c(du)\Psi_2(u,t+),\\
\Psi_2(s,t) &= 1 + \int_{[s,t)} b(du)\Psi_1(u,t-) = 1 + \int_{[s,t)} \Psi_3(s,u)c(du),\\
\Psi_3(s,t) &= \int_{[s,t)} \Psi_2(s,u)b(du) = \int_{[s,t)} b(du)\Psi_0(u,t-).
\end{aligned}$$

For any point τ satisfying Condition 2.0(iii), $\Psi_j, j = 0, 1, 2, 3$ *form bounded monotone increasing interval functions. In particular,* $\Psi_0(s,t) \leq \exp \kappa(\tau)$ *and* $\Psi_1(s,t) \leq \Psi_0(s,t)[c(t) - c(s)]$. *In addition if* τ_0 *is a continuity point of the survival function* $E_P Y(t)$ *and* $\kappa(\tau_0) < \infty$, *then* $\Psi_0(s,t) \leq \exp \kappa(\tau_0)$ *for any* $0 < s < t \leq \tau_0$, *while the remaining functions are locally bounded.*

(ii) *Suppose that* τ *satisfies Condition* 2.0(iii), *or else* $\tau = \tau_0$, τ_0 *is a continuity point of* $E_P Y(t)$ *and* $\kappa(\tau_0) < \infty$. *The resolvent of the kernel* k *is given by* $\tilde{\Delta}(s,t,-1) = \tilde{\Delta}(s,t) = \Psi_0(0,\tau)^{-1}\Psi_1(0, s \wedge t)\Psi_0(s \vee t, \tau)$.

(iii) *Under assumptions in* (ii), *for any* $\tilde{\eta} \in L_2(b)$, *the solution to equation* (2.13) *satisfies* $\tilde{\psi} \in L_2(b)$, *and and* $\|\tilde{\psi}\|_2 \leq \|\tilde{\eta}\|_2[1 + \Psi_0(0,\tau_0)\kappa(\tau_0)]$, *where* $\|\cdot\|_2$ *is the* L_2 *norm with respect to the measure* b.

(iv) *Suppose that* τ *satisfies Condition* 2.0(iii). *If* $\tilde{\eta}$ *is a bounded function or a function of bounded variation, then the solution* $\tilde{\psi}$ *has the same properties and the bounds of part* (ii) *hold in supremum and variation norm, respectively.*

(v) *The solution to equation* (2.7) *is given by*

$$\begin{aligned}\varphi_{\theta_0}(t) &= -\dot{\Gamma}_{\theta_0}(t) \\ &\quad + \int_0^\tau \tilde{\Delta}(t,u)\rho_{-\dot{\Gamma}}(u,\theta_0)EN(du)\mathcal{P}_{\theta_0}(0,u)\mathcal{P}_{\theta_0}(0,t).\end{aligned}$$

Under assumptions of part (ii) *and integrability conditions of Lemma* 2.2, *we have* $\varphi \in L_2(B_{\theta_0})$. *We also have,* $\varphi_{\theta_0}(t) = \int_0^t g_{\theta_0} d\Gamma_{\theta_0}$, *where*

$$\begin{aligned}g_{\theta_0}(u) &= [\dot{s}/s](\Gamma_{\theta_0}(u),\theta_0,u) - [s'/s](\Gamma_{\theta_0}(u),\theta_0,u)\varphi_{\theta_0}(u) \\ &\quad - s(\Gamma_{\theta_0}(u),\theta_0,u)^{-1}\int_u^\tau \mathcal{P}_{\theta_0}(u,t)\rho_\varphi(u,\theta_0)EN(du),\end{aligned}$$

(vi) *If* τ *satisfies Condition* 2.0(iii), *then the solution* φ *is a function of bounded variation. Moreover, the constant* $W = \Psi(0,\tau)$ *satisfies* $W = \Psi_1(0,t) \times \Psi_3(t,\tau) + \Psi_0(0,t)\Psi_0(t,\tau)$ *for any* $0 < t \leq \tau$, *and*

$$\begin{aligned}\varphi_{\theta_0}(t) &= \int_0^\tau \tilde{\Delta}(t,u)\rho(u,\theta_0)EN(du)\mathcal{P}_{\theta_0}(0,u)\mathcal{P}_{\theta_0}(0,t) \\ &\quad + \int_0^\tau \bar{\Delta}(t,u)\dot{s}(\Gamma_0(u),\theta_0,u)c(du)\mathcal{P}_{\theta_0}(0,u)\mathcal{P}_{\theta_0}(0,t),\end{aligned}$$

where $\bar{\Delta}(t,u) = W^{-1}[\Psi_0(0,u \wedge t)\Psi_0(u \vee t,\tau) - \Psi_1(0,u \wedge t)\Psi_3(u \vee t,\tau)]$.

The proof of this proposition is given in Section 8.

We have chosen to transform equation (2.8) in order to simplify calculations. The resolvent of the kernel K corresponding to equation (2.8) can be obtained based on recurrent Fredholm determinant formulas [25] applied to the kernel K. The same arguments can be applied to find the solution to equation (2.8) for $\theta \neq \theta_0$. The only difference is that the kernel function $K_\theta(t,u)$ does not represent the asymptotic covariance function of the process $\sqrt{n}[\Gamma_{n\theta} - \Gamma_\theta]$ for such θ points.

The sample analogue of the function ψ_θ can be obtained in several different manners. Firstly, equations (2.7)–(2.8) can be solved directly by plugging in sample analogues of the functions K, v, ρ etc. If these sample analogues are functions placing mass at each uncensored observation, then this choice is not convenient, because to solve the equation one must eventually invert an $m \times m$ dimensional matrix (here m is the number of uncensored observations in the sample). Proposition 2.3 provides

a simpler form of this equation. Define estimates

$$\begin{aligned} c_{n\theta}(t) &= \int_0^t \tilde{\mathcal{P}}_{n\theta}(0,u)^{-2} C_{n\theta}(du), \\ b_{n\theta}(t) &= \int_0^t \tilde{\mathcal{P}}_{n\theta}(0,u)^{2} B_{n\theta}(du), \\ C_{n\theta}(du) &= \int_0^t S(\Gamma_{n\theta}(u-),\theta,u)^{-2} N_{\cdot}(du), \\ \tilde{\mathcal{P}}_{n\theta}(u,t) &= \exp[-\int_u^t S'(\Gamma_{\theta}(u-),\theta,u) C_{n\theta}(du)], \end{aligned}$$

and let $B_{n\theta}$ be the plug-in analogue of the formula (2.4). Let $X^*_{(1)} < \cdots < X^*_{(m)}$ be the distinct ordered uncensored observations in the sample. Then the discrete version of equations (2.11)–(2.13) is given by

$$\tilde{\psi}_{n\theta}(X^*_{(j)}) + \sum_{i=1}^{m} c_{n\theta}(X^*_{(i)} \wedge X^*_{(j)}) b_{n\theta}(\Delta X^*_{(i)}) \tilde{\psi}_{n\theta}(X^*_{(i)}) = \tilde{\eta}_{n\theta}(X^*_{(j)}).$$

Using Proposition 2.3, we have shown in an earlier version of this text that finding solution to this equation amounts to inversion of a bandsymmetric tridiagonal matrix which can be easily implemented in practice. A numerical example is given in [16]. The formula (2.14) in part (v) gives in this case an estimate of the function φ_θ corresponding to the equation (2.7). We show in [17] that it satisfies Conditions 2.3.

Finally, we show that in the continuous case, equation (2.11) corresponds to a Sturm–Liouville equation. Suppose that the point τ satisfies Condition 2.0(iii). By twice "differentiating" (2.11) with respect to dc, we obtain a Sturm–Liouville equation

$$\frac{d}{dc}[\frac{d}{dc}\tilde{\psi}](t) = \tilde{\psi}(t)\frac{db}{dc}(t) - \tilde{\rho}_{-\dot{G}}(t,\theta_0)\frac{EN}{dc}(t)$$

with boundary conditions $\tilde{\psi}(0) = 0, \frac{d}{dc}\tilde{\psi}(t)_{|t=\tau} = 0$. Its solution is of the form (2.12) with $\tilde{\Delta}$ representing the Green's function associated with the homogeneous equation

$$\frac{d}{dc}[\frac{d}{dc}\tilde{\psi}](t) = \tilde{\psi}(t)\frac{db}{dc}(t) \tag{2.14}$$

and boundary conditions $\tilde{\psi}(0) = 0, \frac{d}{dc}\tilde{\psi}(\tau) = 0$. The Green's function is given by $\tilde{\Delta}(t,u) = \tilde{W}^{-1}[\psi_1(t \wedge u)\psi_0(t \vee u)]$, where ψ_1 and ψ_0 is a pair of fundamental solutions, ψ_1 corresponding to the left boundary ($\psi_1(0) = 0$) and ψ_0 corresponding to the right-boundary ($\frac{d}{dc}\psi_0(\tau) = 0$). Moreover,

$$\tilde{W} = -[\psi_1(t)\frac{d}{dc}\psi_0(t) - \psi_0(t)\frac{d}{dc}\psi_1(t)]$$

is the negative Wronskian (the right-hand side is a constant, not depending on t). By twice integrating the homogeneous equation subject to the boundary conditions, we obtain a pair of Volterra equations whose solutions are $a_1\Psi_1(0,t)$ and $a_0\Psi_0(t,\tau)$, where $a_p \neq 0$ are arbitrary constants. The choice of $a_p = 1$, corresponds to the Volterra equations for $\Psi_0(t,\tau)$ and $\Psi_1(0,t)$ discussed in part (i). We also have $\tilde{W} = a_0 a_1 \Phi_0(0,\tau) = a_0 a_1 W$. Thus the $\tilde{\Delta}$ function of Proposition 2.3 is the Green's function of this Sturm–Liouville equation. Note that this is a different equation than in Bickel [4] and Bickel *et al.* [6]. In particular, it derives its form from the covariance function of a time transformed Brownian motion, rather than Brownian Bridge.

3. Examples

In this section we assume the conditional independence Assumption 2.2 and discuss Condition 2.1(ii) in more detail. It assumes that the hazard rate satisfies $m_1 \leq \alpha(x,\theta,z) \leq m_2$ μ a.e. z. This holds for example in the proportional hazards model, if the covariates are bounded and the regression coefficients vary over a bounded neighbourhood of the true parameter. Recalling that for any $P \in \mathcal{P}$, $\mathcal{X}(\mathcal{P})$ is the set of (sub)-distribution functions whose cumulative hazards satisfy $m_2^{-1}A \leq g \leq m_1^{-1}A$ and A is the cumulative hazard function (2.1), this uniform boundedness is used in Section 6 to verify that equation (2.2) has a unique solution which is defined on the entire support of the withdrawal time distribution. This need not be the case in general, as the equation may have an explosive solution on an interval strictly contained in the support of this distribution ([20]).

We shall consider now the case of hazards $\alpha(x,\theta,z)$ which for μ almost all z are locally bounded and locally bounded away from 0. A continuous nonnegative function f on the positive half-line is referred to here as locally bounded and locally bounded away from 0, if $f(0) > 0$, $\lim_{x\uparrow\infty} f(x)$ exists, and for any $d > 0$ there exists a finite positive constant $k = k(d)$ such that $k^{-1} \leq f(x) \leq k$ for $x \in [0,d]$. In particular, hazards of this form may form unbounded functions growing to infinity or functions decaying to 0 as $x \uparrow \infty$.

To allow for this type of hazards, we note that the transformation model assumes only that the conditional cumulative hazard function of the failure time T is of the form $H(t|z) = A(\tilde{\Gamma}(t),\theta|z)$ for some *unspecified* increasing function $\tilde{\Gamma}$. We can choose it as $\tilde{\Gamma} = \Phi(\Gamma)$, where Φ is a known increasing differentiable function mapping positive half-line onto itself, $\Phi(0) = 0$. This is equivalent to selection of the reparametrized core model with cumulative hazard function $A(\Phi(x),\theta|z)$ and hazard rate $\alpha(\Phi(x),\theta|z)\varphi(\Phi(x)), \varphi = \Phi'$. If in the original model the hazard rate decays to 0 or increases to infinity at its tails, then in the reparametrized model the hazard rate may form a bounded function. Our results imply in this case that we can define a family of transformations $\tilde{\Gamma}_\theta$ bounded between $m_2^{-1}A(t)$ and $m_1^{-1}A(t)$, This in turn defines a family of transformations Γ_θ bounded between $\Phi^{-1}(m_2^{-1}A(t))$ and $\Phi^{-1}(m_1^{-1}A(t))$. More generally, the function Φ may depend on the unknown parameter θ and covariates. Of course selection of this reparametrization is not unique, but this merely means that different core models may generate the same semiparametric transformation model.

Example 3.1. Half-logistic and half-normal scale regression model. The assumption that the conditional distribution of a failure time T given a covariate Z has cumulative hazard function $H(t|z) = A_0(\tilde{\Gamma}(t)\exp[\theta^T z])$, for some unknown increasing function $\tilde{\Gamma}$ (model I), is clearly equivalent to the assumption that this cumulative hazard function is of the form $H(t|z) = A_0(A_0^{-1}(\Gamma(t))\exp[\theta^T z])$, for some unknown increasing function Γ (model II). The corresponding core models have hazard rates

$$\text{model I:} \qquad \alpha(x,\theta,z) = e^{\theta^T z}\alpha_0(xe^{\theta^T z}) \tag{3.1}$$

and

$$\text{model II:} \quad \alpha(x,\theta,z) = e^{\theta^T z}\frac{\alpha_0(A_0^{-1}(x)e^{\theta^T z})}{\alpha_0(A_0^{-1}(x))}, \tag{3.2}$$

respectively. In the case of the core model I, Condition 2.1(ii) is satisfied if the covariates are bounded, θ varies over a bounded neighbourhood of the true parameter and α_0 is a hazard rate that is bounded and bounded away from 0. An example is

provided by the half-logistic transformation model with $\alpha_0(x) = 1/2 + \tanh(x/2)$. This is a bounded increasing function from 1/2 to 1.

Next let us consider the choice of the half-normal transformation model. The half-normal distribution has survival function $F_0(x) = 2(1 - \Phi(x))$, where Φ is the standard normal distribution function. The hazard rate is given by

$$\alpha_0(x) = x + \frac{\int_x^\infty F_0(u)du}{F_0(x)}.$$

The second term represents the residual mean of the half normal distribution, and we have $\alpha_0(x) = x + \ell_0'(x)$. The function α_0 is increasing and unbounded so that the Condition 2.1(ii) fails to be satisfied by hazard rates (3.1). On the other hand the reparameterized transformation model II has hazard rates

$$\alpha(x, \theta, z) = e^{\theta^T z} \frac{A_0^{-1}(x)e^{\theta^T z} + \ell_0'(e^{\theta^T z} A_0^{-1}(x))}{A_0^{-1}(x) + \ell_0'(A_0^{-1}(x))}.$$

It can be shown that the right side satisfies $\exp(\theta^T z) \le \alpha(x, \theta, z) \le \exp(2\theta^T z) + \exp(\theta^T z)$ for $\exp(\theta^T z) > 1$, and $\exp(2\theta^T z)(1+\exp(\theta^T z))^{-1} \le \alpha(x, \theta, z) \le \exp(\theta^T z)$ for $\exp(\theta^T z) \le 1$. These inequalities are used to verify that the hazard rates of the core model II satisfy the remaining conditions 2.1 (ii).

Condition 2.1 assumes that the support of the distribution of the core model corresponds to the whole positive half-line and thus it has a support independent of the unknown parameter. The next example deals with the situation in which this support may depend on the unknown parameter.

Example 3.2. The gamma frailty model [14, 28] has cumulative hazard function

$$\begin{aligned} G(x, \theta|z) &= \frac{1}{\eta} \log[1 + \eta x e^{\beta^T z}], \quad \theta = (\eta, \beta), \eta > 0, \\ &= x e^{\beta^T z}, \qquad\qquad \eta = 0, \\ &= \frac{1}{\eta} \log[1 + \eta x e^{\beta^T z}], \quad \text{for} \quad \eta < 0 \quad \text{and} \quad -1 < \eta e^{\beta^T z} x \le 0. \end{aligned}$$

The right-hand side can be recognized as inverse cumulative hazard rate of Gompertz distribution.

For $\eta < 0$ the model is not invariant with respect to the group of strictly increasing transformations of R_+ onto itself. The unknown transformation Γ must satisfy the constraint $-1 < \eta \exp(\beta^T z)\Gamma(t) \le 0$ for μ a.e. z. Thus its range is bounded and depends on (η, β) and the covariates. Clearly, in this case the transformation model, assuming that the function Γ does not depend on covariates and parameters does not make any sense. When specialized to the transformation $\Gamma(t) = t$, the model is also not regular. For example, for $\eta = -1$ the cumulative hazard function is the same as that of the uniform distribution on the interval $[0, \exp(-\beta^T Z)]$. Similarly to the uniform distribution without covariates, the rate of convergence of the estimates of the regression coefficient is n rather than $\sqrt{n}$. For other choices of the $\tilde{\eta} = -\eta$ parameter, the Hellinger distance between densities corresponding to parameters β_1 and β_2 is determined by the magnitude of

$$E_Z 1(h^T Z > 0)[1 - \tilde{\eta} \exp(-h^T Z)]^{1/\tilde{\eta}} + E_Z 1(h^T Z < 0)[1 - \tilde{\eta} \exp(h^T Z)]^{1/\tilde{\eta}},$$

where $h = \beta_2 - \beta_1$. After expanding the exponents, this difference is of order $O(E_Z|hZ|^{1/\tilde{\eta}})$ so that for $\tilde{\eta} \le 1/2$ the model is regular, and irregular for $\tilde{\eta} > 1/2$.

For $\eta \geq 0$, the model is Hellinger differentiable both in the presence of covariates and in the absence of them ($\beta = 0$). The densities are supported on the whole positive half-line. The hazard rates are given by $g(x, \theta|z) = \exp(\beta^T z)\ [1 + \eta \exp(\beta^T z)x]^{-1}$. These are decreasing functions decaying to zero as $x \uparrow \infty$. Using Gompertz cumulative hazard function $G_\eta^{-1}(x) = \eta^{-1}[e^{\eta x} - 1]$ to reparametrize the model, we get $A(x, \theta|z) = G(G_\eta^{-1}(x), \theta|z) = \eta^{-1} \log[1+(e^{\eta x}-1) \exp(\beta^T z)]$. The hazard rate of this model is given by $\alpha(x, \theta|z) = \exp(\beta^T z+\eta x)[1+(e^{\eta x}-1) \exp(\beta^T z)]^{-1}$. Pointwise in β, this function is bounded between $\max\{\exp(e\beta^T Z), 1\}$ and from below by $\min\{\exp(\beta^T Z), 1\}$. The bounds are uniform for all $\eta \in [0, \infty)$ and the reparametrization preserves regularity of the model.

Note that the original core model has the property that for each parameter $\eta, \eta \geq 0$ it describes a distribution with different shape and upper tail behaviour. As a result of this, in the case of transformation model, the unknown function Γ is confounded by the parameter η. For example, at $\eta = 0$, the unknown transformation Γ represents a cumulative hazard function whereas at $\eta = 1$, it represents an odds ratio function. For any continuous variable X having a nondefective distribution, we have $E\Gamma(X) = 1$, if Γ is a cumulative hazard function, and $E\Gamma(X) = \infty$, if Γ is an odds ratio function. Since an odds ratio function diverges to infinity at a much faster rate than a cumulative hazard function, these are clearly very different parameters.

The preceding entails that when $\eta, \eta \geq 0$, is unknown we are led to a constrained optimization problem and our results fail to apply. Since the parameter η controls the shape and growth-rate of the transformation, it is not clear why this parameter could be identifiable based on rank statistics instead of order statistics. But if omission of constraints is permissible, then results of the previous section apply so long as the true regression coefficient satisfies $\beta_0 \neq 0$ and there exists a preliminary $\sqrt{n}$-consistent estimator of θ. At $\beta_0 = 0$, the parameter η is not identifiable based on ranks, if the unknown transformation is only assumed to be continuous and completely specified. We do not know if such initial estimators exist, and rank invariance arguments used in [14] suggest that the parameter η is not identifiable based on rank statistics because the models assuming that the cumulative hazard function is of the form $\eta^{-1} \log[1 + c\eta \exp(\beta^T z)\Gamma(t)]$ and $\eta^{-1} \log[1 + \exp(\beta^T z)\Gamma(t)], c > 0, \eta > 0$ all represent the same transformation model corresponding to log-Burr core model with different scale parameter c. Because this scale parameter is not identifiable based on ranks, the restriction $c = 1$ does not imply, that η may be identifiable based on rank statistics.

The difficulties arising in analysis of the gamma frailty with fixed frailty parameter disappear if we assume that the frailty parameter η depends on covariates. One possible choice corresponds to the assumption that the frailty parameter is of the form $\eta(z) = \exp \xi^T z$. The corresponding cumulative hazard function is given by $\exp[-\xi^T z] \log[1 + \exp(\xi^T z + \beta^T z)\Gamma(t)]$. This is a frailty model assuming that conditionally on Z and an unobserved frailty variable U, the failure time T follows a proportional hazards model with cumulative hazard function $U\Gamma(t) \exp(\beta^T Z)$, and conditionally on Z, the frailty variable U has gamma distribution with shape and scale parameter equal to $\exp(\xi^T z)$.

Example 3.3. Linear hazard model. The core model has hazard rate $h(x, \theta|z) = a_\theta(z) + x b_\theta(z)$ where $a_\theta(z), b_\theta(z)$ are nonnegative functions of the covariates dependent on a Euclidean parameter θ. The cumulative hazard function is equal to $H(t|z) = a_\theta(z)t + b_\theta(z)t^2/2$. Note that the shape of the density depends on the parameters a and b: it may correspond to both a decreasing and a non-monotone

function.

Suppose that $b_\theta(z) > 0, a_\theta(z) > 0$. To reparametrize the model we use $G^{-1}(x) = [(1+2x)^{1/2}-1]$. The reparametrized model has cumulative hazard function $A(x,\theta|z) = H(G^{-1}(x),\theta|z)$ with hazard rate $\alpha(x,\theta,z) = a_\theta(z)(1+2x)^{-1/2} + b_\theta(z)[1-(1+2x)^{-1/2}]$. The hazard rates are decreasing in x if $a_\theta(z) > b_\theta(z)$, constant in x if $a_\theta(z) = b_\theta(z)$ and bounded increasing if $a_\theta(z) < b_\theta(z)$. Pointwise in z the hazard rates are bounded from above by $\max\{a_\theta(z), b_\theta(z)\}$ and from below by $\min\{a_\theta(z), b_\theta(z)\}$. Thus our regularity conditions are satisfied, so long as in some neighbourhood of the true parameter θ_0 these maxima and minima stay bounded and bounded away from 0 and the functions a_θ, b_θ satisfy appropriate differentiability conditions. Finally, a sufficient condition for identifiability of parameters is that at a known reference point z_0 in the support of covariates, we have $a_\theta(z_0) = 1 = b_\theta(z_0)$, $\theta \in \Theta$ and

$$[a_\theta(z) = a_{\theta'}(z) \quad \text{and} \quad b_\theta(z) = b_{\theta'}(z) \quad \mu \text{ a.e. } z] \Rightarrow \theta = \theta'.$$

Returning to the original linear hazard model, we have excluded the boundary region $a_\theta(z) = 0$ or $b_\theta(z) = 0$. These boundary regions lead to lack of identifiability. For example,

$$\begin{aligned} &\text{model 1:} \qquad a_\theta(z) = 0 \quad \mu \text{ a.e. } z, \\ &\text{model 2:} \qquad b_\theta(z) = 0 \quad \mu \text{ a.e. } z, \\ &\text{model 3:} \qquad a_\theta(z) = cb_\theta(z) \quad \mu \text{ a.e. } z, \end{aligned}$$

where $c > 0$ is an arbitrary constant, represent the same proportional hazards model. The reparametrized model does not include the first two models, but, depending on the choice of the parameter θ, it may include the third model (with $c = 1$).

Example 3.4. Half-t and polynomial scale regression models. In this example we assume that the core model has cumulative hazard $A_0(x\exp[\theta^T z])$ for some known function A_0 with hazard rate α_0. Suppose that $c_1 \le \exp(\theta^T z) \le c_2$ for μ a.e. z.

For fixed $\xi \ge -1$ and $\eta \ge 0$, let G^{-1} be the inverse cumulative hazard function corresponding to the hazard rate $g(x) = [1+\eta x]^\xi$. If α_0/g is a function locally bounded and locally bounded away from zero such that $\lim_{x\uparrow\infty} \alpha_0(x)/g(x) = c$ for a finite positive constant c, then for any $\varepsilon \in (0,c)$ there exist constants $0 < m_1(\varepsilon) < m_2(\varepsilon) < \infty$, such that the hazard rate of $A_0(G^{-1}(x)\exp[\theta^T z])$ is bounded between $m_1(\varepsilon)$ and $m_2(\varepsilon)$. Indeed, using $c_1 \le \exp(\theta^T z) \le c_2$ and monotonicity properties of the function $g(x)$, we can find finite positive constants b_1, b_2 such that $b_1 \le e^{\theta^T z} g(x\exp[\theta^T z])/g(x) \le b_2$ for μ a.e. z and $x \ge 0$. The claim follows by setting $m_1(\varepsilon) = b_1 \max(c-\varepsilon, k^{-1})$ and $m_2(\varepsilon) = b_2 \min(c+\varepsilon, k)$, where $k = k(d), k > 0$ and $d > 0$ are such that $c-\varepsilon \le \alpha_0(x)/g(x) \le c+\varepsilon$ for $x > d$, and $k^{-1} \le \alpha_0(x)/g(x) \le k$, for $x \le d$.

In the case of half-logistic distribution, we choose $g(x) \equiv 1$. The function $g(x) = 1+x$ applies to the half-normal scale regression, while the choice $g(x) = (1+n^{-1}x)^{-1}$ applies to the half-t_n scale regression model. Of course in the case of gamma, inverse Gaussian frailty models (with fixed frailty parameters) and linear hazard model the choice of the $g(x)$ function is obvious.

In the case of polynomial hazards $\alpha_0(x) = 1 + \sum_{p=1}^m a_p x^p, m > 1$, where a_p are fixed nonnegative coefficients and $a_m > 0$, we choose $g(x) = [1+a_m x]^m$. Note however, that polynomial hazards may be also well defined when some of the coefficients a_p are negative. We do not know under what conditions polynomial hazards

define regular parametric models, but we expect that in such models parameters are estimated subject to added constraints in both parametric and semiparametric setting. Evidently, our results do not apply to such complicated problems.

The choice of $g(x) = [1 + \eta x]^{\xi}, \xi < -1$ was excluded in this example because it forms a defective hazard rate. Gronwall's inequalities in [3] show that hazard rates of the form $\exp(\theta^T z)[1 + x\exp(\theta^T z)]^{\xi}, \xi < -1$, lead to a Volterra equation whose solution is on a finite interval dependent on θ which may be strictly contained in the support of the withdrawal time distribution. Our results do not apply to this setting.

4. Discussion

In Section 2 we discussed properties of the estimate of the unknown transformation under no special regularity on the model representing the "true" distribution P of the data. Examples of Section 3 show that the class of transformation models to which these results apply is quite large and allows hazards of core models to have a variety of shapes.

To estimate the unknown Euclidean θ parameter, we made the additional assumption that the failure and censoring times are conditionally independent given the covariates and the failure times follow the transformation model. These conditions are sufficient to ensure that the score process is asymptotically unbiased, and the solution to the score equation forms a consistent estimate of the "true" parameter θ_0. However, only first two moment characteristics of certain stochastic integrals are used for this purpose in Section 7, so that the results may also be valid under different assumptions on the true distribution P.

We also showed that the class of M-estimators includes a special choice corresponding to an estimate whose asymptotic variance is equal to the inverse of the asymptotic variance of the score function $\sqrt{n}U_n(\theta_0)$. In [17] we show that this estimate is asymptotically efficient. Therein we discuss alternative ad hoc estimators of the unknown transformation and consider a larger class of M-estimators, allowing to adjust common inefficient estimates of the θ parameter to efficient one-step MLE estimates. Note that asymptotic variance of an M-estimator is usually of a "sandwich" form : As. var $\sqrt{n}(\hat{\theta} - \theta_0) = A^{-1}(\theta_0)$ As. var$\sqrt{n}U(\theta_0)[A^T(\theta_0)]^{-1}$, where $A(\theta)$ is the limit in probability of the derivative of $U_n(\theta)$ with respect to θ. However, it is quite common that estimators derived from conditional likelihoods of type (2.6) satisfy $A(\theta_0) =$ As. var $\sqrt{n}U_n(\theta_0)$ but are inefficient, so that results of Proposition 2.3 do not imply asymptotic efficiency of the corresponding estimate of the parameter θ.

The proofs of Propositions 2.1 and 2.2 are based on empirical and U-process techniques and are given in Sections 6 and 7. The next section collects some auxiliary results. The proof of consistency and weak convergence of the estimate of the unknown transformation relies also on Gronwall's inequalities collected in Section 9. The proof of Proposition 2.3 uses Fredholm determinant formula for resolvents of linear integral equations [25].

5. Some auxiliary results

We denote by $P_n = n^{-1}\sum_{i=1}^n \varepsilon_{X_i,\delta_i,Z_i}$ the empirical measure corresponding to a sequence of n iid observations (X_i, δ_i, Z_i) representing withdrawal times, censoring indicators and covariates. Set $N_{\cdot}(t) = n^{-1}\sum_{i=1}^n 1(X_i \le t, \delta_i = 1)$, $Y_{\cdot}(t) =$

$n^{-1}\sum_{i=1}^n 1(X_i \geq t)$ and Further, let $\|\cdot\|$ be the supremum norm in the set $\ell^\infty([0,\tau]\times\Theta)$, and let $\|\cdot\|_\infty$ be the supremum norm $\ell^\infty([0,\tau])$. We assume that the point τ satisfies Condition 2.0(iii).

Define

$$\begin{aligned}
R_n(t,\theta) &= \int_0^t \frac{N_\cdot(du)}{S(\Gamma_\theta(u-),\theta,u)} - \int_0^t \frac{EN(du)}{s(\Gamma_\theta(u-),\theta,u)},\\
R_{pn}(t,\theta) &= \int_0^t h(\Gamma_\theta(u-),\theta,u)[N_\cdot - EN](du), \quad p=5,6,\\
R_{pn}(t,\theta) &= \int_0^t H_n(\Gamma_\theta(u-),\theta,u)N(du)\\
&\quad - \int_0^t h(\Gamma_\theta(u-),\theta,u)EN(du), \quad p=7,8,\\
R_{9n}(t,\theta) &= \int_{[0,t)} EN(du)\Big|\int_{(u,t]} \mathcal{P}_\theta(u,w)R_{5n}(dw,\theta)\Big|,\\
R_{10n}(t,\theta) &= \int_0^t \sqrt{n}R_n(u-,\theta)R_{5n}(du,\theta),\\
R_{pn}(t,\theta) &= \int_0^t H_n(\Gamma_{n\theta}(u-),\theta,u)N_\cdot(du)\\
&\quad - \int_0^t h(\Gamma_\theta(u-),\theta,u)EN(du), \quad p=11,12,\\
B_{pn}(t,\theta) &= \int_0^t F_{pn}(u,\theta)R_n(du,\theta), \quad p=1,2,
\end{aligned}$$

where $\mathcal{P}_\theta(u,w)$ is given by (2.3). In addition, $H_n = K_n'$ for $p=7$ or $p=11$, $H_n = \dot K_n$ for $p=8$ or $p=12$, $h=k'$ for $p=5,7$ or $p=11$, and $h=\dot k$ for $p=6,8$ or 12. Here $k' = -[s'/s^2]$, $\dot k = -[\dot s/s^2]$, $K_n' = -[S'/S^2]$, $\dot K_n = -[\dot S/S^2]$. Further, set $F_{1n}(u,\theta) = [\dot S - \bar e S](\Gamma_\theta(u),\theta,u)$ and $F_{2n}(u,\theta) = [S' - eS'](\Gamma_\theta(u),\theta,u)$.

Lemma 5.1. *Suppose that Conditions* 2.0 *and* 2.1 *are satisfied.*

(i) $\sqrt{n}R_n(t,\theta)$ *converges weakly in* $\ell^\infty([0,\tau]\times\Theta)$ *to a mean zero Gaussian process* R *whose covariance function is given below.*
(ii) $\|R_{pn}\| \to 0$ *a.s., for* $p=5,\ldots,12$.
(iii) $\sqrt{n}\|B_{pn}\| \to 0$ *a.s. for* $p=1,2$.
(iv) *The processes* $V_n(\Gamma_\theta(t-),\theta,t)$ *and* $V_n(\Gamma_\theta(t),\theta,t)$, *where* $V_n = S/s - 1$ *satisfy* $\|V_n\| = O(b_n)$ *a.s. In addition,* $\|V_n\| \to 0$ *a.s. for* $V_n = [S'-s']/s, [S''-s'']/s, [\dot S - \dot s]/s, [\ddot S - \ddot s]/s$ *and* $[\dot S' - \dot s']/s$.

Proof. The Volterra identity (2.2), which defines Γ_θ as a parameter dependent on P, is used in the foregoing to compute the asymptotic covariance function of the process R_{1n}. In Section 6 we show that the solution to the identity (2.2) is unique and, for some positive constants d_0, d_1, d_2, we have

$$\begin{aligned}
&\Gamma_\theta(t) \leq d_0 A_P(t), \qquad |\Gamma_\theta(t) - \Gamma_{\theta'}(t)| \leq |\theta-\theta'|d_1\exp[d_2A_P(t)],\\
&|\Gamma_\theta(t) - \Gamma_\theta(t')| \leq d_0|A_P(t) - A_P(t')| \qquad (5.1)\\
&\qquad\qquad \leq \frac{d_0}{E_P Y(\tau)} P(X \in (t\wedge t', t\vee t'], \delta = 1),
\end{aligned}$$

with similar inequalities holding for the left continuous version of $\Gamma_\theta = \Gamma_{\theta,P}$. Here $A_P(t)$ is the cumulative hazard function corresponding to observations (X,δ).

To show part (i), we use the quadratic expansion, similar to the expansion of the ordinary Aalen–Nelson estimator in [19]. We have $R_n = \sum_{j=1}^4 R_{jn}$,

$$\begin{aligned}
R_{1n}(t,\theta) &= \frac{1}{n}\sum_{i=1}^n \int_0^t \left[\frac{N_i(du)}{s(\Gamma_\theta(u-),\theta,u)} - \frac{S_i}{s^2}(\Gamma_\theta(u-),\theta,u)EN(du)\right] \\
&= \frac{1}{n}\sum_{i=1}^n R_{1n}^{(i)}(t,\theta), \\
R_{2n}(t,\theta) &= \frac{-1}{n^2}\sum_{i\neq j}\int_0^t \left(\frac{S_i - s}{s^2}\right)(\Gamma_\theta(u-),\theta,u)[N_j - EN_j](du), \\
R_{3n}(t,\theta) &= \frac{-1}{n^2}\sum_{i=1}^n\int_0^t \left(\frac{S_i - s}{s^2}\right)(\Gamma_\theta(u-),\theta,u)[N_i - EN_i](du), \\
R_{4n}(t,\theta) &= \int_0^t \left(\frac{S-s}{s}\right)^2(\Gamma_\theta(u-),\theta,u)\frac{N_\cdot(du)}{S(\Gamma_\theta(u-),\theta,u)},
\end{aligned}$$

where $S_i(\Gamma_\theta(u-),\theta,u) = Y_i(u)\alpha(\Gamma_\theta(u-),\theta,Z_i)$.

The term R_{3n} has expectation of order $O(n^{-1})$. Using Conditions 2.1, it is easy to verify that R_{2n} and $n[R_{3n} - ER_{3n}]$ form canonical U-processes of degree 2 and 1 over Euclidean classes of functions with square integrable envelopes. We have $\|R_{2n}\| = O(b_n^2)$ and $n\|R_{3n} - ER_{3n}\| = O(b_n)$ almost surely, by the law of iterated logarithm for canonical U processes [1]. The term R_{4n} can be bounded by $\|R_{4n}\| \le \|[S/s]-1\|^2 m_1^{-1}A_n(\tau)$. But for a point τ satisfying Condition 2.0(iii), we have $A_n(\tau) = A(\tau)+O(b_n)$ a.s. Therefore part (iv) below implies that $\sqrt{n}\|R_{4n}\| \to 0$ a.s.

The term R_{1n} decomposes into the sum $R_{1n} = R_{1n;1} - R_{1n;2}$, where

$$\begin{aligned}
R_{1n;1}(t,\theta) &= \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{N_i(du) - Y_i(u)A(du)}{s(\Gamma_\theta(u-),\theta,u)}, \\
R_{1n;2}(t,\theta) &= \int_0^t G(u,\theta)C_\theta(du)
\end{aligned}$$

and $G(t,\theta) = [S(\Gamma_\theta(u-),\theta,u) - s(\Gamma_\theta(u-),\theta,u)Y_\cdot(u)/EY(u)]$. The Volterra identity (2.2) implies

$$\begin{aligned}
&n\mathrm{cov}(R_{1n;1}(t,\theta), R_{1n;1}(t',\theta')) = \int_0^{t\wedge t'} \frac{[1-A(\Delta u)]\Gamma_\theta(du)}{s(\Gamma_{\theta'}(u-),\theta',u)}, \\
&n\mathrm{cov}(R_{1n;1}(t,\theta), R_{1n;2}(t',\theta')) \\
&\quad = \int_0^t\int_0^{u\wedge t'} E[\alpha(\Gamma_{\theta'}(v-),Z,\theta'|X=u,\delta=1]C_{\theta'}(dv)\Gamma_\theta(du) \\
&\qquad - \int_0^t\int_0^{u\wedge t'} E\alpha(\Gamma_{\theta'}(v-),Z,\theta'|X\ge u]]C_{\theta'}(dv)\Gamma_\theta(du), \\
&n\mathrm{cov}(R_{1n;2}(t,\theta), R_{1n;2}(t',\theta')) \\
&\quad = \int_0^t\int_0^{t'\wedge u} f(u,v,\theta,\theta')C_\theta(du)C_{\theta'}(dv) \\
&\qquad + \int_0^{t'}\int_0^{t\wedge v} f(v,u,\theta',\theta)C_\theta(du)C_{\theta'}(dv) \\
&\qquad - \int_0^{t\wedge t'} f(u,u,\theta,\theta')C_{\theta'}(\Delta u)C_\theta(du),
\end{aligned}$$

where $f(u,v,\theta,\theta') = EY(u)\mathrm{cov}(\alpha(\Gamma_\theta(u-),\theta,Z),\alpha(\Gamma_{\theta'}(v-),\theta',Z)|X \geq u)$. Using CLT and Cramer-Wold device, the finite dimensional distributions of $\sqrt{n}R_{1n}(t,\theta)$ converge in distribution to finite dimensional distributions of a Gaussian process. The process R_{1n} can be represented as $R_{1n}(t,\theta) = [P_n - P]h_{t,\theta}$, where $\mathcal{H} = \{h_{t,\theta}(x,d,z) : t \leq \tau, \theta \in \Theta\}$ is a class of functions such that each $h_{t,\theta}$ is a linear combination of 4 functions having a square integrable envelope and such that each is monotone with respect to t and Lipschitz continuous with respect to θ. This is a Euclidean class of functions [29] and $\{\sqrt{n}R_{1n}(t,\theta) : \theta \in \Theta, t \leq \tau\}$ converges weakly in $\ell^\infty([0,\tau]\times\Theta)$ to a tight Gaussian process. The process $\sqrt{n}R_{1n}(t,\theta)$ is asymptotically equicontinuous with respect to the variance semimetric ρ. The function ρ is continuous, except for discontinuity hyperplanes corresponding to a finite number of discontinuity points of EN. By the law of iterated logarithm [1], we also have $\|R_{1n}\| = O(b_n)$ a.s.

Remark 5.1. Under Condition 2.2, we have the identity

$$\begin{aligned} &n\mathrm{cov}(R_{1n;2}(t,\theta_0), R_{1n;2}(t',\theta_0)) \\ &\quad = \sum_{p=1}^{2} n\mathrm{cov}(R_{1n;p}(t,\theta_0, R_{1n;3-p}(t',\theta_0)) \\ &\qquad - \int_{[0,t\wedge t']} EY(u)var(\alpha(\Gamma_{\theta_0}(u-)|X \geq u)C_{\theta_0}(\Delta u)C_{\theta_0}(du). \end{aligned}$$

Here θ_0 is the true parameter of the transformation model. Therefore, using the assumption of continuity of the EN function and adding up all terms, $n\mathrm{cov}(R_{1n}(t,\theta_0), R_{1n}(t',\theta_0)) = n\mathrm{cov}(R_{1n;1}(t,\theta_0), R_{1n;1}(t',\theta_0)) = C_{\theta_0}(t\wedge t')$.

Next set $b_\theta(u) = h(\Gamma_\theta(u-),\theta,u)$, $h = k'$ or $h = \dot{h}$. Then $\int_0^t b_\theta(u)N_\cdot(du) = P_n f_{t,\theta}$, where $f_{t,\theta} = 1(X \leq t, \delta = 1)h(\Gamma_\theta(X\wedge\tau-),\theta, X\wedge\tau-)$. The conditions 2.1 and the inequalities (5.1) imply that the class of functions $\{f_{t,\theta} : t \leq \tau, \theta \in \Theta\}$ is Euclidean for a bounded envelope, for it forms a product of a VC-subgraph class and a class of Lipschitz continuous functions with a bounded envelope. The almost sure convergence of the terms $R_{pn}, p = 5, 6$ follows from Glivenko–Cantelli theorem [29].

Next, set $b_\theta(u) = k'(\Gamma_\theta(u-),\theta,u)$ for short. Using Fubini theorem and $|\mathcal{P}_\theta(u,w)| \leq \exp[\int_u^w |b_\theta(s)|EN(ds)]$, we obtain

$$\begin{aligned} R_{9n}(t,\theta) \;&\leq\; \int_{(0,t)} EN(du)|R_{5n}(t,\theta) - R_{5n}(u,\theta)| \\ &\quad + \int_{(0,t)} EN(du)|\int_{(u,t]} \mathcal{P}_\theta(u,s-)b_\theta(s)EN(ds)[R_{5n}(t,\theta) - R_{5n}(s,\theta)]| \\ &\leq\; 2\|R_{5n}\| \int_{[0,t)} EN(du)[1 + \int_{(u,t]} |\mathcal{P}(u,w-)||b_\theta(w)|EN(dw)] \\ &\leq\; 2\|R_{5n}\| \int_0^\tau EN(du)\exp[\int_{(u,\tau]} |b_\theta|(s)EN(ds)] \to 0 \quad \text{a.s.} \end{aligned}$$

uniformly in $t \leq \tau, \theta \in \Theta$.

Further, we have $R_{10n}(t,\theta) = \sqrt{n}\sum_{p=1}^{4} R_{10n;p}(t,\theta)$, where

$$\begin{aligned} R_{10n;p}(t;\theta) &= \int_0^t R_{pn}(u-,\theta)R_{5n}(du;\theta) = \\ &= \int_0^t R_{pn}(u-;\theta)k'(\Gamma_\theta(u-),\theta,u)[N_\cdot - EN](du). \end{aligned}$$

We have $\|\sqrt{n}R_{10n;p}\| = O(1)\sup_{\theta,t}|\sqrt{n}R_{pn}(u-,\theta)| \to 0$ a.s. for $p = 2,3,4$. Moreover, $\sqrt{n}R_{10n;1}(t;\theta) = \sqrt{n}R_{10n;11}(t;\theta) + \sqrt{n}R_{10n;12}(t;\theta)$, where $R_{10n;11}$ is equal to

$$n^{-2}\sum_{i\neq j}\int_0^t R_{1n}^{(i)}(u-,\theta)k'(\Gamma_\theta(u-),\theta,u)[N_j - EN_j](du),$$

while $R_{10n;12}(t,\theta)$ is the same sum taken over indices $i = j$. These are U-processes over Euclidean classes of functions with square integrable envelopes. By the law of iterated logarithm [1], we have $\|R_{10n;11}\| = O(b_n^2)$ and $n\|R_{10n;12} - ER_{10n;12}\| = O(b_n)$ a.s. We also have $ER_{10n;12}(t,\theta) = O(1/n)$ uniformly in $\theta \in \Theta$, and $t \le \tau$.

The analysis of terms B_{1n} and B_{2n} is quite similar. Suppose that $\ell'(x,\theta) \not\equiv 0$. We have $B_{2n} = \sum_{p=1}^{4} B_{2n;p}$, where in the term $B_{2n;p}$ integration is with respect to R_{np}. For $p = 1$, we obtain $B_{2n;1} = B_{2n;11} + B_{2n;12}$, where

$$B_{2n;11}(t,\theta) = \frac{1}{n^2}\int_0^t \sum_{i\neq j}[S_i' - eS_i](\Gamma_\theta(u),\theta,u)R_{1n}^{(j)}(du,\theta),$$

whereas the term $B_{2n;12}$ represents the same sum taken over indices $i = j$. These are U-processes over Euclidean classes of functions with square integrable envelopes. By the law of iterated logarithm [1], we have $\|B_{2n;11}\| = O(b_n^2)$ and $n\|B_{2n;12} - EB_{2n;12}\| = O(b_n)$ a.s. We also have $EB_{2n;12}(t,\theta) = O(1/n)$ uniformly in $\theta \in \Theta$, and $t \le \tau$. Thus $\sqrt{n}\|B_{2n;1}\| \to 0$ a.s. A similar analysis, leading to U-statistics of degree 1, 2, 3 can be applied to the integrals $\sqrt{n}B_{2n;p}(t,\theta)$, $p = 2,3$. On the other hand, assumption 2.1 implies that for $p = 4$, we have the bound

$$\begin{aligned} |B_{2n;4}(t,\theta)| &\le 2\int_0^\tau \psi(A_2(u-))\frac{(S-s)^2}{s^2}EN(du) \\ &\le O(1)\int_0^\tau \frac{(S-s)^2}{s^2}EN(du), \end{aligned}$$

where, under Condition 2.1, the function ψ bounding ℓ' is either a constant c or a bounded decreasing function (thus bounded by some c). The right-hand side can further be expanded to verify that $\|\sqrt{n}B_{2n;4}\| \to 0$ a.s. Alternatively, we can use part (iv).

A similar expansion can also be applied to show that $\|R_{7n}\| \to 0$ a.s. Alternatively we have, $|R_{7n}(t,\theta)| \le \int_0^\tau |K_n' - k'|(\Gamma_\theta(u-),\theta,u)N_\cdot(du) + |R_{5n}(t,\theta)|$ and by part (iv), we have uniform almost sure convergence of the term R_{7n}. We also have $|R_{11n} - R_{7n}|(t,\theta) \le \int_0^\tau O(|\Gamma_{n\theta} - \Gamma_\theta|)(u)N_\cdot(du)$ a.s., so that part (i) implies $\|R_{11n}\| \to 0$ a.s. The terms R_{8n} and R_{12n} can be handled analogously.

Next, $[S/s](\Gamma_\theta(t-),\theta,t) = P_n f_{\theta,t}$, where

$$f_{\theta,t}(x,\delta,z) = 1(x \ge t)\frac{\alpha(\Gamma_\theta(t-),\theta,z)}{EY(u)\alpha(\Gamma_\theta(t-),\theta,Z)} = 1(x \ge t)g_{\theta,t}(z).$$

Suppose that Condition 2.1 is satisfied by a decreasing function ψ and an increasing function ψ_1. The inequalities (5.1) and Condition 2.1, imply that $|g_{\theta,t}(Z)| \le$

$m_2[m_1 E_P Y(\tau)]^{-1}$, $|g_{\theta,t}(Z) - g_{\theta',t}(Z)| \le |\theta - \theta'| h_1(\tau)$, $|g_{\theta,t}(Z) - g_{\theta,t'}(Z)| \le [P(X \in [t \wedge t', t \vee t')) + P(X \in (t \wedge t', t \vee t'], \delta = 1)] h_2(\tau)$, where

$$\begin{aligned} h_1(\tau) &= 2m_2[m_1 E_P Y(\tau)]^{-1}[\psi_1(d_0 A_P(\tau)) + \psi(0) d_1 \exp[d_2 A_P(\tau)], \\ h_2(\tau) &= m_2[m_1 E_P Y(\tau)]^{-2}[m_2 + 2\psi(0)]. \end{aligned}$$

Setting $h(\tau) = \max[h_1(\tau), h_2(\tau), m_2(m_1 E_P Y(\tau))^{-1}]$, it is easy to verify that the class of functions $\{f_{\theta,t}(x, \delta, z)/h(\tau) : \theta \in \Theta, t \le \tau\}$ is Euclidean for a bounded envelope. The law of iterated logarithm for empirical processes over Euclidean classes of functions [1] implies therefore that part (iii) is satisfied by the process $V = S/s - 1$. For the remaining choices of the V processes the proof is analogous and follows from the Glivenko–Cantelli theorem for Euclidean classes of functions [29]. □

6. Proof of Proposition 2.1

6.1. Part (i)

For $P \in \mathcal{P}$, let $A(t) = A_P(t)$ be given by (2.1) and let $\tau_0 = \sup\{t : E_P Y(t) > 0\}$. The condition 2.1 (ii) assumes that there exist constants $m_1 < m_2$ such that the hazard rate $\alpha(x, \theta|z)$ is bounded from below by m_1 and from above by m_2. Put $A_1 = m_1^{-1} A(t)$ and $A_2(t) = m_2^{-1}(t)$. Then $A_2 \le A_1$. Further, Condition 2(iii) assumes that the function $\ell(x, \theta, z) = \log \alpha(x, \theta, z)$ has a derivative $\ell'(x, \theta, z)$ with respect to x satisfying $|\ell'(x, \theta, z)| \le \psi(x)$ for some bounded decreasing function. Suppose that $\psi \le c$ and define $\rho(t) = \max(c, 1) A_1(t)$. Finally, the derivative $\dot{\ell}(x, \theta, z)$ satisfies $|\dot{\ell}(x, \theta, z)| \le \psi_1(x)$ for some bounded function or a function that is continuous strictly increasing, bounded at origin and satisfying $\int_0^\infty \psi_1(x)^2 e^{-x} dx < \infty$. Let

$$d = \int_0^\infty \psi_1(x) e^{-x} dx < \infty.$$

In the inequalities (5.1) of Lemma 5.1 we take $d_0 = m_1^{-1}, d_1 = \max(1, c)$ and $d_2 = d$.

Let $\mathcal{T} = [0, \tau_0]$ if τ_0 is a discontinuity point of the survival function $E_P Y(t)$, and let $\mathcal{T} = [0, \tau_0)$, if τ_0 is a continuity point of this survival function. Consider the set of functions $\mathcal{X}(P) = \{g : g \text{ monotone increasing}, e^{-g} \in D(\mathcal{T}), g \ll E_P N, A_2 \le g \le A_1\}$. Since for each $g \in \mathcal{X}(\mathcal{P})$, the function e^{-g} is a subsurvival function satisfying $\exp[-A_1] \le \exp[-g] \le \exp[-A_2]$, we can consider $\mathcal{X}(P)$ as a subset of $D(\mathcal{T})$, endowed with supremum norm. Next, for $\tau < \tau_0$, let $\mathcal{X}(P, \tau) \subset D([0, \tau])$ consist of functions $g \in \mathcal{X}(P)$ restricted to the interval $[0, \tau]$. For fixed $\theta \in \Theta$ and $g \in \mathcal{X}(P, \tau)$, define

$$\Psi_\theta(g)(t) = \int_0^t [EY(u)\alpha(g(u-), \theta, Z)]^{-1} EN(du), \quad 0 \le t \le \tau.$$

Using bounds $A_1 \le g \le A_2$ it is easy to verify that for fixed $\theta \in \Theta$, Ψ_θ maps $\mathcal{X}(P, \tau)$ into itself. Since $A_p(0-) = 0$, we have $g(0-) = 0$ and $\Psi_\theta(g)(0-) = 0$ as well.

Consider the equation $\Psi_\theta(g) = g, g(0-) = 0$. Using Helly selection theorem, it is easy to verify that for fixed $\theta \in \Theta$, the operator Ψ_θ maps $\mathcal{X}(P, \tau)$ into itself, is continuous (with respect to g) and has compact range. Since $\mathcal{X}(P, \tau)$ forms a bounded, closed convex set of functions, Schauder's fixed point theorem implies that Ψ_θ has a fixed point in $\mathcal{X}(P, \tau)$.

To show uniqueness of the solution and its continuity with respect to θ, we consider first the case of continuous $EN(t)$ function. Then $\mathcal{X}(P,\tau) \subset C([0,\tau])$. Define a norm in $C([0,\tau])$ by setting $\|x\|_\rho^\tau = \sup_{t\le\tau} e^{-\rho(t)}|x(t)|$. Then $\|\cdot\|_\rho^\tau$ is equivalent to the sup norm in $C([0,\tau])$. For $g, g' \in \mathcal{X}(P,\tau)$ and $\theta \in \Theta$, we have

$$\begin{aligned}|\Psi_\theta(g) - \Psi_\theta(g')|(t) &\le \int_0^t |g-g'|(u)\psi(A_2(u))A_1(du)\\ &\le \int_0^t |g-g'|(u)\rho(du) \le \|g-g'\|_\rho^\tau \int_0^t e^{\rho(u)}\rho(du)\\ &\le \|g-g'\|_\rho^\tau e^{\rho(t)}(1-e^{-\rho(\tau)})\end{aligned}$$

and hence $\|\Psi_\theta(g) - \Psi_{\theta'}(g')\|_\rho^\tau \le \|g-g'\|_\rho^\tau(1-e^{-\rho(\tau)})$. For any $g \in \mathcal{X}(P,\tau)$ and $\theta, \theta' \in \Theta$, we also have

$$\begin{aligned}|\Psi_\theta(g) - \Psi_{\theta'}(g)|(t) &\le |\theta-\theta'|\int_0^t \psi_1(g(u))A_1(du)\\ &\le |\theta-\theta'|\int_0^t \psi_1(\rho(u))\rho(du)\\ &\le |\theta-\theta'|e^{\rho(t)}\int_0^t \psi_1(\rho(u))e^{-\rho(u)}\rho(du) \le |\theta-\theta'|e^{\rho(t)}d,\end{aligned}$$

so that $\|\Psi_\theta(g) - \Psi_{\theta'}(g)\|_\rho^\tau \le |\theta-\theta'|d$. It follows that $\{\Psi_\theta : \theta \in \Theta\}$, restricted to $C[0,\tau]$), forms a family of continuously contracting mappings. Banach fixed point theorem for continuously contracting mappings [24] implies therefore that there exists a unique solution Γ_θ to the equation $\Phi_\theta(g)(t) = g(t)$ for $t \le \tau$, and this solution is continuous in θ. Since $A(0) = A(0-) = 0$, and the solution is bounded between two multiples of $A(t)$, we also have $\Gamma_\theta(0) = 0$.

Because $\|\cdot\|_\rho^\tau$ is equivalent to the supremum norm in $C[0,\tau]$, we have that for fixed $\tau < \tau_0$, there exists a unique (in sup norm) solution to the equation, and the solution is continuous with respect to θ. It remains to consider the behaviour of these functions at τ_0. Fix $\theta \in \Theta$ again. If $A(\tau_0) < \infty$, then Γ_θ is unique on the whole interval $[0,\tau_0]$ (the preceding argument can be applied to the interval $[0,\tau_0]$). So let us consider the case of $A(\tau) \uparrow \infty$ as $\tau \uparrow \tau_0$. If $\tau^{(1)} < \tau^{(2)} < \tau_0$, then $\mathcal{X}(P,\tau^{(1)}) \subset \mathcal{X}(P,\tau^{(2)})$. Let $\Gamma_\theta^{(p)} \in \mathcal{X}(P,\tau^{(p)}), p = 1,2$ be the solutions obtained on intervals $[0,\tau^{(1)}]$ and $[0,\tau^{(2)}]$, respectively. Then the function $\Gamma_\theta^{(2)}$ satisfies $\Gamma_\theta^{(2)}(t) = \Gamma_\theta^{(1)}(t)$ for $t \in [0,\tau^{(1)}]$. If $\tau^{(n)} \uparrow \tau_0$, then the inequalities $\exp[-A_1(\tau^{(n)})] \le \exp[-\Gamma_\theta^{(n)}(\tau^{(n)})] \le \exp[-A_2(\tau^{(n)})]$ imply $\Gamma_\theta^{(n)}(\tau^{(n)}) \uparrow \infty$. Since this holds for any such sequence $\tau^{(n)}$, there exists a unique locally bounded solution to the equation on the interval $\mathcal{T} = [0,\tau_0)$.

Next let us consider the case of discrete $EN(t)$ with a finite number of discontinuity points. In this case $A_P(\tau_0)$ is bounded and satisfies $A_p(0-) = 0$. Fix θ. Using induction on jumps, it is easy to verify that for any $g \in \mathcal{X}(P,\tau_0)$, we have $\Psi_\theta(g) \in \mathcal{X}(P,\tau_0)$, and for any $g, g' \in \mathcal{X}(P,\tau_0)$, we also have $\Psi_\theta(g) = \Psi_\theta(g')$. Hence $\Psi_\theta^2(g) = \Psi_\theta(g)$. Alternatively, that the solution Γ_θ to the equation $\Psi_\theta(g) = g$, $g(0-) = 0$ is uniquely defined follows also from the recurrent formula $\Gamma_\theta(t) = \Gamma_\theta(t-) + EN(\Delta t)[EY(t)\alpha(\Gamma_\theta(t-),\theta,Z)]^{-1}$, $\Gamma_\theta(0-) = 0$.

For any $g \in \mathcal{X}(P, \tau_0)$ and $\theta, \theta' \in \Theta$, we also have

$$\begin{aligned} |\Psi_\theta(g) - \Psi_{\theta'}(g)|(t-) &\le |\theta - \theta'| \int_{[0,t)} \psi_1(g(u-)) A_1(du) \\ &\le |\theta - \theta'| e^{\rho(t)} \int_{[0,t)} \psi_1(\rho(u-)) e^{-\rho(u)} \rho(du) \\ &\le |\theta - \theta'| e^{\rho(t-)} d. \end{aligned}$$

To see the last inequality, we define

$$\Psi_1(x) = \int_0^x e^{-y} \psi_1(y) dy.$$

Then $\Psi_1(\rho(t)) - \Psi_1(0) = \Sigma \rho(\Delta u) \psi_1(\rho(u^*)) \exp -\rho(u^*)$, where the sum extends over discontinuity points less than t, and $\rho(u^*)$ is between $\rho(u-)$ and $\rho(u)$. The right-hand side is bounded from below by the corresponding sum $\sum \rho(\Delta u) \psi_1(\rho(u-)) \exp[-\rho(u)]$, because $\psi_1(x)$ is increasing and $\exp(-x)$ is decreasing. Since $\Psi_\theta(g) = \Gamma_\theta$ for any θ, we have $\sup_{t \le \tau_0} e^{-\rho(t-)} |\Gamma_\theta - \Gamma_{\theta'}|(t-) \le |\theta - \theta'| d$.

Finally, for both the continuous and discrete case, we have

$$\begin{aligned} |\Gamma_\theta - \Gamma_{\theta'}|(t) &\le |\Psi_\theta(\Gamma_\theta) - \Psi_\theta(\Gamma_{\theta'})|(t) + |\Psi_\theta(\Gamma_{\theta'}) - \Psi_{\theta'}(\Gamma_{\theta'})|(t) \le \\ &\le \int_0^t |\Gamma_\theta - \Gamma_{\theta'}|(u-) \rho(du) + |\theta - \theta'| \int_0^t \psi_1(\rho(u-)) \rho(du), \end{aligned}$$

and Gronwall's inequality (Section 9) yields

$$|\Gamma_\theta - \Gamma_{\theta'}|(t) \le |\theta - \theta'| e^{\rho(t)} \int_{(0,t]} \psi_1(\rho(u-)) e^{-\rho(u-)} \rho(du) \le d |\theta - \theta'| e^{\rho(t)}.$$

Hence $\sup_{t \le \tau} e^{-\rho(t)} |\Gamma_\theta - \Gamma_{\theta'}|(t) \le |\theta - \theta'| d$. In the continuous case this holds for any $\tau < \tau_0$, in the discrete case for any $\tau \le \tau_0$.

Remark 6.1. We have chosen the ρ function as equal to $\rho(t) = \max(c, 1) A_1$, where c is a constant bounding the function $\ell_i'(x, \theta)$. Under Condition 2.1, this function may also be bounded by a continuous decreasing function ψ. The proof, assuming that $\rho(t) = \int_0^t \psi(A_2(u-)) A_1(du)$ is quite similar. In the foregoing we consider the simpler choice, because in Proposition 2.2 we have assumed Condition 2.0(iii). Further, in the discrete case the assumption that the number of discontinuity points is finite is not needed but the derivations are longer.

To show consistency of the estimate $\Gamma_{n\theta}$, we assume now that the point τ satisfies Condition 2.0(iii). Let $A_n(t)$ be the Aalen–Nelson estimator and set $A_{pn} = m_p^{-1} A_n, p = 1, 2$. We have $A_{2n}(t) \le \Gamma_{n\theta}(t) \le A_{1n}(t)$ for all $\theta \in \Theta$ and $t \le \max(X_i, i = 1, \ldots, n)$. Setting $K_n(\Gamma_{n\theta}(u-), \theta, u) = S(\Gamma_{n\theta}(u-), \theta, u)^{-1}$, we have

$$\begin{aligned} \Gamma_{n\theta}(t) - \Gamma_\theta(t) &= R_n(t, \theta) \\ &\quad + \int_{(0,t]} [K_n(\Gamma_{n\theta}(u-), \theta, u) - K_n(\Gamma_\theta(u-), \theta, u)] \, N_\cdot(du). \end{aligned}$$

Hence $|\Gamma_{n\theta}(t) - \Gamma_\theta(t)| \le |R_n(t, \theta)| + \int_0^t |\Gamma_{n\theta} - \Gamma_\theta|(u-) \rho_n(du)$, where $\rho_n = \max(c, 1) A_{1n}$. Gronwall's inequality implies $\sup_{t,\theta} \exp[-\rho_n(t)] |\Gamma_{n\theta} - \Gamma_\theta|(t) \to 0$ a.s., where the supremum is over $\theta \in \Theta$ and $t \le \tau$. If τ_0 is a discontinuity point of the survival function $E_P Y(t)$ then this holds for $\tau = \tau_0$.

Next suppose that τ_0 is a continuity point of this survival function, and let $\mathcal{T} = [0, \tau_0)$. We have $\sup_{t\in\mathcal{T}} |\exp[-A_{pn}(t) - \exp[-A_p(t)| = o_P(1)$. In addition, for any $\tau < \tau_0$, we have $\exp[-A_{1n}(\tau)] \leq \exp[-\Gamma_{n\theta}(\tau)] \leq \exp[-A_{2n}(\tau)]$. Standard monotonicity arguments imply $\sup_{t\in\mathcal{T}} |\exp[-\Gamma_{n\theta}(t) - \exp[-\Gamma_\theta(t)| = o_P(1)$, because $\Gamma_\theta(\tau) \uparrow \infty$ as $\tau \uparrow \infty$.

6.2. Part (iii)

The process $\hat{W}(t,\theta) = \sqrt{n}[\Gamma_{n\theta} - \Gamma_\theta](t)$ satisfies

$$\hat{W}(t,\theta) = \sqrt{n}R_n(t,\theta) - \int_{[0,t]} \hat{W}(u-,\theta)N_\cdot(du)b^*_{n\theta}(u),$$

where

$$b^*_{n\theta}(u) = \left[\int_0^1 \left(S'/S^2\right)(\theta, \Gamma_\theta(u-) + \lambda[\Gamma_{n\theta} - \Gamma_\theta](u-), u)d\lambda\right].$$

Define

$$\tilde{W}(t,\theta) = \sqrt{n}R_n(t,\theta) - \int_0^t \tilde{W}(u-,\theta)b_\theta(u)EN(du),$$

where $b_\theta(u) = [s'/s^2](\Gamma_\theta(u), \theta, u)$. We have

$$\tilde{W}(t,\theta) = \sqrt{n}R_n(t,\theta) - \int_0^t \sqrt{n}R_n(u-,\theta)b_\theta(u)EN(du)\mathcal{P}_\theta(u,t)$$

and

$$\hat{W}(t,\theta) - \tilde{W}(t,\theta) = -\int_0^t [\hat{W} - \tilde{W}](u-,\theta)b^*_{n\theta}(u)N_\cdot(du) + \mathrm{rem}(t,\theta),$$

where

$$\mathrm{rem}(t,\theta) = -\int_{[0,t]} \tilde{W}(u-,\theta)[b^*_{n\theta}(u)N_\cdot(du) - b_\theta(u)EN(du)].$$

The remainder term is bounded by

$$\int_0^\tau |\tilde{W}(u-,\theta)||[b^*_{n\theta} - b_\theta](u)|N_\cdot(du) + R_{10n}(t,\theta)$$
$$+ \int_0^{t-} |\sqrt{n}R_n(u-,\theta)||b_\theta(u)|R_{9n}(du,\theta).$$

By noting that $R_{9n}(\cdot,\theta)$ is a nonnegative increasing process, we have $\|\mathrm{rem}\| = o_P(1) + \|R_{10n}\| + O_P(1)\|R_{9n}\| = o_P(1)$. Finally,

$$|\hat{W}(t,\theta) - \tilde{W}(t,\theta)| \leq |\mathrm{rem}(t,\theta)| + \int_0^t |\hat{W} - \tilde{W}|(u-,\theta)\rho_n(du).$$

By Gronwall's inequality (Section 9), we have $\hat{W}(t,\theta) = \tilde{W}(t,\theta) + o_P(1)$ uniformly in $t \leq \tau, \theta \in \Theta$. This verifies that the process $\sqrt{n}[\Gamma_{n\theta} - \Gamma_\theta]$ is asymptotically Gaussian, under the assumption that observations are iid, but Condition 2.2 does not necessarily hold.

6.3. Part (ii)

Put

$$\dot{\Gamma}_{n\theta}(t) = \int_0^t \dot{K}_n(\Gamma_{n\theta}(u-),\theta,u)N_{\cdot}(du) + \int_0^t K_n'(\Gamma_{n\theta}(u-),\theta,u)\dot{\Gamma}_{n\theta}(u-)N_{\cdot}(du), \tag{6.1}$$

$$\dot{\Gamma}_{\theta}(t) = \int_0^t \dot{k}(\Gamma_{\theta}(u-),\theta,u)EN(du) + \int_0^t k'(\Gamma_{\theta}(u-),\theta,u)\dot{\Gamma}_{\theta}(u-)EN(du). \tag{6.2}$$

Here $\dot{K} = \dot{S}/S^2$, $K' = -S'/S^2$, $\dot{k} = \dot{s}/s^2$ and $k' = -s'/s^2$. Assumption 2.0(iii) implies that $\Gamma_\theta(\tau) \le m_1^{-1}(\tau) < \infty$. For $G = k', \dot{k}$, Conditions 2.1 imply that $\sup_{\theta,t}\int_0^t |G(\Gamma_\theta(u-),\theta,u)|EN(du) < \infty$ so that $\sup_\theta \|\dot{\Gamma}_\theta\|_\infty < \infty$. Uniform consistency of the $\Gamma_{n\theta}$ process implies also that for $G_n = K_n', \dot{K}_n$, we have

$$\limsup_n \sup_{\theta,t} \int_0^t |G_n(\Gamma_{n\theta}(u-),\theta,u)|N_{\cdot}(du) < \infty$$

almost surely. Substracting equation (3.2) from (3.3), we get

$$[\dot{\Gamma}_{n\theta} - \dot{\Gamma}_\theta](t) = \Psi_n(\theta,t) + \int_0^t [\dot{\Gamma}_{n\theta} - \dot{\Gamma}_\theta](u-)K_n'(\Gamma_\theta(u-),\theta,u)N_{\cdot}(du),$$

where

$$\Psi_n(t,\theta) = R_{12n}(t,\theta) + \int_0^t \dot{\Gamma}_\theta(u-)R_{11n}(du,\theta).$$

By Lemma 5.1 and Fubini theorem, we have $\|\Psi_n\| \to 0$ a.s. And using Gronwall's inequality (Section 9), $|\dot{\Gamma}_{n\theta} - \dot{\Gamma}_\theta|(t) \to 0$ a.s. uniformly in t and θ.

Further, consider the remainder term $\mathrm{rem}_n(h,\theta,t) = \Gamma_{n\theta+h}(t) - \Gamma_{n\theta}(t) - h^T\dot{\Gamma}_{n\theta}(t)$ for $\theta, \theta+h \in \Theta$. Set $h_{2n} = \Gamma_{n,\theta+h} - \Gamma_{n,\theta}$. We have

$$\mathrm{rem}_n(h,\theta,t) = h^T\int_0^t \psi_{2n}(h,\theta,u)N_{\cdot}(du) + \int_0^t \mathrm{rem}_n(h,\theta,u-)\psi_{1n}(h,\theta,u)N_{\cdot}(du),$$

$$\psi_{1n}(h,\theta,u) = \int_0^1 K_n'(\Gamma_{n\theta}(u-) + \lambda h_{2n}(u-), \theta+\lambda h, u)d\lambda,$$

$$\psi_{2n}(h,\theta,u) = \int_0^1 \left[\dot{K}_n(\Gamma_{n\theta}(u-) + \lambda h_{2n}(u-), \theta+\lambda h, u) - \dot{K}_n(\Gamma_{n\theta}(u-),\theta,u)\right] d\lambda + \int_0^1 [K_n'(\Gamma_{n\theta}(u-)+\lambda h_{2n}(u-), \theta+\lambda h, u) - K_n'(\Gamma_{n\theta}(u-),\theta,u)] \times \dot{\Gamma}_{n\theta}(u-)d\lambda.$$

We have $\int_0^t |\psi_{1n}(h,\theta,u)|N_\cdot(du) \le \rho_n(t)$ and $\int_0^t |\psi_{2n}(h,\theta,u)|N_\cdot(du) \le h^T \int_0^t B_n(u)\times N_\cdot(du)$, for a process B_n with $\limsup_n \int_0^\tau B_n(u)N_\cdot(du) = O(1)$ a.s. This follows from condition 2.1 and some elementary algebra. By Gronwall's inequality, $\limsup_n \sup_{t\le\tau} |\mathrm{rem}_n(h,\theta,t)| = O(|h|^2) = o(|h|)$ a.s. A similar argument shows that if h_n is a nonrandom sequence with $h_n = O(n^{-1/2})$, then $\limsup_n \sup_{t\le\tau} |\mathrm{rem}_n(h_n, \theta, t)| = O(n^{-1})$ a.s. If $\hat{h}_n$ is a random sequence with $|\hat{h}_n| \xrightarrow{P} 0$, then $\limsup_n \sup_{t\le\tau} |\mathrm{rem}_n(\hat{h}_n,\theta,t)| = O_p(|\hat{h}_n|^2)$.

6.4. Part (iv)

Next suppose that θ_0 is a fixed point in Θ, $EN(t)$ is continuous, and $\hat{\theta}$ is a $\sqrt{n}$-consistent estimate of θ_0. Since $EN(t)$ is a continuous function, $\{\hat{W}(t,\theta) : t \le \tau, \theta \in \Theta\}$ converges weakly to a process W whose paths can be taken to be continuous with respect to the supremum norm. Because $\sqrt{n}[\hat{\theta} - \theta_0]$ is bounded in probability, we have $\sqrt{n}[\Gamma_{n\hat{\theta}} - \Gamma_{\theta_0}] - \sqrt{n}[\hat{\theta} - \theta_0]\dot{\Gamma}_{\theta_0} = \hat{W}(\cdot,\hat{\theta}) + \sqrt{n}[\Gamma_{\hat{\theta}} - \Gamma_{\theta_0} - [\hat{\theta} - \theta_0]\dot{\Gamma}_{\theta_0}] = \hat{W}(\cdot,\hat{\theta}) + O_P(\sqrt{n}|\hat{\theta} - \theta_0|^2) \Rightarrow W(\cdot,\theta_0)$ by weak convergence of the process $\{\hat{W}(t,\theta) : t \le \tau, \theta \in \Theta\}$ and [8].

7. Proof of Proposition 2.2

The first part follows from Remark 3.1 and part (iv) of Proposition 2.1. Note that at the true parameter value $\theta = \theta_0$, we have $\sqrt{n}[\Gamma_{n\theta_0} - \Gamma_{\theta_0}](t) = n^{1/2}\int_0^t R_{1n}(du,\theta_0)\times \mathcal{P}_{\theta_0}(u,t) + o_P(1)$, where R_{1n} is defined as in Lemma 5.1,

$$R_{1n}(t,\theta) = \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{M_i(du,\theta)}{s(\Gamma_\theta(u-),\theta,u)}.$$

and $M_i(t,\theta) = N_i(t) - \int_0^t Y_i(u)\alpha(\Gamma_\theta,\theta,Z_i)\Gamma_\theta(du)$.

We shall consider now the score process. Define

$$\begin{aligned}
\tilde{U}_{n1}(\theta) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \tilde{b}_i(\Gamma_\theta(u),\theta)M_i(dt,\theta),\\
\tilde{U}_{n2}(\theta) &= \int_0^\tau R_{1n}(du,\theta)\int_{(u,\tau]} \mathcal{P}_\theta(u,v-)r(dv,\theta).
\end{aligned}$$

Here $\tilde{b}_i(\Gamma_\theta(u),\theta) = \tilde{b}_{i1}(\Gamma_\theta(u),\theta) - \tilde{b}_{i2}(\Gamma_\theta(u),\theta)\varphi_{\theta_0}(t)$ and $\tilde{b}_{1i}(\Gamma_\theta(t),\theta) = \dot{\ell}(\Gamma_\theta(t),\theta,Z_i) - [\dot{s}/s](\Gamma_\theta(t),\theta,t)$, $\tilde{b}_{2i}(\Gamma_\theta(t),\theta) = \ell'(\Gamma_\theta(t),\theta,Z_i) - [s'/s](\Gamma_\theta(t),\theta,t)$. The function $r(\cdot,\theta)$ is the limit in probability of the term $\hat{r}_1(t,\theta)$ given below. Under Condition 2.2, it reduces at $\theta = \theta_0$ to

$$r(\cdot,\theta_0) = -\int_0^t \rho_\varphi(u,\theta_0)EN(du)$$

and $\rho_\varphi(u,\theta_0)$ is the conditional correlation defined in Section 2.3. The terms $\sqrt{n}\tilde{U}_{1n}(\theta_0)$ and $\sqrt{n}\tilde{U}_{2n}(\theta_0)$ are uncorrelated sums of iid mean zero variables and their sum converges weakly to a mean zero normal variable with covariance matrix $\Sigma_{2,\varphi}(\theta_0,\tau)$ given in the statement of Proposition 2.2.

We decompose the process $U_n(\theta)$ as $U_n(\theta) = \hat{U}_n(\theta) + \bar{U}_n(\theta)$, where

$$\begin{aligned}
\hat{U}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [b_i(\Gamma_{n\theta}(t),\theta,t) - b_{2i}(\Gamma_{n\theta}(t),\theta,t)\varphi_{\theta_0}(t)]N_i(dt),\\
\bar{U}_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau b_{2i}(\Gamma_{n\theta}(t),\theta,t)[\varphi_{n\theta} - \varphi_{\theta_0}](t)]N_i(dt).
\end{aligned}$$

We have $\hat{U}_n(\theta) = \sum_{j=1}^{3} U_{nj}(\theta)$, where

$$\begin{aligned}
U_{n1}(\theta) &= \tilde{U}_{n1}(\theta) + B_{n1}(\tau,\theta) - \int_0^\tau \varphi_{\theta_0}(u)B_{n2}(du,\theta),\\
U_{n2}(\theta) &= \int_0^\tau [\Gamma_{n\theta} - \Gamma_\theta](t)\hat{r}_1(dt,\theta),\\
U_{n3}(\theta) &= \int_0^\tau [\Gamma_{n\theta} - \Gamma_\theta](t)\hat{r}_2(dt,\theta).
\end{aligned}$$

As in Section 2.4, $b_{1i}(x,\theta,t) = \dot{\ell}(x,\theta,Z_i) - [\dot{S}/S](x,\theta,t)$ and $b_{2i}(x,\theta,t) = \ell'(x,\theta,Z_i) - [S'/S](x,\theta,t)$. If $\dot{b}_{pi}$ and b'_{pi} are the derivatives of these functions with respect to θ and x, then

$$\begin{aligned}
\hat{r}_1(s,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^s [b'_{1i}(\Gamma_\theta(t),\theta,t) - b'_{2i}(\Gamma_\theta(t),\theta,t)\varphi_{\theta_0}(t)]N_i(dt),\\
\hat{r}_2(s,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^s\int_0^1 \hat{r}_{2i}(t,\theta,\lambda)d\lambda N_i(dt),\\
\hat{r}_{2i}(t,\theta,\lambda) &= [b'_{1i}(\Gamma_\theta(t) + \lambda(\Gamma_{n\theta} - \Gamma_\theta)(t),\theta,t) - b'_{1i}(\Gamma_\theta(t),\theta,t)]\\
&\quad - [b'_{2i}(\Gamma_\theta(t) + \lambda(\Gamma_{n\theta} - \Gamma_\theta)(t),\theta,t) - b'_{2i}(\Gamma_\theta(t),\theta,t)]\,\varphi_{\theta_0}(t).
\end{aligned}$$

We also have $\bar{U}_n(\theta) = U_{n4}(\theta) + U_{n5}(\theta)$, where

$$\begin{aligned}
U_{n4}(\theta) &= -\int_0^\tau [\varphi_{n\theta} - \varphi_{\theta_0}](t)B_n(dt,\theta),\\
U_{n5}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [\varphi_{n\theta} - \varphi_{\theta_0}](t)[b_{2i}(\Gamma_{n\theta}(u),\theta,u) - b_{2i}(\Gamma_\theta(u),\theta,u)]N_i(dt),\\
B_n(t,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t b_{2i}(\Gamma_\theta(u),\theta,u)N_i(du)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \tilde{b}_{2i}(\Gamma_\theta(u),\theta,u)M_i(du,\theta) + B_{2n}(t,\theta).
\end{aligned}$$

We first show that $\bar{U}_n(\theta_0) = o_P(n^{-1/2})$. By Lemma 5.1, $\sqrt{n}B_{2n}(t,\theta_0)$ converges in probability to 0, uniformly in t. At $\theta = \theta_0$, the first term multiplied by $\sqrt{n}$ converges weakly to a mean zero Gaussian martingale. We have $\|\varphi_{n\theta_0} - \varphi_{\theta_0}\|_\infty = o_P(1)$, $\|\varphi_{\theta_0}\|_v < \infty$ and $\limsup_n \|\varphi_{n\theta_0}\|_v < \infty$. Integration by parts, Skorohod–Dudley construction and arguments similar to Lemma A.3 in [7], show that $\sqrt{n}U_{n4}(\theta_0) = o_P(1)$. We also have $\sqrt{n}U_{n5}(\theta_0) = \int_0^\tau O_P(\sqrt{n}|\Gamma_{n\theta_0} - \Gamma_{\theta_0}|(t)|\varphi_{n\theta_0} - \varphi_0|(t)N_i(dt) = o_P(1)$.

We consider now the term $\hat{U}_n(\theta_0)$. We have $\sqrt{n}U_{n3}(\theta_0) = \sqrt{n}\int_0^\tau O_P(|\Gamma_{n\theta_0} - \Gamma_{\theta_0}|(t)^2)N_\cdot(dt) = o_P(1)$. We also have $\|r(\cdot,\theta_0)\|_v < \infty, \limsup_n \|\hat{r}(\cdot,\theta_0)\|_v < \infty$ and

$\|[\hat{r}_1 - r](\cdot, \theta_0)\|_\infty = o_P(1)$, so that the same integration by parts argument implies that $\sqrt{n}U_{n2}(\theta_0) = \sqrt{n}\tilde{U}_{n2}(\theta_0) + o_P(1)$. Finally, $\sqrt{n}U_{n1}(\theta_0) = \sqrt{n}\tilde{U}_{n1}(\theta_0) + o_P(1)$, by Lemma 5.1 and Fubini theorem.

Suppose now that θ varies over a ball $B(\theta_0, \varepsilon_n)$ centered at θ_0 and having radius $\varepsilon_n, \varepsilon_n \downarrow 0, \sqrt{n}\varepsilon_n \to \infty$. It is easy to verify that for $\theta, \theta' \in B(\theta_0, \varepsilon_n)$ we have $U_n(\theta') - U_n(\theta) = -(\theta' - \theta)^T \Sigma_{1n}(\theta_0) + (\theta' - \theta)^T R_n(\theta, \theta')$, where $R_n(\theta, \theta')$ is a remainder term satisfying $\sup\{|R_n(\theta, \theta')| : \theta, \theta' \in B(\theta_0, \varepsilon_n)\} = o_P(1)$. The matrix $\Sigma_{1n}(\theta)$ is equal to the sum $\Sigma_{1n}(\theta) = \Sigma_{11n}(\theta) + \Sigma_{12n}(\theta)$,

$$\begin{aligned}
\Sigma_{11n}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [g_{1i}g_{2i}^T](\Gamma_{n\theta}(u), \theta, u)^T N_i(du), \\
\Sigma_{12n}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [f_i - S_f/S](\Gamma_{n\theta}(u), \theta, u) N_i(du),
\end{aligned}$$

where $S_f(\Gamma_{n\theta}(u), \theta, u) = n^{-1}\sum_{i=1}^n Y_i(u)[\alpha_i f_i](\Gamma_{n\theta}(u), \theta, u)$ and

$$\begin{aligned}
g_{1i}(\theta, \Gamma_{n\theta}(u), u) &= b_{1i}(\Gamma_{n\theta}(u), \theta) - b_{2i}(\Gamma_{n\theta}(u), \theta)\varphi_{\theta_0}(u), \\
g_{2i}(\theta, \Gamma_{n\theta}(u), u) &= b_{1i}(\Gamma_{n\theta}(u), \theta) + b_{2i}(\Gamma_{n\theta}(u), \theta)\dot{\Gamma}_{n\theta}(u) \\
f_i(\theta, \Gamma_{n\theta}(u), u) &= \frac{\ddot{\alpha}}{\alpha}(\Gamma_{n\theta}(u), \theta, Z_i) - \frac{\dot{\alpha}'}{\alpha}(\Gamma_{n\theta}(u), \theta, Z_i)\varphi_{\theta_0}(u)^T \\
&\quad + \dot{\Gamma}_{n\theta}(u)[\frac{\dot{\alpha}'}{\alpha}(\Gamma_{n\theta}(u), \theta, Z_i)]^T \\
&\quad + \frac{\alpha''}{\alpha}(\Gamma_{n\theta}(u), \theta, Z_i)\dot{\Gamma}_{n\theta}(u)\varphi_{\theta_0}(u)^T.
\end{aligned}$$

These matrices satisfy $\Sigma_{11n}(\theta_0) \to_P \Sigma_{1,\varphi}(\theta_0, \tau)$ and $\Sigma_{12n}(\theta_0) \to_P 0$, and $\Sigma_{1,\varphi}(\theta_0, \tau) = \Sigma_1(\theta_0)$ is defined in the statement of Proposition 2.2. By assumption this matrix is non-singular. Finally, set $h_n(\theta) = \theta + \Sigma_1(\theta_0)^{-1}U_n(\theta)$. It is easy to verify that this mapping forms a contraction on the set $\{\theta : |\theta - \theta_0| \le A_n/(1 - a_n)\}$, where $A_n = |\Sigma_1(\theta_0)^{-1}U_n(\theta_0)| = O_P(n^{-1/2})$ and $a_n = \sup\{|I - \Sigma_1(\theta_0)^{-1}\Sigma_{1n}(\theta_0) + \Sigma_1(\theta_0)^{-1}R_n(\theta, \theta')| : \theta, \theta' \in B(\theta_0, \varepsilon_n)\} = o_P(1)$. The argument is similar to Bickel *et al.* ([6], p.518), though note that we cannot apply their mean value theorem arguments.

Next consider Condition 2.3(v.2). In this case we have $\hat{U}_n(\theta') - \hat{U}_n(\theta) = -(\theta' - \theta)^T\Sigma_{1n}(\theta_0) + (\theta' - \theta)^T\hat{R}_n(\theta, \theta')$, where $\sup\{|\hat{R}_n(\theta, \theta') : \theta, \theta' \in B(\theta_0, \varepsilon_n)\} = o_P(1)$. In addition, for $\theta \in B_n(\theta_0, \varepsilon)$, we have the expansion $\bar{U}_n(\theta) = [\bar{U}_n(\theta) - \bar{U}_n(\theta_0)] + \bar{U}_n(\theta_0) = o_P(|\theta - \theta_0| + n^{-1/2})$. The same argument as above shows that the equation $\hat{U}_n(\theta)$ has, with probability tending to 1, a unique root in the ball $B(\theta_0, \varepsilon_n)$. But then, we also have $U_n(\hat{\theta}_n) = \hat{U}_n(\hat{\theta}_n) + \bar{U}_n(\hat{\theta}_n) = o_P(|\hat{\theta}_n - \theta_0| + n^{-1/2}) = o_P(O_P(n^{-1/2}) + n^{-1/2}) = o_p(n^{-1/2})$.

Part (iv) can be verified analogously, i.e. it amounts to showing that if $\sqrt{n}[\hat{\theta} - \theta_0]$ is bounded in probability, then the remainder term $\hat{R}_n(\hat{\theta}, \theta_0)$ is of order $o_P(|\hat{\theta} - \theta_0|)$, and $\bar{U}_n(\hat{\theta}) = o_P(|\hat{\theta} - \theta_0| + n^{-1/2})$.

8. Proof of Proposition 2.3

Part (i) is verified at the end of the proof. To show part (ii), define

$$\begin{aligned} D(\lambda) &= \sum_{m\geq 0} \frac{(-1)^m}{m!}\lambda^m d_m, \\ D(t,u,\lambda) &= \sum_{m\geq 0} \frac{(-1)^m}{m!}\lambda^m D_m(t,u). \end{aligned}$$

The numbers d_m and the functions $D_m(t,u)$ are given by $d_m = 1, D_m(t,u) = k(t,u)$ for $m = 0$. For $m \geq 1$ set

$$\begin{aligned} d_m &= \int \ldots \int_{(s_1,\ldots,s_m)\in(0,\tau]} \det \bar{d}_m(\mathbf{s}) b(ds_1)\cdot\ldots\cdot b(ds_m), \\ D_m(t,u) &= \int \ldots \int_{(s_1,\ldots,s_m)\in(0,\tau]} \det \bar{D}_m(t,u;\mathbf{s}) b(ds_1)\cdot\ldots\cdot b(ds_m), \end{aligned}$$

where for any $\mathbf{s} = (s_1,\ldots,s_m)$, $\bar{d}_m(\mathbf{s})$ is an $m\times m$ matrix with entries $\bar{d}_m(\mathbf{s}) = [k(s_i,s_j)]$, and $\bar{D}_m(t,u;\mathbf{s})$ is an $(m+1)\times(m+1)$ matrix

$$\bar{D}_m(t,u;\mathbf{s}) = \begin{pmatrix} k(t,u), & U_m(t;\mathbf{s}) \\ V_m(\mathbf{s};u), & \bar{d}_m(\mathbf{s}) \end{pmatrix},$$

where $U_m(t;\mathbf{s}) = [k(t,s_1),\ldots,k(t,s_m)]$, $V_m(\mathbf{s};u) = [k(s_1,u),\ldots,k(s_m,u)]^T$. By Fredholm determinant formula [25], the resolvent of the kernel k is given by $\tilde{\Delta}(t,u,\lambda) = D(t,u,\lambda)/D(\lambda)$, for all λ such that $D(\lambda)\neq 0$, so that

$$d_m = \int \ldots \int_{\substack{s_1,\ldots,s_m\in(0,\tau] \\ distinct}} \det \bar{d}_m(\mathbf{s}) b(ds_1)\cdot\ldots\cdot b(ds_m),$$

because the determinant is zero whenever two or more points $s_i, i = 1,\ldots,m$ are equal. By Fubini theorem, the right-hand side of the above expression is equal to

$$\begin{aligned} &\sum_{\pi} \int \ldots \int_{0<s_1<s_2<\cdots<s_m\leq\tau} \det \bar{d}_m(s_{\pi(1)},\ldots,s_{\pi(m)}) b(ds_1)\cdot\ldots\cdot b(ds_m) \\ &\quad = m! \int \ldots \int_{0<s_1<s_2<\cdots<s_m\leq\tau} \det \bar{d}_m(s_1,\ldots,s_m) b(ds_1)\cdot\ldots\cdot b(ds_m). \end{aligned}$$

The first sum extends over the $m!$ possible permutations $\pi = (\pi(1),\ldots,\pi(m))$ of the index set $\{1,\ldots,m\}$. The second line follows by noting that

$$\det \bar{d}_m(s_{\pi(1)},\ldots,s_{\pi(m)}) = \det \bar{d}_m(s_1,\ldots,s_m)$$

for any such permutation, because the matrix $\bar{d}_m(s_{\pi(1)},\ldots,s_{\pi(m)})$ is symmetric and to rearrange it into the matrix $\bar{d}_m(s_1,\ldots,s_m)$, we need the same number of transpositions of rows and columns. Since the total number of such transpositions is even, the determinants have the same sign. In the same way, the function $D_m(t,u)$ is equal to

$$m! \int \ldots \int_{0<s_1<s_2<\cdots<s_m\leq\tau} \det \bar{D}_m(t,u;s_1,\ldots,s_m) b(ds_1)\cdot\ldots\cdot b(ds_m),$$

so that in both cases it is enough to consider the determinants for ordered sequences $\mathbf{s} = (\mathrm{s}_1, \ldots, \mathrm{s}_\mathrm{m}), \mathrm{s}_1 < \mathrm{s}_2 < \ldots < \mathrm{s}_\mathrm{m}$ of points in $(0, \tau]^m$.

For any such sequence $\mathbf{s}$, the matrix $d_m(\mathbf{s})$ has a simple pattern:

$$\bar{d}_m(\mathbf{s}) = \begin{pmatrix} c(s_1) & c(s_1) & c(s_1) & \ldots & c(s_1) \\ c(s_1) & c(s_2) & c(s_2) & \ldots & c(s_2) \\ c(s_1) & c(s_2) & c(s_3) & \ldots & c(s_3) \\ \vdots & & & & \vdots \\ c(s_1) & c(s_2) & c(s_3) & \ldots & c(s_m) \end{pmatrix}.$$

We have $\bar{d}_m(\mathbf{s}) = A_m^T C_m(\mathbf{s}) A_m$ where $C_m(\mathbf{s})$ is a diagonal matrix of increments

$$C_m(\mathbf{s}) = \mathrm{diag}\,[c(s_1) - c(s_0), c(s_2) - c(s_1), \ldots c(s_m) - c(s_{m-1})],$$

$(c(s_0) = 0, s_0 = 0)$ and A_m is an upper triangular matrix

$$A_m = \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ 0 & 1 & \ldots & 1 & 1 \\ \vdots & & & & \vdots \\ 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

To see this it is enough to note that Brownian motion forms a process with independent increments, and the kernel $k(s,t) = c(s \wedge t)$ is the covariance function of a time transformed Brownian motion.

Apparently, $\det A_m = 1$. Therefore

$$\det \bar{d}_m(\mathbf{s}) = \prod_{j=1}^{m} [c(s_j) - c(s_{j-1})]$$

and

$$\begin{aligned} \det \bar{D}_m(t,u;\mathbf{s}) &= \det \bar{d}_m(\mathbf{s})[c(t \wedge u) - U_m(t;\mathbf{s})[\bar{d}_m(\mathbf{s})]^{-1} V_m(\mathbf{s};u)] \\ &= \det \bar{d}_m(\mathbf{s})[c(t \wedge u) - U_m(t;\mathbf{s}) A_m^{-1} C_m^{-1}(\mathbf{s})(A_m^T)^{-1} V_m(\mathbf{s};u)]. \end{aligned}$$

The inverse A_m^{-1} is given by Jordan matrix

$$A_m^{-1} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ & \vdots & & & \vdots & \\ 0 & 0 & 0 & \ldots & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

and a straightforward multiplication yields

$$\begin{aligned} \det \bar{D}_m(t,u;\mathbf{s}) &= c(t \wedge u) \prod_{j=1}^{m} [c(s_j) - c(s_{j-1})] \\ &\quad - \sum_{i=1}^{m} [c(t \wedge s_i) - c(t \wedge s_{i-1})][c(u \wedge s_i) - c(u \wedge s_{i-1})] \\ &\quad \times \prod_{j=1, j \neq i}^{m} [c(s_j) - c(s_{j-1})]. \end{aligned}$$

By noting that the i-th summand is zero whenever $t \wedge u < s_{i-1}$ and using induction on m, it is easy to verify that for $t \le u$ the determinant reduces to the sum

$$\begin{aligned}\det \bar{D}_m(t,u;\mathbf{s}) &= 1(t \le u < s_1)c(t)[c(s_1)-c(u)]\prod_{j=2}^{m}[c(s_j)-c(s_{j-1})] \\ &\quad + 1(s_m < t \le u)\prod_{j=1}^{m}[c(s_j)-c(s_{j-1})][c(t)-c(s_m)] \\ &\quad + \sum_{i=1}^{m-1} 1(s_i < t \le u < s_{i+1})\Bigg(\prod_{j=1}^{i}[c(s_j)-c(s_{j-1})][c(t)-c(s_i)]\Bigg) \\ &\quad \times \Bigg([c(s_{i+1})-c(u)]\prod_{j=i+2}^{m}[c(s_j)-c(s_{j-1})]\Bigg),\end{aligned}$$

where in the last sum, product over an empty set of indices is interpreted as equal to 1. Thus we have a simple expression for the two determinants. Integration with respect to the product measure $b(ds_1)\dots b(ds_m)$ and induction on m yields also

$$\begin{aligned}\frac{1}{m!}d_m &= \Psi_{0m}(0,\tau), \\ \frac{1}{m!}D_m(t,u) &= \sum_{l=0}^{m}\Psi_{1l}(0,t\wedge u)\Psi_{0,m-l}(t\vee u,\tau),\end{aligned}$$

for $m \ge 0$. The numerator and denominator of the Fredholm determinant formula are bounded functions for any point τ satisfying the condition 2.0 (iii). For $\lambda = -1$, the ratio $\tilde{\Delta}(t,u,\lambda) = D(t,u,\lambda)/D(\lambda)$ reduces to the function $\tilde{\Delta}$ given in the statement of part (ii). Using monotonicity of the functions Ψ_0 with respect to the length of the interval $(s,t]$, we also have $\tilde{\Delta}(t,u) \le \Psi_0(0,\tau)c(t\wedge u)$. If τ_0 is a continuity point of $EY(t)$ and $\kappa(\tau_0) < \infty$, then the denominator is bounded, and the inequality is satisfied for any $u, t < \tau_0$.

Parts (iii) and (iv) are easy to verify using this last observation. For example, if $\kappa(\tau_0) < \infty$ then for any $\tilde{\eta} \in L_2(b)$, the Fredholm equation has a unique solution $\tilde{\psi}$ and

$$\|\tilde{\psi}\|_2 \le \|\tilde{\eta}\|_2 + \left[\int_0^{\tau_0}\left[\int_0^{\tau_0}\tilde{\Delta}(t,u)b(du)\tilde{\eta}(u)\right]^2 b(dt)\right]^{1/2}.$$

By Cauchy–Schwarz inequality and monotone convergence, the second term is bounded by

$$\begin{aligned}&\|\tilde{\eta}\|_2\left[\int_0^{\tau_0}\int_0^{\tau_0}\tilde{\Delta}^2(t,u)b(du)b(dt)\right]^{1/2} \\ &\quad \le \|\tilde{\eta}\|_2\Psi_0(0,\tau_0)\left[\int_0^{\tau_0}\int_0^{\tau_0}c(t\wedge u)b(du)b(dt)\right]^{1/2} \\ &\quad \le \|\tilde{\eta}\|_2\Psi_0(0,\tau_0)\left[\int_0^{\tau_0}\int_0^{\tau_0}c(t)c(u)b(du)b(dt)\right]^{1/2} = \|\tilde{\eta}\|_2\Psi_0(0,\tau_0)\kappa(\tau_0).\end{aligned}$$

Part (v) follows from part (iv) and Lemma 2.1. Part (vi) can be verified using straightforward but laborious algebra.

Part (i). For $u > s$, set $c((s,u]) = c(u) - c(s)$. The n-the term of the series $\Psi_{0n}(s,t)$ is given by the multiple integral

$$\int_{s<s_1<\cdots<s_n\le t} c((s,s_1])b(ds_1)c((s_1,s_2])b(ds_2)\cdots c((s_{n-1},s_n])b(ds_n),$$

and satisfies $\Psi_{0n}(s,t) \le (n!)^{-1}[I(s,t)]^n$, $I(s,t) = \int_s^t c((s,u])b(u)$. The integral $I(s,t)$ is increasing with the width of the interval $(s,t]$ and is bounded by $\kappa(\tau)$. Thus $\Psi_0(s,t) \le \exp\kappa(\tau) < \infty$ for all $0 < s < t \le \tau$. If in addition $\kappa(\tau_0) < \infty$, then $\Psi_0(s,t)$ is bounded for all $0 < s < t \le \tau_0$. In both circumstances, this implies that the remaining interval functions $\Psi_j(s,t)$, $0 < s \le t \le \tau$ are finite for any point τ satisfying Condition 2.0(iii), and monotonically increasing with the size of the interval.

While the identities can be verified by applying Fubini to each term of the $\Psi_j, j = 0,1,2,3$ series, the following provides an interpretation in terms of linear Volterra equations. First, it is easy to see that the "odd" functions satisfy

$$\begin{aligned}
\Psi_1(s,t) &= c((s,t]) + \int_{(s,t]} c((s,u])b(du)\Psi_1(u,t) \\
&= c((s,t]) + \int_{(s,t]} \Psi_1(s,u)b(du)c((u,t]), \\
\Psi_3(s,t) &= b([s,t)) + \int_{[s,t)} b([s,u))c(du)\Psi_3(u,t) \\
&= b([s,t)) + \int_{[s,t)} \Psi_3(s,u)c(du)b([u,t)),
\end{aligned}$$

so that they form resolvents of linear Volterra equations. The "even" functions Ψ_0 and Ψ_2 satisfy such equations

$$\begin{aligned}
\Psi_0(s,t) &= 1 + \int_{(s,t]} c((s,u])b(du)\Psi_0(u,t) \\
&= 1 + \int_{(s,t]} \Psi_0(s,u-)c(du)b([u,t]), \\
\Psi_2(s,t) &= 1 + \int_{[s,t)} b([s,u))c(du)\Psi_2(u,t) \\
&= 1 + \int_{[s,t)} \Psi_2(s,u+)b(du)c((u,t)).
\end{aligned}$$

With fixed t, the equations

$$\begin{aligned}
h_1(s,t) - \int_{(s,t]} c((s,u])b(du)h_1(u,t) &= g_1(s,t), \\
h_3(s,t) - \int_{(s,t]} b([s,u))c(du)h_3(u,t) &= g_3(s,t),
\end{aligned}$$

have unique solutions

$$\begin{aligned}
h_1(s,t) &= g_1(s,t) + \int_{(s,t]} \Psi_1(s,u)b(du)g_1(u,t), \\
h_3(s,t) &= g_3(s,t) + \int_{[s,t)} \Psi_3(s,u)c(du)g_3(u,t).
\end{aligned}$$

The first pair of equations for Ψ_0 and Ψ_2 in part (i) follows by setting $g_1(s,t) = 1 = g_3(s,t)$. With s fixed, the equations

$$\begin{aligned}\bar{h}_1(s,t) - \int_{[s,t)} \bar{h}_1(s,u+)b(du)c_1((u,t)) &= \bar{g}_1(s,t),\\ \bar{h}_3(s,t) - \int_{(s,t]} \bar{h}_3(s,u-)c(du)b_3([u,t]) &= \bar{g}_3(s,t),\end{aligned}$$

have solutions

$$\begin{aligned}\bar{h}_1(s,t) &= \bar{g}_1(s,t) + \int_{[s,t)} \bar{g}_1(s,u+)b(du)\Psi_1(u,t-),\\ \bar{h}_3(s,t) &= \bar{g}_3(s,t) + \int_{(s,t]} \bar{g}_3(s,u-)c(du)\Psi_3(u,t+).\end{aligned}$$

The second pair of equations for Ψ_0 and Ψ_2 in part (i) follows by setting $\bar{g}_1(s,t) \equiv 1 \equiv \bar{g}_3(s,t)$. Next, the "odd" functions can be represented in terms of "even" functions using Fubini.

9. Gronwall's inequalities

Following Gill and Johansen [18], recall that if b is a cadlag function of bounded variation, $\|b\|_v \le r_1$ then the associated product integral $\mathcal{P}(s,t) = \prod_{(s,t]}(1+b(du))$ satisfies the bound $|\mathcal{P}(s,t)| \le \prod_{(s,t]}(1 + \|b\|_v(dw)) \le \exp \|b\|_v(s,t]$ uniformly in $0 < s < t \le \tau$. Moreover, the functions $s \to \mathcal{P}(s,t), s \le t \le \tau$ and $t \to \mathcal{P}(s,t), t \in (s,\tau]$ are of bounded variation with variation norm bounded by $r_1 e^{r_1}$.

The proofs use the following consequence of Gronwall's inequalities in Beesack [3] and Gill and Johansen [18]. If b is a nonnegative measure and $y \in D([0,\tau])$ is a nonnegative function then for any $x \in D([0,\tau])$ satisfying

$$0 \le x(t) \le y(t) + \int_{(0,t]} x(u-)b(du), \quad t \in [0,\tau],$$

we have

$$0 \le x(t) \le y(t) + \int_{(0,t]} y(u-)b(du)\mathcal{P}(u,t), \quad t \in [0,\tau].$$

Pointwise in t, $|x(t)|$ is bounded by

$$\max\{\|y\|_\infty, \|y^-\|_\infty\}[1 + \int_{(0,t]} b(du)\mathcal{P}(u,t)] \le \{\|y\|_\infty, \|y^-\|_\infty\} \exp[\int_0^t b(du)].$$

We also have $\|e^{-b}|x|\|_\infty \le \max\{\|y\|_\infty, \|y^-\|_\infty\}$. Further, if $0 \not\equiv y \in D([0,\tau])$ and b is a function of bounded variation then the solution to the linear Volterra equation

$$x(t) = y(t) + \int_0^t x(u-)b(du)$$

is unique and given by

$$x(t) = y(t) + \int_{(0,t]} y(u-)b(du)\mathcal{P}(u,t).$$

We have $|x(t)| \leq \max\{\|y\|_\infty, \|y^-\|_\infty\} \exp \int_0^t d\|b\|_v$ and $\| \exp[-\int_\cdot d\|b\|_v]|x|\|_\infty \leq \max\{\|y\|_\infty, \|y^-\|_\infty\}$. If $y_\theta(t)$, and $b_\theta(t) = \int_0^t k_\theta(u)n(du)$ are functions dependent on a Euclidean parameter $\theta \in \Theta \subset R^d$, and $|k_\theta|(t) \leq k(t)$, then these bounds hold pointwise in θ and

$$\sup_{\substack{t\leq\tau \\ \theta\in\Theta}} \{\exp[-\int_0^t k(u)n(du)]|x_\theta(t)|\} \leq \max\{\sup_{\substack{u\leq\tau \\ \theta\in\Theta}} |y_\theta|(u), \sup_{\substack{u\leq\tau \\ \theta\in\Theta}} |y_\theta(u-)|\}.$$

Acknowledgement

The paper was presented at the First Erich Leh–mann Symposium, Guanajuato, May 2002. I thank Victor Perez Abreu and Javier Rojo for motivating me to write it. I also thank Kjell Doksum, Misha Nikulin and Chris Klaassen for some discussions. The paper benefited also from comments of an anonymous reviewer and the Editor Javier Rojo.

References

[1] Arcones, M. A. and Giné, E. (1995). On the law of iterated logarithm for canonical U-statistics and processes. *Stochastic Processes Appl.* **58**, 217–245.
[2] Bennett. S. (1983). Analysis of the survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.
[3] Beesack, P. R. (1975). *Gronwall Inequalities.* Carlton Math. Lecture Notes **11**, Carlton University, Ottawa.
[4] Bickel, P. J. (1986) Efficient testing in a class of transformation models. In *Proceedings of the 45th Session of the International Statistical Institute.* ISI, Amsterdam, 23.3-63–23.3-81.
[5] Bickel, P. J. and Ritov, Y. (1995). Local asymptotic normality of ranks and covariates in transformation models. In *Festschrift for L. LeCam* (D. Pollard and G. Yang, eds). Springer.
[6] Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press.
[7] Bilias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Ann. Statist.* **25**, 662–683.
[8] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley.
[9] Bogdanovicius, V. and Nikulin, M. (1999). Generalized proportional hazardss model based on modified partial likelihood. *Lifetime Data Analysis* **5**, 329–350.
[10] Bogdanovicius, M. Hafdi, M. A. and Nikulin, M. (2004). Analysis of survival data with cross-effects of survival functions. *Biostatistics* **5**, 415–425.
[11] Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *J. Amer. Statist. Assoc.* **92**, 227–235.
[12] Cox, D. R. (1972). Regression models in life tables. *J. Roy. Statist. Soc. Ser. B.* **34**, 187–202.
[13] Cuzick, J. (1988) Rank regression. *Ann. Statist.* **16**, 1369–1389.
[14] Dabrowska, D. M. and Doksum, K.A. (1988). Partial likelihood in transformation models. *Scand. J. Statist.* **15**, 1–23.

[15] Dabrowska, D. M., Doksum, K. A. and Miura, R. (1989). Rank estimates in a class of semiparametric two–sample models. *Ann. Inst. Statist. Math.* **41**, 63–79.

[16] Dabrowska, D. M. (2005). Quantile regression in transformation models. *Sankhyā* **67**, 153–187.

[17] Dabrowska, D. M. (2006). Information bounds and efficient estimation in a class of transformation models. Manuscript in preparation.

[18] Gill, R. D. and Johansen, S. (1990). A survey of product integration with a view toward application in survival analysis. *Ann. Statist.* **18**, 1501–1555.

[19] Giné, E. and Guillou, A. (1999). Laws of iterated logarithm for censored data. *Ann. Probab.* **27**, 2042–2067.

[20] Gripenberg, G., Londen, S. O. and Staffans, O. (1990). *Volterra Integral and Functional Equations.* Cambridge University Press.

[21] Klaassen, C. A. J. (1993). Efficient estimation in the Clayton–Cuzick model for survival data. Tech. Report, University of Amsterdam, Amsterdam, Holland.

[22] Kosorok, M. R. , Lee, B. L. and Fine, J. P. (2004). Robust inference for univariate proportional hazardss frailty regression models. *Ann. Statist.* **32**, 1448–1449.

[23] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24**, 23–43.

[24] Maurin, K. (1976). *Analysis*. Polish Scientific Publishers and D. Reidel Pub. Co, Dodrecht, Holland.

[25] Mikhlin, S. G. (1960). *Linear Integral Equations.* Hindustan Publ. Corp., Delhi.

[26] Murphy, S. A. (1994). Consistency in a proportional hazardss model incorporating a random effect. *Ann. Statist.* **25**, 1014–1035.

[27] Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92**, 968–976.

[28] Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**, 25–44.

[29] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of the optimization estimators. *Econometrica* **57**, 1027–1057.

[30] Parner, E. (1998). Asymptotic theory for the correlated gamma model. *Ann. Statist.* **26**, 183–214.

[31] Scharfstein, D. O., Tsiatis, A. A. and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time to event data. *Lifetime and Data Analysis* **4**, 355–393.

[32] Serfling, R. (1981). *Approximation Theorems of Mathematical Statistics.* Wiley.

[33] Slud, E. and Vonta, F. (2004). Consistency of the NMPL estimator in the right censored transformation model. *Scand. J. Statist.* **31** , 21–43.

[34] Yang, S. and Prentice, R. (1999). Semiparametric inference in the proportional odds regression model. *J. Amer. Statist. Assoc.* **94**, 125–136.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 170–182
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000446

Bayesian transformation hazard models

Gousheng Yin[1] **and Joseph G. Ibrahim**[2]

M. D. Anderson Cancer Center and University of North Carolina

Abstract: We propose a class of transformation hazard models for right-censored failure time data. It includes the proportional hazards model (Cox) and the additive hazards model (Lin and Ying) as special cases. Due to the requirement of a nonnegative hazard function, multidimensional parameter constraints must be imposed in the model formulation. In the Bayesian paradigm, the nonlinear parameter constraint introduces many new computational challenges. We propose a prior through a conditional-marginal specification, in which the conditional distribution is univariate, and absorbs all of the nonlinear parameter constraints. The marginal part of the prior specification is free of any constraints. This class of prior distributions allows us to easily compute the full conditionals needed for Gibbs sampling, and hence implement the Markov chain Monte Carlo algorithm in a relatively straightforward fashion. Model comparison is based on the conditional predictive ordinate and the deviance information criterion. This new class of models is illustrated with a simulation study and a real dataset from a melanoma clinical trial.

1. Introduction

In survival analysis and clinical trials, the Cox [10] proportional hazards model has been routinely used. For a subject with a possibly time-dependent covariate vector $\mathbf{Z}(t)$, the proportional hazards model is given by,

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}(t)\}, \tag{1.1}$$

where $\lambda_0(t)$ is the unknown baseline hazard function and $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector of interest. Cox [11] proposed to estimate $\boldsymbol{\beta}$ under model (1.1) by maximizing the partial likelihood function and its large sample theory was established by Andersen and Gill [1]. However, the proportionality of hazards might not be a valid modeling assumption in many situations. For example, the true relationship between hazards could be parallel, which leads to the additive hazards model (Lin and Ying [24]),

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}'\mathbf{Z}(t). \tag{1.2}$$

As opposed to the hazard ratio yielded in (1.1), the hazard difference can be obtained from (1.2), which formulates a direct association between the expected num-

[1]Department of Biostatistics & Applied Mathematics, M. D. Anderson Cancer Center, The University of Texas, 1515 Holcombe Boulevard 447, Houston, TX 77030, USA, e-mail: gsyin@mdanderson.org

[2]Department of Biostatistics, The University of North Carolina, Chapel Hill, NC 27599, USA, e-mail: ibrahim@bios.unc.edu

AMS 2000 subject classifications: primary 62N01; secondary 62N02, 62C10.

Keywords and phrases: additive hazards, Bayesian inference, constrained parameter, CPO, DIC, piecewise exponential distribution, proportional hazards.

ber of events or death occurrences and risk exposures. O'Neill [28] showed that use of the Cox model can result in serious bias when the additive hazards model is correct. Both the multiplicative and additive hazards models have sound biological motivations and solid statistical bases.

Lin and Ying [25], Martinussen and Scheike [26] and Scheike and Zhang [30] proposed general additive-multiplicative hazards models in which some covariates impose the proportional hazards structure and others induce an additive effect on the hazards. In contrast, we link the additive and multiplicative hazards models in a completely different fashion. Through a simple transformation, we construct a class of hazard-based regression models that includes those two commonly used modeling schemes. In the usual linear regression model, the Box–Cox transformation [4] may be applied to the response variable,

$$\phi(Y) = \begin{cases} (Y^\gamma - 1)/\gamma & \gamma \neq 0 \\ \log(Y) & \gamma = 0, \end{cases} \tag{1.3}$$

where $\lim_{\gamma \to 0}(Y^\gamma - 1)/\gamma = \log(Y)$. This transformation has been used in survival analysis as well [2, 3, 5, 7, 13, 32]. Breslow and Storer [7] and Barlow [3] applied this family of power transformations to the covariate structure to model the relative risk $R(\mathbf{Z})$,

$$\log R(\mathbf{Z}) = \begin{cases} \{(1 + \boldsymbol{\beta}'\mathbf{Z})^\gamma - 1\}/\gamma & \gamma \neq 0 \\ \log(1 + \boldsymbol{\beta}'\mathbf{Z}) & \gamma = 0, \end{cases}$$

where $R(\mathbf{Z})$ is the ratio of the incidence rate at one level of the risk factor to that at another level. Aranda-Ordaz [2] and Breslow [5] proposed a compromise between these two special cases, $\gamma = 0$ or 1, while their focus was only on grouped survival data by analyzing sequences of contingency tables. Sakia [29] gave an excellent review on this power transformation.

The proportional and additive hazards models may be viewed as two extremes of a family of regression models. On a basis that is very different from the available methods in the literature, we propose a class of regression models for survival data by imposing the Box–Cox transformation on both the baseline hazard $\lambda_0(t)$ and the hazard $\lambda(t|\mathbf{Z})$. This family of transformation models is very general, which includes the Cox proportional hazards model and the additive hazards model as special cases. By adding a transformation parameter, the proposed modeling structure allows a broad class of hazard patterns. In many applications where the hazards are neither proportional nor parallel, our proposed transformation model provides a unified and flexible methodology for analyzing survival data.

The rest of this article is organized as follows. In Section 2.1, we introduce notation and a class of regression models based on the Box–Cox transformed hazards. In Section 2.2, we derive the likelihood function for the proposed model using piecewise constant hazards. In Section 2.3, we propose a prior specification scheme incorporating the parameter constraints within the Bayesian paradigm. In Section 3, we derive the full conditional distributions needed for Gibbs sampling. In Section 4, we introduce model selection methods based on the conditional predictive ordinate (CPO) in Geisser [14] and the deviance information criterion (DIC) proposed by Spiegelhalter *et al.* [31]. We illustrate the proposed methods with data from a melanoma clinical trial, and examine the model using a simulation study in Section 5. We give a brief discussion in Section 6.

2. Transformation hazard models

2.1. A new class of models

For n independent subjects, let T_i $(i = 1, \ldots, n)$ be the failure time for subject i and $\mathbf{Z}_i(t)$ be the corresponding $p \times 1$ covariate vector. Let C_i be the censoring variable and define $Y_i = \min(T_i, C_i)$. The censoring indicator is $\nu_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. Assume that T_i and C_i are independent conditional on $\mathbf{Z}_i(t)$, and that the triplets $\{(T_i, C_i, \mathbf{Z}_i(t)), i = 1, \ldots, n\}$ are independent and identically distributed.

For right-censored failure time data, we propose a class of Box–Cox transformation hazard models,

$$\phi\{\lambda(t|\mathbf{Z}_i)\} = \phi\{\lambda_0(t)\} + \boldsymbol{\beta}'\mathbf{Z}_i(t), \tag{2.1}$$

where $\phi(\cdot)$ is a known link function given by (1.3). We take γ as fixed throughout our development for the following reasons. First, our main goal is to model selection on γ, by fitting separate models for each value of γ and evaluating them through a model selection criterion. Once the best γ is chosen according to a model selection criterion, posterior inference regarding $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is then based on that γ. Second, in real data settings, there is typically very little information contained in the data to estimate γ directly. Third, posterior estimation of γ is computationally difficult and often numerically unstable due to the constraint (2.3) as well as its weak identifiability property. To understand how the hazard varies with respect to γ, we carried out a numerical study as follows. We assume that $\lambda_0(t) = t/3$ in one case, and $\lambda_0(t) = t^2/5$ in another case. A single covariate Z takes a value of 0 or 1 with probability .5, and $\gamma = (0, .25, .5, .75, 1)$. Model (2.1) can be written as

$$\lambda(t|\mathbf{Z}_i) = \{\lambda_0(t)^\gamma + \gamma\boldsymbol{\beta}'\mathbf{Z}_i(t)\}^{1/\gamma}.$$

As shown in Figure 1, there is a broad family of models for $0 \leq \gamma \leq 1$. Our primary interest for γ lies in $[0, 1]$, which covers the two popular cases and a family of intermediate modeling structures between the proportional $(\gamma = 0)$ and the additive $(\gamma = 1)$ hazards models.

Misspecified models may lead to severe bias and wrong statistical inference. In many applications where neither the proportional nor the parallel hazards assumption holds, one can apply (2.1) to the data with a set of prespecified γ's, and choose

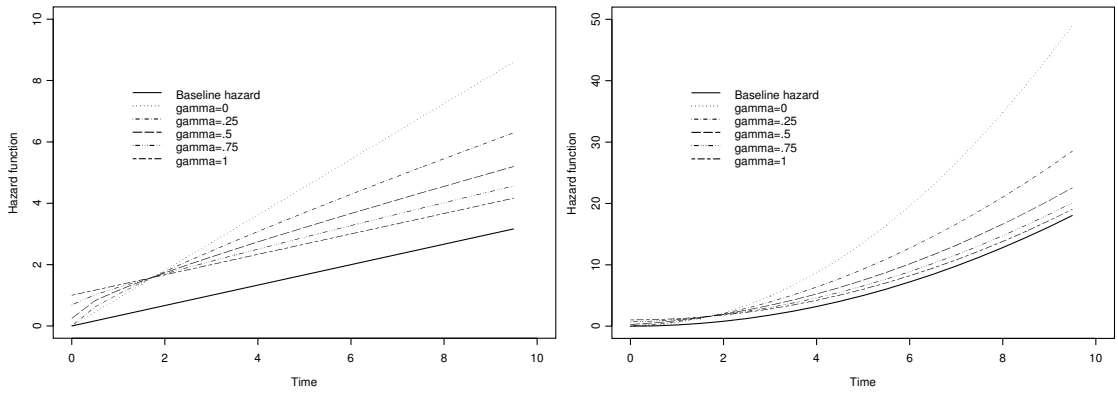


FIG 1. *The relationships between* $\lambda_0(t)$ *and* $\lambda(t|Z) = \{\lambda_0(t)^\gamma + \gamma Z\}^{1/\gamma}$, *with* $Z = 0, 1$. *Left:* $\lambda_0(t) = t/3$; *right:* $\lambda_0(t) = t^2/5$.

the best fitting model according to a suitable model selection criterion. The need for the general class of models in (2.1) can be demonstrated by the E1690 data from the Eastern Cooperative Oncology Group (ECOG) phase III melanoma clinical trial (Kirkwood *et al.* [23]). The objective of this trial was to compare high-dose interferon to observation (control). Relapse-free survival was a primary outcome variable, which was defined as the time from randomization to progression of tumor or death. As shown in Section 5, the best choice of γ in the E1690 data is indeed neither 0 nor 1, but $\gamma = .5$.

Due to the extra parameter γ, $\boldsymbol{\beta}$ is intertwined with $\lambda_0(t)$ in (2.1). As a result, the model is very different from either the proportional hazards model, which can be solved through the partial likelihood procedure, or the additive hazards model, where the estimating equation can be constructed based on martingale integrals. Here, we propose to conduct inference with this transformation model using a Bayesian approach.

2.2. Likelihood function

The piecewise exponential model is chosen for $\lambda_0(t)$. This is a flexible and commonly used modeling scheme and usually serves as a benchmark for the comparison of parametric and nonparametric approaches (Ibrahim, Chen and Sinha [21]). Other nonparametric Bayesian methods for modeling $\lambda_0(t)$ are available in the literature [20, 22, 27]. Let y_i be the observed time for the ith subject, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$, and $\mathbf{Z}(t) = (\mathbf{Z}_1(t), \ldots, \mathbf{Z}_n(t))'$. Let J denote the number of partitions of the time axis, i.e. $0 < s_1 < \cdots < s_J$, $s_J > y_i$ for $i = 1, \ldots, n$, and that $\lambda_0(y) = \lambda_j$ for $y \in (s_{j-1}, s_j]$, $j = 1, \ldots, J$. When $J = 1$, the model reduces to a parametric exponential model. By increasing J, the piecewise constant hazard formulation can essentially model any shape of the underlying hazard. The usual way to partition the time axis is to obtain an approximately equal number of failures in each interval, and to guarantee that each time interval contains at least one failure. Define $\delta_{ij} = 1$ if the ith subject fails or is censored in the jth interval, and 0 otherwise. Let $D = (n, \mathbf{y}, \mathbf{Z}(t), \boldsymbol{\nu})$ denote the observed data, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$. For ease of exposition and computation, let $\mathbf{Z}_i \equiv \mathbf{Z}_i(t)$, then the likelihood function is

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\lambda} | D) \;=\; & \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_j^{\gamma} + \gamma \boldsymbol{\beta}' \mathbf{Z}_i)^{\delta_{ij} \nu_i / \gamma} \\
& \times e^{-\delta_{ij} \{ (\lambda_j^{\gamma} + \gamma \boldsymbol{\beta}' \mathbf{Z}_i)^{1/\gamma} (y_i - s_{j-1}) + \sum_{g=1}^{j-1} (\lambda_g^{\gamma} + \gamma \boldsymbol{\beta}' \mathbf{Z}_i)^{1/\gamma} (s_g - s_{g-1}) \}}.
\end{aligned} \tag{2.2}
$$

2.3. Prior distributions

The joint prior distribution of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ needs to accommodate the nonnegativity constraint for the hazard function, that is,

$$
\lambda_j^{\gamma} + \gamma \boldsymbol{\beta}' \mathbf{Z}_i \geq 0 \quad (i = 1, \ldots, n;\ j = 1, \ldots, J). \tag{2.3}
$$

Constrained parameter problems typically make Bayesian computation and analysis quite complicated [8, 9, 16]. For example, the order constraint on a set of parameters (e.g., $\theta_1 \leq \theta_2 \leq \cdots$) is very common in Bayesian hierarchical models. In these settings, closed form expressions for the normalizing constants in the full conditional

distributions are typically available. However, for our model, this is not the case; the normalizing constant involves a complicated intractable integral. The nonnegativity of the hazard constraint is very different from the usual order constraints. If the hazard is negative, the likelihood function and the posterior density are not well defined. One way to proceed with this nonlinear constraint is to specify an appropriately truncated joint prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, such as a truncated multivariate normal prior $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $(\boldsymbol{\beta}|\boldsymbol{\lambda})$ to satisfy this constraint. This would lead to a prior distribution of the form

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\beta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})I(\lambda_j^\gamma + \gamma\boldsymbol{\beta}'\mathbf{Z}_i \geq 0,\ i = 1, \ldots, n;\ j = 1, \ldots, J).$$

Following this route, we would need to analytically compute the normalizing constant,

$$c(\boldsymbol{\lambda}) = \int \cdots \int_{\lambda_j^\gamma + \gamma\boldsymbol{\beta}'\mathbf{Z}_i \geq 0 \text{ for all } i,j} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})\right\} d\beta_1 \cdots d\beta_p$$

to construct the full conditional distribution of $\boldsymbol{\lambda}$. However, $c(\boldsymbol{\lambda})$ involves a p-dimensional integral on a complex nonlinear constrained parameter space, which cannot be obtained in a closed form. Such a prior would lead to intractable full conditionals, therefore making Gibbs sampling essentially impossible.

To circumvent the multivariate constrained parameter problem, we reduce our prior specification to a one-dimensional truncated distribution, and thus the normalizing constant can be obtained in a closed form. Without loss of generality, we assume that all the covariates are positive. Let $\mathbf{Z}_{i(-k)}$ denote the covariate $\mathbf{Z}_i$ with the kth component Z_{ik} deleted, and let $\boldsymbol{\beta}_{(-k)}$ denote the $(p-1)$-dimensional parameter vector with β_k removed, and define

$$h_\gamma(\lambda_j, \boldsymbol{\beta}_{(-k)}, \mathbf{Z}_i) = \min_{i,j}\left\{\frac{\lambda_j^\gamma + \gamma\boldsymbol{\beta}'_{(-k)}\mathbf{Z}_{i(-k)}}{\gamma Z_{ik}}\right\}.$$

We propose a joint prior for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ of the form

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \pi(\beta_k|\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})I\left(\beta_k \geq -h_\gamma(\lambda_j, \boldsymbol{\beta}_{(-k)}, \mathbf{Z}_i)\right)\pi(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}). \tag{2.4}$$

We see that β_k and $(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})$ are not independent a priori due to the nonlinear parameter constraint. This joint prior specification only involves one parameter β_k in the constraints and makes all the other parameters $(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})$ free of constraints.

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution. Specifically, we take $(\beta_k|\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})$ to have a truncated normal distribution,

$$\pi(\beta_k|\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}) = \frac{\exp\{-\frac{\beta_k^2}{2\sigma_k^2}\}}{c(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})}I\left(\beta_k \geq -h_\gamma(\lambda_j, \boldsymbol{\beta}_{(-k)}, \mathbf{Z}_i)\right), \tag{2.5}$$

where the normalizing constant depends on $\boldsymbol{\beta}_{(-k)}$ and $\boldsymbol{\lambda}$, given by

$$c(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}) = \sqrt{2\pi}\sigma_k\left[1 - \Phi\left(-\frac{h_\gamma(\lambda_j, \boldsymbol{\beta}_{(-k)}, \mathbf{Z}_i)}{\sigma_k}\right)\right]. \tag{2.6}$$

Thus, we need only to constrain one parameter β_k to guarantee the nonnegativity of the hazard function and allow the other parameters, $(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})$, to be free.

Although not required for the development, we can take $\boldsymbol{\beta}_{(-k)}$ and $\boldsymbol{\lambda}$ to be independent a priori in (2.4), $\pi(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\beta}_{(-k)})\pi(\boldsymbol{\lambda})$. In addition, we can specify a normal prior distribution for each component of $\boldsymbol{\beta}_{(-k)}$. We assume that the components of $\boldsymbol{\lambda}$ are independent a priori, and each λ_j has a Gamma(α, ξ) distribution.

3. Gibbs sampling

For $0 \leq \gamma \leq 1$, it can be shown that the full conditionals of $(\beta_1, \ldots, \beta_p)$ are log-concave, in which case we only need to use the adaptive rejection sampling (ARS) algorithm proposed by Gilks and Wild [19]. Due to the non-log-concavity of the full conditionals of the λ_j's, a Metropolis step is required within the Gibbs steps, for details see Gilks, Best and Tan [18]. For each Gibbs sampling step, the support for the parameter to be sampled is set to satisfy the constraint (2.3), such that the likelihood function is well defined within the sampling range. For $i = 1, \ldots, n; j = 1, \ldots, J; k = 1, \ldots, p$, the following inequalities need to be satisfied,

$$\beta_k \geq -h_\gamma(\lambda_j, \boldsymbol{\beta}_{(-k)}, \mathbf{Z}_i), \quad \lambda_j \geq -\min_i\{(\gamma \boldsymbol{\beta}' \mathbf{Z}_i)^{1/\gamma}, 0\}.$$

Suppose that the kth component of $\boldsymbol{\beta}$ has a truncated normal prior as given in (2.5), and all other parameters are left free. The full conditionals of the parameters are given as follows:

$$\begin{aligned}
\pi(\beta_k|\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}, D) &\propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\beta_k|\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}) \\
\pi(\beta_l|\boldsymbol{\beta}_{(-l)}, \boldsymbol{\lambda}, D) &\propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\beta_l)/c(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda}) \\
\pi(\lambda_j|\boldsymbol{\beta}, \boldsymbol{\lambda}_{(-j)}, D) &\propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\lambda_j)/c(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})
\end{aligned}$$

where

$$\begin{aligned}
\pi(\beta_l) &\propto \exp\{-\beta_l^2/(2\sigma_l^2)\}, \quad l \neq k, l = 1, \ldots, p, \\
\pi(\lambda_j) &\propto \lambda_j^{\alpha-1}\exp(-\xi\lambda_j), \quad j = 1, \ldots, J.
\end{aligned}$$

These full conditionals have nice tractable structures, since $c(\boldsymbol{\beta}_{(-k)}, \boldsymbol{\lambda})$ has a closed form with our proposed prior specification. Posterior estimation is very robust with respect to the conditioning scheme (the choice of k) in (2.4).

4. Model assessment

It is crucial to compare a class of competing models for a given dataset and select the model that best fits the data. After fitting the proposed models for a set of pre-specified γ's, we compute the CPO and DIC statistics, which are the two commonly used measures of model adequacy [14, 15, 12, 31].

We first introduce the CPO as follows. Let $\mathbf{Z}^{(-i)}$ denote the $(n-1) \times p$ covariate matrix with the ith row deleted, let $\mathbf{y}^{(-i)}$ denote the $(n-1) \times 1$ response vector with y_i deleted, and $\boldsymbol{\nu}^{(-i)}$ is defined similarly. The resulting data with the ith case deleted can be written as $D^{(-i)} = \{(n-1), \mathbf{y}^{(-i)}, \mathbf{Z}^{(-i)}, \boldsymbol{\nu}^{(-i)}\}$. Let $f(y_i|\mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})$ denote the density function of y_i, and let $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})$ denote the posterior density of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ given $D^{(-i)}$. Then, CPO_i is the marginal posterior predictive density of

y_i given $D^{(-i)}$, which can be written as

$$\begin{aligned}\mathrm{CPO}_i &= f(y_i|\mathbf{Z}_i, D^{(-i)}) \\ &= \int\int f(y_i|\mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})d\boldsymbol{\beta}d\boldsymbol{\lambda} \\ &= \left\{\int\int \frac{\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)}{f(y_i|\mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})}d\boldsymbol{\beta}d\boldsymbol{\lambda}\right\}^{-1}.\end{aligned}$$

For the proposed transformation model, a Monte Carlo approximation of CPO_i is given by,

$$\widehat{\mathrm{CPO}}_i = \left\{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_i(\boldsymbol{\beta}_{[m]}, \boldsymbol{\lambda}_{[m]}|y_i, \mathbf{Z}_i, \nu_i)}\right\}^{-1},$$

where

$$\begin{aligned}L_i(\boldsymbol{\beta}_{[m]}, \boldsymbol{\lambda}_{[m]}|y_i, \mathbf{Z}_i, \nu_i) &= \prod_{j=1}^{J}(\lambda_{j,[m]}^{\gamma} + \gamma\boldsymbol{\beta}_{[m]}'\mathbf{Z}_i)^{\delta_{ij}\nu_i/\gamma} \\ &\quad \times \exp\Big[-\delta_{ij}\big\{(\lambda_{j,[m]}^{\gamma} + \gamma\boldsymbol{\beta}_{[m]}'\mathbf{Z}_i)^{1/\gamma}(y_i - s_{j-1}) \\ &\quad + \sum_{g=1}^{j-1}(\lambda_{g,[m]}^{\gamma} + \gamma\boldsymbol{\beta}_{[m]}'\mathbf{Z}_i)^{1/\gamma}(s_g - s_{g-1})\big\}\Big].\end{aligned}$$

Note that M is the number of Gibbs samples after burn-in, and $\boldsymbol{\lambda}_{[m]} = (\lambda_{1,[m]}, \ldots, \lambda_{J,[m]})'$ and $\boldsymbol{\beta}_{[m]}$ are the samples of the mth Gibbs iteration. A common summary statistic based on the CPO_i's is $B = \sum_{i=1}^{n}\log(\mathrm{CPO}_i)$, which is often called the logarithm of the pseudo Bayes factor. A larger value of B indicates a better fit of a model.

Another model assessment criterion is the DIC (Spiegelhalter et al. [31]), defined as

$$\mathrm{DIC} = 2\overline{\mathrm{Dev}(\boldsymbol{\beta}, \boldsymbol{\lambda})} - \mathrm{Dev}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\lambda}}),$$

where $\mathrm{Dev}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -2\log L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)$ is the deviance, and $\overline{\mathrm{Dev}(\boldsymbol{\beta}, \boldsymbol{\lambda})}$, $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\lambda}}$ are the corresponding posterior means. Specifically, in our proposed model,

$$\mathrm{DIC} = -\frac{4}{M}\sum_{m=1}^{M}\log L(\boldsymbol{\beta}_{[m]}, \boldsymbol{\lambda}_{[m]}|D) + 2\log L(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\lambda}}|D).$$

The smaller the DIC value, the better the fit of the model.

5. Numerical studies

5.1. Application

As an illustration, we applied the transformation models to the E1690 data. There were a total of $n = 427$ patients on these combined treatment arms. The covariates in this analysis were treatment (high-dose interferon or observation), age (a continuous variable which ranged from 19.13 to 78.05 with mean 47.93 years), sex (male or female) and nodal category (1 if there were no positive nodes, or 2 otherwise). Figure 2 shows the estimated cumulative hazard curves for the interferon and observation groups based on the Nelson–Aalen estimator.

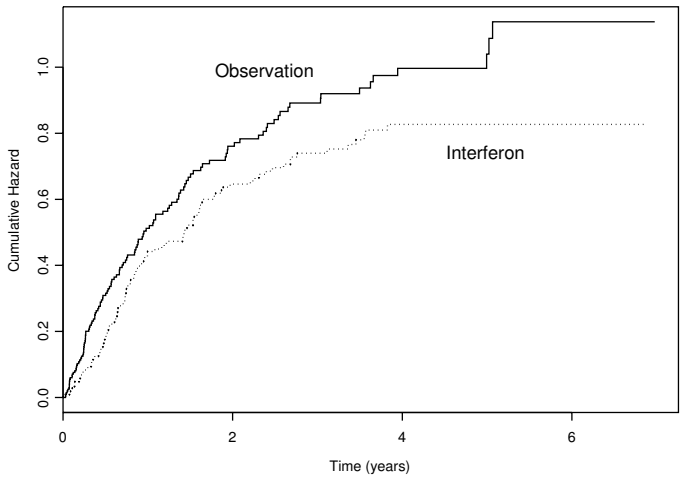


FIG 2. *The estimated cumulative hazard curves for the two arms in E1690*

TABLE 1
The B/DIC statistics with respect to γ and J in the E1690 data

		J		
		1	5	10
	0	**−567.43/1129.19**	−528.36/1051.84	−555.46/1105.48
	.25	−567.96/1131.71	−523.74/1045.68	−534.57/1066.86
γ	.5	−568.47/1133.72	**−522.55/1043.64**	−529.13/1056.44
	.75	−568.89/1135.16	−522.66/1043.86	−527.47/1053.17
	1	−569.46/1136.54	−523.04/1044.84	**−526.80/1052.06**

We constrained the regression coefficient for treatment, β_1, to have the truncated normal prior. We prespecified $\gamma = (0, .25, .5, .75, 1)$ and took the priors for $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$ to be noninformative. For example, $(\beta_1|\boldsymbol{\lambda}, \boldsymbol{\beta}_{(-1)})$ was assigned the truncated $N(0, 10,000)$ prior as defined in (2.5), $(\beta_l,\ l = 2, 3, 4)$ were taken to have independent $N(0, 10,000)$ prior distributions, and $\lambda_j \sim$ Gamma$(2, .01)$, and independent for $j = 1, \ldots, J$. To allow for a fair comparison between different models using different γ's, we used the same noninformative priors across all the targeted models.

The shape of the baseline hazard function is controlled by J. The finer the partition of the time axis, the more general the pattern of the hazard function that is captured. However, by increasing J, we introduce more unknown parameters (the λ_j's). For the proposed transformation model, γ also directly affects the shape of the hazard function, and specifically, there is much interplay between J and γ in controlling the shape of the hazard, and in some sense γ and J are somewhat confounded. Thus when searching for the best fitting model, we must find suitable J and γ simultaneously. Similar to a grid search, we set $J = (1, 5, 10)$, and located the point (J, γ) that yielded the largest B statistic and the smallest DIC.

After a burn-in of 2,000 samples and thinned by 5 iterations, the posterior computations were based on 10,000 Gibbs samples. The B and DIC statistics for model selection are summarized in Table 1. The two model selection criteria are quite consistent with each other, and both lead to the same best model with $J = 5$ and $\gamma = .5$. Table 2 summarizes the posterior means, standard deviations and the 95% highest posterior density (HPD) intervals for $\boldsymbol{\beta}$ using $J = (1, 5, 10)$ and $\gamma = (0, .5, 1)$. For the best model (with $J = 5$ and $\gamma = .5$), we see that the treatment effect has a 95% HPD interval that does not include 0, confirming that treatment with high-dose

TABLE 2
Posterior means, standard deviations, and 95% HPD intervals for the E1690 data

J	γ	Covariate	Mean	SD	95% HPD Interval
1	0	Treatment	−.2888	.1299	(−.5369, −.0310)
		Age	.0117	.0050	(.0016, .0214)
		Sex	−.3479	.1375	(−.6372, −.0962)
		Nodal Category	.5267	.1541	(.2339, .8346)
	.5	Treatment	−.1398	.0626	(−.2588, −.0111)
		Age	.0056	.0024	(.0011, .0103)
		Sex	−.1464	.0644	(−.2791, −.0254)
		Nodal Category	.2179	.0688	(.0835, .3529)
	1	Treatment	−.0655	.0299	(−.1245, −.0078)
		Age	.0026	.0011	(.0004, .0047)
		Sex	−.0593	.0293	(−.1155, −.0007)
		Nodal Category	.0863	.0296	(.0304, .1471)
5	0	Treatment	−.4865	.1295	(−.7492, −.2408)
		Age	−.0036	.0050	(−.0133, .0061)
		Sex	−.4423	.1421	(−.7196, −.1684)
		Nodal Category	.1461	.1448	(−.1307, .4298)
	.5	Treatment	−.1835	.0626	(−.3066, −.0604)
		Age	.0017	.0024	(−.0030, .0064)
		Sex	−.1557	.0655	(−.2853, −.0310)
		Nodal Category	.1141	.0685	(−.0179, .2510)
	1	Treatment	−.0525	.0274	(−.1058, .0007)
		Age	.0011	.0009	(−.0006, .0027)
		Sex	−.0334	.0249	(−.0818, .0148)
		Nodal Category	.0265	.0224	(−.0169, .0705)
10	0	Treatment	−.7238	.1260	(−.9639, −.4710)
		Age	−.0175	.0047	(−.0269, −.0084)
		Sex	−.6368	.1439	(−.9158, −.3544)
		Nodal Category	.1685	.1302	(−.4184, .0859)
	.5	Treatment	−.2272	.0629	(−.3581, −.1094)
		Age	−.0009	.0023	(−.0056, .0035)
		Sex	−.1791	.0649	(−.3094, −.0546)
		Nodal Category	.0534	.0670	(−.0814, .1798)
	1	Treatment	−.0610	.0274	(−.1142, −.0070)
		Age	.0006	.0008	(−.0010, .0021)
		Sex	−.0334	.0256	(−.0850, .0155)
		Nodal Category	.0107	.0225	(−.0325, .0569)

interferon indeed substantially reduced the risk of melanoma relapse compared to observation.

In Figure 3, we present the estimated hazards for the interferon and observation arms for $\gamma = 0, .5$ and 1 using $J = 5$. It is important to note that, when $\gamma = .5$, the hazard ratio increases over time while the hazard difference decreases.

The proportional hazards model yields a hazard ratio of 1.63, the additive hazards model gives a hazard difference of .05, and the model with $\gamma = .5$ shows hazard ratios of 1.27, 1.36 and 1.61, and hazard differences of .14, .11 and .07 at .5, 1 and 3 years, respectively. This interesting feature between the hazards cannot be captured through a conventional modeling structure. An opposite phenomenon in which the difference of the hazards increases in t whereas their ratio decreases, was noted in the British doctors study (Breslow and Day [6], p.112, pp. 336-338), which examined the effects of cigarette smoking on mortality. We also computed the half year and one year posterior predictive survival probabilities for a 48 years old male patient under the high-dose interferon treatment with one or more positive nodes. When $\gamma = .5$, the .5 year posterior predictive survival probabilities are .8578, .7686 and .7804 for $J = 1$, 5 and 10; the 1 year survival probabilities are .7357, .6043 and .6240, respectively. When J is large enough, the posterior inference becomes stable.

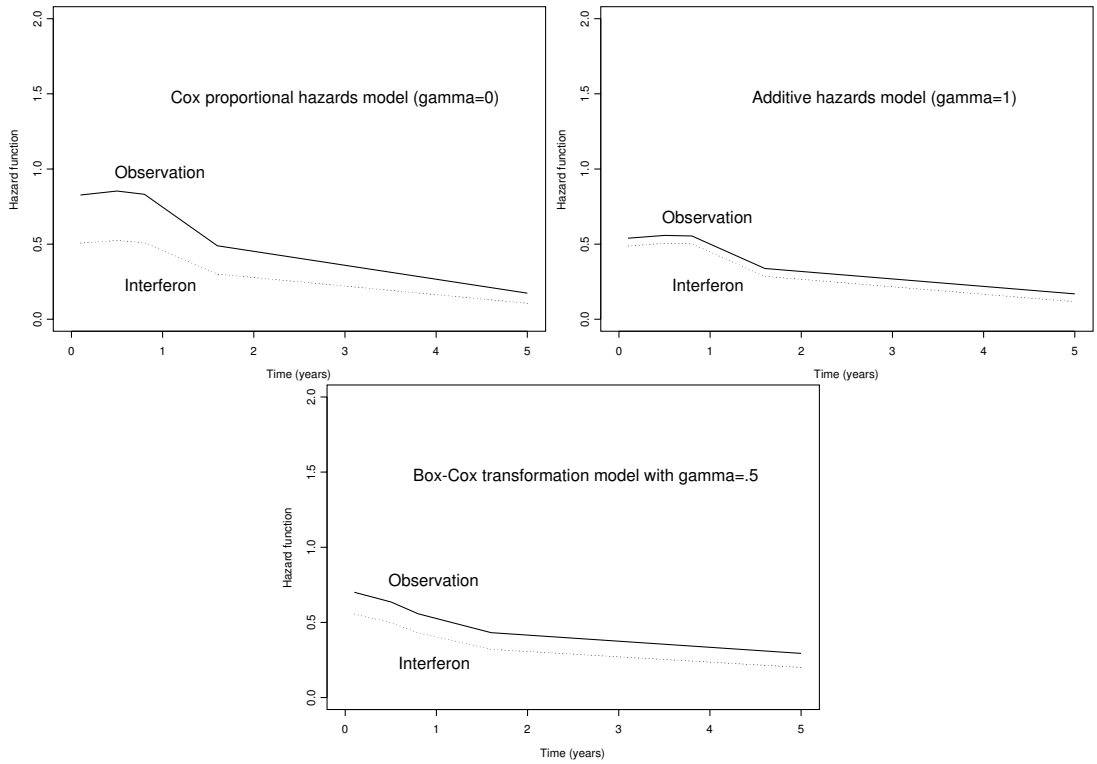


FIG 3. *Estimated hazards under models with* $\gamma = 0$, .5 *and* 1, *for male subjects at age*= 47.93 *years and with one or more positive nodes, using* $J = 5$.

TABLE 3
Sensitivity analysis with β_k *having a truncated normal prior using* $J = 5$ *and* $\gamma = .5$

Truncated Covariate	Regression Coefficient	Mean	SD	95% HPD Interval
Age	Treatment	−.1862	.0633	(−.3122, −.0627)
	Age	.0016	.0024	(−.0032, .0063)
	Sex	−.1551	.0665	(−.2802, −.0187)
	Nodal Category	.1132	.0697	(−.0229, .2511)
Sex	Treatment	−.1883	.0634	(−.3107, −.0592)
	Age	.0017	.0024	(−.0032, .0063)
	Sex	−.1572	.0651	(−.2801, −.0296)
	Nodal Category	.1131	.0672	(−.0165, .2448)
Nodal Category	Treatment	−.1850	.0633	(−.3037, −.0566)
	Age	.0017	.0024	(−.0030, .0062)
	Sex	−.1519	.0662	(−.2819, −.0236)
	Nodal Category	.1124	.0679	(−.0223, .2416)

We examined MCMC convergence based on the method proposed by Geweke [17]. The Markov chains mixed well and converged fast. We conducted a sensitivity analysis on the choice of the conditioning scheme in the prior (2.5) by choosing the regression coefficient of each covariate to have a truncated normal prior. The results in Table 3 show the robustness of the model to the choice of the constrained parameter in the prior specification. This demonstrates the appealing feature of the proposed prior specification, which thus facilitates an attractive computational procedure.

5.2. Simulation

We conducted a simulation study to examine properties of the proposed model. The failure times were generated from model (2.1) with $\gamma = .5$. We assumed a constant

TABLE 4
Simulation results based on 500 replications, with the true values $\beta_1 = .7$ and $\beta_2 = 1$

n	$c\%$	Mean (β_1)	SD (β_1)	Mean (β_2)	SD (β_2)
300	0	.7705	.2177	1.0556	.4049
	25	.7430	.2315	1.0542	.4534
500	0	.7424	.1989	1.0483	.3486
	25	.7510	.2084	1.0503	.3781
1000	0	.7273	.1784	1.0412	.2920
	25	.7394	.1869	1.0401	.3100

baseline hazard, i.e., $\lambda_0(t) = .5$, and two covariates were generated independently: $Z_1 \sim N(5, 1)$ and Z_2 is a binary random variable taking a value of 1 or 2 with probability .5. The corresponding regression parameters were $\beta_1 = .7$ and $\beta_2 = 1$. The censoring times were simulated from a uniform distribution to achieve approximately a 25% censoring rate. The sample sizes were $n = 300$, 500 and 1,000, and we replicated 500 simulations for each configuration.

Noninformative prior distributions were specified for the unknown parameters as in the E1690 example. For each Markov chain, we took a burn-in of 200 samples and the posterior estimates were based on 5,000 Gibbs samples. The posterior means and standard deviations are summarized in Table 4, which show the convergence of the posterior means of the parameters to the true values. As the sample size increases, the posterior means of β_1 and β_2 approach their true values and the corresponding standard deviations decrease. As the censoring rate increases, the posterior standard deviation also increases.

6. Discussion

We have proposed a class of survival models based on the Box–Cox transformed hazard functions. This class of transformation models makes hazard-based regression more flexible, general, and versatile than other methods, and opens a wide family of relationships between the hazards. Due to the complexity of the model, we have proposed a joint prior specification scheme by absorbing the non-linear constraint into one parameter while leaving all the other parameters free of constraints. This prior specification is quite general and can be applied to a much broader class of constrained parameter problems arising from regression models. It is usually difficult to interpret the parameters in the proposed model except when $\gamma = 0$ or 1. However, if the primary aim is for prediction of survival, the best fitting Box–Cox transformation model could be useful.

Acknowledgements

We would like to thank Professor Javier Rojo and anonymous referees for helpful comments which led to great improvement of the article.

References

[1] ANDERSEN, P. K. AND GILL, R. D. (1982). Cox's regression model for counting processes: A large-sample study. *Ann. Statist.* **10**, 1100–1120.
[2] ARANDA-ORDAZ, F. J. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics* **39**, 109–117.

[3] Barlow, W. E. (1985). General relative risk models in stratified epidemiologic studies. *Appl. Statist.* **34**, 246–257.
[4] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26**, 211–252.
[5] Breslow, N. E. (1985). Cohort analysis in epidemiology. In *A Celebration of Statistics* (A. C. Atkinson and S. E. Fienberg, eds.). Springer, New York, 109–143.
[6] Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research*, 2, *The Design and Analysis of Case-Control Studies*, IARC, Lyon.
[7] Breslow, N. E. and Storer, B. E. (1985). General relative risk functions for case-control studies. *Amer. J. Epidemi.* **122**, 149–162.
[8] Chen, M. and Shao, Q. (1998). Monte Carlo methods for Bayesian analysis of constrained parameter problems. *Biometrika* **85**, 73–87.
[9] Chen, M., Shao, Q. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.
[10] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.
[11] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.
[12] Dey, D. K., Chen, M. and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**, 1239–1252.
[13] Foster, A. M., Tian, L. and Wei, L. J. (2001). Estimation for the Box–Cox transformation model without assuming parametric error distribution. *J. Amer. Statist. Assoc.* **96**, 1097–1101.
[14] Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall, London.
[15] Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press, Oxford, 147–167.
[16] Gelfand, A. E., Smith, A. F. M. and Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87**, 523–532.
[17] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press, Oxford, 169–193.
[18] Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Statist.* **44**, 455–472.
[19] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–348.
[20] Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.
[21] Ibrahim, J. G., Chen, M. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.
[22] Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc. Ser. B* **40**, 214–221.
[23] Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M. and Blum, R. H. (2000). High- and low-dose interferon Alfa-2b in high-risk melanoma: first analysis of intergroup trial E1690/S9111/C9190. *J. Clinical*

Oncology **18**, 2444–2458.

[24] Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

[25] Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann. Statist.* **23**, 1712–1734.

[26] Martinussen, T and Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283–298.

[27] Nieto-Barajas, L. E. and Walker, S. G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scand. J. Statist.* **29**, 413–424.

[28] O'Neill, T. J. (1986). Inconsistency of the misspecified proportional hazards model. *Statist. Probab. Lett.* **4**, 219-22.

[29] Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician* **41**, 169–178.

[30] Scheike, T. H. and Zhang, M.-J. (2002). An additive-multiplicative Cox–Aalen regression model. *Scand. J. Statist.* **29**, 75–88.

[31] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. Ser. B* **64**, 583–616.

[32] Yin, G. and Ibrahim, J. (2005). A general class of Bayesian survival models with zero and non-zero cure fractions. *Biometrics* **61**, 403–412.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 183–209
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000455

Characterizations of joint distributions, copulas, information, dependence and decoupling, with applications to time series

Victor H. de la Peña[1,*], **Rustam Ibragimov**[2,†] **and Shaturgun Sharakhmetov**[3]

Columbia University, Harvard University and Tashkent State Economics University

Abstract: In this paper, we obtain general representations for the joint distributions and copulas of arbitrary dependent random variables absolutely continuous with respect to the product of given one-dimensional marginal distributions. The characterizations obtained in the paper represent joint distributions of dependent random variables and their copulas as sums of U-statistics in independent random variables. We show that similar results also hold for expectations of arbitrary statistics in dependent random variables. As a corollary of the results, we obtain new representations for multivariate divergence measures as well as complete characterizations of important classes of dependent random variables that give, in particular, methods for constructing new copulas and modeling different dependence structures.

The results obtained in the paper provide a device for reducing the analysis of convergence in distribution of a sum of a double array of dependent random variables to the study of weak convergence for a double array of their independent copies. Weak convergence in the dependent case is implied by similar asymptotic results under independence together with convergence to zero of one of a series of dependence measures including the multivariate extension of Pearson's correlation, the relative entropy or other multivariate divergence measures. A closely related result involves conditions for convergence in distribution of m-dimensional statistics $h(X_t, X_{t+1}, \ldots, X_{t+m-1})$ of time series $\{X_t\}$ in terms of weak convergence of $h(\xi_t, \xi_{t+1}, \ldots, \xi_{t+m-1})$, where $\{\xi_t\}$ is a sequence of independent copies of X_t's, and convergence to zero of measures of intertemporal dependence in $\{X_t\}$. The tools used include new sharp estimates for the distance between the distribution function of an arbitrary statistic in dependent random variables and the distribution function of the statistic in independent copies of the random variables in terms of the measures of dependence of the random variables. Furthermore, we obtain new sharp complete decoupling moment and probability inequalities for dependent random variables in terms of their dependence characteristics.

*Supported in part by NSF grants DMS/99/72237, DMS/02/05791, and DMS/05/05949.

†Supported in part by a Yale University Graduate Fellowship; the Cowles Foundation Prize; and a Carl Arvid Anderson Prize Fellowship in Economics.

[1]Department of Statistics, Columbia University, Mail Code 4690, 1255 Amsterdam Avenue, New York, NY 10027, e-mail: vp@stat.columbia.edu

[2]Department of Economics, Harvard University, 1805 Cambridge St., Cambridge, MA 02138, e-mail: ribragim@fas.harvard.edu

[3]Department of Probability Theory, Tashkent State Economics University, ul. Uzbekistanskaya, 49, Tashkent, 700063, Uzbekistan, e-mail: tim001@tseu.silk.org

AMS 2000 subject classifications: primary 62E10, 62H05, 62H20; secondary 60E05, 62B10, 62F12, 62G20.

Keywords and phrases: joint distribution, copulas, information, dependence, decoupling, convergence, relative entropy, Kullback–Leibler and Shannon mutual information, Pearson coefficient, Hellinger distance, divergence measures.

1. Introduction

In recent years, a number of studies in statistics, economics, finance and risk management have focused on dependence measuring and modeling and testing for serial dependence in time series. It was observed in several studies that the use of the most widely applied dependence measure, the correlation, is problematic in many setups. For example, Boyer, Gibson and Loretan [9] reported that correlations can provide little information about the underlying dependence structure in the cases of asymmetric dependence. Naturally (see, e.g., Blyth [7] and Shaw [71]), the linear correlation fails to capture nonlinear dependencies in data on risk factors. Embrechts, McNeil and Straumann [22] presented a rigorous study concerning the problems related to the use of correlation as measure of dependence in risk management and finance. As discussed in [22] (see also Hu [32]), one of the cases when the use of correlation as measure of dependence becomes problematic is the departure from multivariate normal and, more generally, elliptic distributions. As reported by Shaw [71], Ang and Chen [4] and Longin and Solnik [54], the departure from Gaussianity and elliptical distributions occurs in real world risks and financial market data. Some of the other problems with using correlation is that it is a bivariate measure of dependence and even using its time varying versions, at best, leads to only capturing the pairwise dependence in data sets, failing to measure more complicated dependence structures. In fact, the same applies to other bivariate measures of dependence such as the bivariate Pearson coefficient, Kullback-Leibler and Shannon mutual information, or Kendall's tau. Also, the correlation is defined only in the case of data with finite second moments and its reliable estimation is problematic in the case of infinite higher moments. However, as reported in a number of studies (see, e.g., the discussion in Loretan and Phillips [55], Cont [11] and Ibragimov [33, 34] and references therein), many financial and commodity market data sets exhibit heavy-tailed behavior with higher moments failing to exist and even variances being infinite for certain time series in finance and economics. A number of frameworks have been proposed to model heavy-tailedness phenomena, including stable distributions and their truncated versions, Pareto distributions, multivariate t-distributions, mixtures of normals, power exponential distributions, ARCH processes, mixed diffusion jump processes, variance gamma and normal inverse Gamma distributions (see [11, 33, 34] and references therein), with several recent studies suggesting modeling a number of financial time series using distributions with "semiheavy tails" having an exponential decline (e.g., Barndorff–Nielsen and Shephard [5] and references therein). The debate concerning the values of the tail indices for different heavy-tailed financial data and on appropriateness of their modeling based on certain above distributions is, however, still under way in empirical literature. In particular, as discussed in [33, 34], a number of studies continue to find tail parameters less than two in different financial data sets and also argue that stable distributions are appropriate for their modeling.

Several approaches have been proposed recently to deal with the above problems. For example, Joe [42, 43] proposed multivariate extensions of Pearson's coefficient and the Kullback–Leibler and Shannon mutual information. A number of papers have focused on statistical and econometric applications of mutual information and other dependence measures and concepts (see, among others, Lehmann [52], Golan [26], Golan and Perloff [27], Massoumi and Racine [57], Miller and Liu [58], Soofi and Retzer [73] and Ullah [76] and references therein). Several recent papers in econometrics (e.g., Robinson [66], Granger and Lin [29] and Hong and White [31]) considered problems of estimating entropy measures of serial dependence in time

series. In a study of multifractals and generalizations of Boltzmann-Gibbs statistics, Tsallis [75] proposed a class of generalized entropy measures that include, as a particular case, the Hellinger distance and the mutual information measure. The latter measures were used by Fernandes and Flôres [24] in testing for conditional independence and noncausality. Another approach, which is also becoming more and more popular in econometrics and dependence modeling in finance and risk management is the one based on copulas. Copulas are functions that allow one, by a celebrated theorem due to Sklar [72], to represent a joint distribution of random variables (r.v.'s) as a function of marginal distributions (see Section 3 for the formulation of the theorem). Copulas, therefore, capture all the dependence properties of the data generating process. In recent years, copulas and related concepts in dependence modeling and measuring have been applied to a wide range of problems in economics, finance and risk management (e.g., Taylor [74], Fackler [23], Frees, Carriere and Valdez [25], Klugman and Parsa [46], Patton [61, 62], Richardson, Klose and Gray [65], Embrechts, Lindskog and McNeil [21], Hu [32], Reiss and Thomas [64], Granger, Teräsvirta and Patton [30] and Miller and Liu [58]). Patton [61] studied modeling time-varying dependence in financial markets using the concept of conditional copula. Patton [62] applied copulas to model asymmetric dependence in the joint distribution of stock returns. Hu [32] used copulas to study the structure of dependence across financial markets. Miller and Liu [58] proposed methods for recovery of multivariate joint distributions and copulas from limited information using entropy and other information theoretic concepts.

The multivariate measures of dependence and the copula-based approaches to dependence modeling are two interrelated parts of the study of joint distributions of r.v.'s in mathematical statistics and probability theory. A problem of fundamental importance in the field is to determine a relationship between a multivariate cumulative distribution function (cdf) and its lower dimensional margins and to measure degrees of dependence that correspond to particular classes of joint cdf's. The problem is closely related to the problem of characterizing the joint distribution by conditional distributions (see Gouriéroux and Monfort [28]). Remarkable advances have been made in the latter research area in recent years in statistics and probability literature (see, e.g., papers in Dall'Aglio, Kotz and Salinetti [13], Beneš and Štěpán [6] and the monographs by Joe [44], Nelsen [60] and Mari and Kotz [56]).

Motivated by the recent surge in the interest in the study and application of dependence measures and related concepts to account for the complexity in problems in statistics, economics, finance and risk management, this paper provides the first characterizations of joint distributions and copulas for multivariate vectors. These characterizations represent joint distributions of dependent r.v.'s and their copulas as sums of U-statistics in independent r.v.'s. We use these characterizations to introduce a unified approach to modeling multivariate dependence and provide new results concerning convergence of multidimensional statistics of time series. The results provide a device for reducing the analysis of convergence of multidimensional statistics of time series to the study of convergence of the measures of intertemporal dependence in the time series (e.g., the multivariate Pearson coefficient, the relative entropy, the multivariate divergence measures, the mean information for discrimination between the dependence and independence, the generalized Tsallis entropy and the Hellinger distance). Furthermore, they allow one to reduce the problems of the study of convergence of statistics of intertemporally dependent time series to the study of convergence of corresponding statistics in the case of intertemporally independent time series. That is, the characterizations for copulas obtained in the

paper imply results which associate with each set of arbitrarily dependent r.v.'s a sum of U-statistics in independent r.v.'s with canonical kernels. Thus, they allow one to reduce problems for dependent r.v.'s to well-studied objects and to transfer results known for independent r.v.'s and U-statistics to the case of arbitrary dependence (see, e.g., Ibragimov and Sharakhmetov [36-40], Ibragimov, Sharakhmetov and Cecen [41], de la Peña, Ibragimov and Sharakhmetov [16, 17] and references therein for general moment inequalities for sums of U-statistics and their particular important cases, sums of r.v.'s and multilinear forms, and Ibragimov and Phillips [35] for a new and conceptually simple method for obtaining weak convergence of multilinear forms, U-statistics and their non-linear analogues to stochastic integrals based on general asymptotic theory for semimartingales and for applications of the method in a wide range of linear and non-linear time series models).

As a corollary of the results for copulas, we obtain new complete characterizations of important classes of dependent r.v.'s that give, in particular, methods for constructing new copulas and modeling various dependence structures. The results in the paper provide, among others, complete positive answers to the problems raised by Kotz and Seeger [47] concerning characterizations of density weighting functions (d.w.f.) of dependent r.v.'s, existence of methods for constructing d.w.f.'s, and derivation of d.w.f.'s for a given model of dependence (see also [58] for a discussion of d.w.f.'s).

Along the way, a general methodology (of intrinsic interest within and outside probability theory, economics and finance) is developed for analyzing key measures of dependence among r.v.'s. Using the methodology, we obtain sharp decoupling inequalities for comparing the expectations of arbitrary (integrable) functions of dependent variables to their corresponding counterparts with independent variables through the inclusion of multivariate dependence measures.

On the methodological side, the paper shows how the results in theory of U-statistics, including inversion formulas for these objects that provide the main tools for the argument for representations in this paper (see the proof of Theorem 1), can be used in the study of joint distributions, copulas and dependence.

The paper is organized as follows. Sections 2 and 3 contain the results on general characterizations of copulas and joint distributions of dependent r.v.'s. Section 4 presents the results on characterizations of dependence based on U-statistics in independent r.v.'s. In Sections 5 and 6, we apply the results for copulas and joint distributions to characterize different classes of dependent r.v.'s. Section 7 contains the results on reduction of the analysis of convergence of multidimensional statistics of time series to the study of convergence of the measures of intertemporal dependence in time series as well as the results on sharp decoupling inequalities for dependent r.v.'s. The proofs of the results obtained in the paper are in the Appendix.

2. General characterizations of joint distributions of arbitrarily dependent random variables

In the present section, we obtain explicit general representations for joint distributions of arbitrarily dependent r.v.'s absolutely continuous with respect to products of marginal distributions. Let $F_k : \mathbf{R} \to [0,1]$, $k = 1, \ldots, n$, be one-dimensional cdf's and let $\xi_1, \ldots, \xi_n$ be independent r.v.'s on some probability space $(\Omega, \Im, P)$ with $P(\xi_k \leq x_k) = F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, \ldots, n$ (we formulate the results for the case of right-continuous cdf's; however, completely similar results hold in the left-continuous case).

In what follows, $F(x_1,\dots,x_n)$, $x_i \in \mathbf{R}$, $i=1,\dots,n$, stands for a function satisfying the following conditions:

(a) $F(x_1,\dots,x_n) = P(X_1 \le x_1,\dots,X_n \le x_n)$ for some r.v.'s $X_1,\dots,X_n$ on a probability space $(\Omega,\Im,P)$;

(b) the one-dimensional marginal cdf's of F are $F_1,\dots,F_n$;

(c) F is absolutely continuous with respect to $dF(x_1)\cdots dF_n(x_n)$ in the sense that there exists a Borel function $G:\mathbf{R}^n \to [0,\infty)$ such that

$$F(x_1,\dots,x_n) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} G(t_1,\dots,t_n)dF_1(t_1)\cdots dF_n(t_n).$$

As usual, throughout the paper, we denote G in (c) by $\frac{dF}{dF_1\cdots dF_n}$. In addition, $F(x_{j_1},\dots,x_{j_k})$, $1 \le j_1 < \cdots < j_k \le n$, $k=2,\dots,n$, stands for the k-dimensional marginal cdf of $F(x_1,\dots,x_n)$. Also, in what follows, if not stated otherwise, $\frac{dF(x_{j_1},\dots,x_{j_k})}{dF_{j_1}\cdots dF_{j_k}}$, $1 \le j_1 < \cdots < j_k \le n$, $k=2,\dots,n$, is to be taken to be 1 if at least one $x_{j_1},\dots,x_{j_k}$ is not a point of increase of the corresponding $F_{j_1},\dots,F_{j_k}$ (that is, if $(x_{j_1},\dots,x_{j_k})$ is not in the support of $dF_{j_1}\cdots dF_{j_k}$).

Throughout the paper, the functions g appearing in the representations obtained are assumed to be Borel measurable.

Theorem 2.1. *A function $F:\mathbf{R}^n \to [0,1]$ is a joint cdf with one-dimensional marginal cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k=1,\dots,n$, absolutely continuous with respect to the product of marginal cdf's $\prod_{k=1}^n F_k(x_k)$, if and only if there exist functions $g_{i_1,\dots,i_c}:\mathbf{R}^c \to \mathbf{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c=2,\dots,n$, satisfying conditions*

A1 (integrability):

$$E|g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c})| < \infty,$$

A2 (degeneracy)*:*

$$E(g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_{k-1}},\xi_{i_k},\xi_{i_{k+1}},\dots,\xi_{i_c})|\xi_{i_1},\dots,\xi_{i_{k-1}},\xi_{i_{k+1}},\dots,\xi_{i_c}) =$$

$$\int_{-\infty}^{\infty} g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_{k-1}},x_{i_k},\xi_{i_{k+1}},\dots,\xi_{i_c})dF_{i_k}(x_{i_k}) = 0,\ (a.s.)$$

$1 \le i_1 < \cdots < i_c \le n$, $k=1,2,\dots,c$, $c=2,\dots,n$,

A3 (positive definiteness)*:*

$$U_n(\xi_1,\dots,\xi_n) \equiv \sum_{c=2}^{n}\sum_{1\le i_1<\cdots<i_c\le n} g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}) \ge -1\ (a.s.)$$

and such that the following representation holds for F :

$$F(x_1,\dots,x_n) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}(1+U_n(t_1,\dots,t_n))\prod_{i=1}^{n}dF_i(t_i). \tag{2.1}$$

Moreover, $g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}) = f_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c})$ (a.s.), $1 \le i_1 < \cdots < i_c \le n$, $c=2,\dots,n$, where

$$f_{i_1,\dots,i_c}(x_{i_1},\dots,x_{i_c}) = \sum_{k=2}^{c}(-1)^{c-k}\sum_{1\le j_1<\cdots<j_k\in\{i_1,\dots,i_c\}}\left(\frac{dF(x_{j_1},\dots,x_{j_k})}{dF_{j_1}\cdots dF_{j_k}}-1\right).$$

Remark 2.1. It is not difficult to see that if r.v.'s $X_1, \dots, X_n$ have a joint cdf given by (2.1) then the r.v.'s $X_{j_1}, \dots, X_{j_k}$, $1 \le j_1 < \cdots < j_k \le n$, $k = 2, \dots, n$, have the joint cdf

$$F(x_{j_1}, \dots, x_{j_k}) = \int_{-\infty}^{x_{j_1}} \cdots \int_{-\infty}^{x_{j_k}} (1 + \sum_{c=2}^{n} \sum_{\{i_1 < \dots < i_c\} \in B_k} g_{i_1,\dots,i_c}(t_{i_1}, \dots, t_{i_c})) \prod_{i=1}^{k} dF_{j_i}(t_{j_i})$$

with the same functions $g_{i_1,\dots,i_c}$, and where $B_k = \{j_1, \dots, j_k\}$.

Theorem 2.1 can be equivalently formulated as in the following remark.

Remark 2.2. A function $F : \mathbf{R}^n \to [0,1]$ is a joint cdf with the one-dimensional marginal cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, \dots, n$, absolutely continuous with respect to the product of marginal cdf's $\prod_{k=1}^{n} F_k(x_k)$, if and only if there exist functions $g_{i_1,\dots,i_c} : \mathbf{R}^c \to \mathbf{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \dots, n$, satisfying conditions A1–A3 and such that the element of frequency $dF(x_1, \dots, x_n)$ can be expressed in the form

$$dF(x_1, \dots, x_n) = \prod_{i=1}^{n} dF_i(x_i)(1 + U_n(x_1, \dots, x_n)). \tag{2.2}$$

Remark 2.3. Sharakhmetov [69] provided proof of (2.2) in the case of density functions (of r.v.'s absolutely continuous with respect to *Lebesgue* measure), with a mention that a similar representation holds for distributions of discrete r.v.'s. The setup considered in this paper includes (among others) the class of vectors of dependent absolutely continuous and discrete r.v.'s as well as vectors of mixtures of absolutely continuous and discrete r.v.'s. Furthermore, our proof easily extends to the case of general Banach spaces, in particular, the spaces $\mathbf{R}^k$.

3. Applications to copulas

Let us start with the definition of copulas and the formulation of Sklar's theorem mentioned in the introduction (see e.g., [22] and [60]).

Definition 3.1. A function $C : [0,1]^n \to [0,1]$ is called a n-dimensional *copula* if it satisfies the following conditions:

1. $C(u_1, \dots, u_n)$ is increasing in each component u_i.
2. $C(u_1, \dots, u_{k-1}, 0, u_{k+1}, \dots, u_n) = 0$ for all $u_i \in [0,1]$, $i \ne k$, $k = 1, \dots, n$.
3. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $u_i \in [0,1]$, $i = 1, \dots, n$.
4. For all $(a_1, \dots, a_n)$, $(b_1, \dots, b_n) \in [0,1]^n$ with $a_i \le b_i$,

$$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1 + \cdots + i_n} C(x_{1i_1}, \dots, x_{ni_n}) \ge 0,$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ for all $j \in \{1, \dots, n\}$. Equivalently, C is a n-dimensional copula if it is a joint cdf of n r.v.'s each of which is uniformly distributed on $[0,1]$.

Definition 3.2. A copula $C : [0,1]^n \to [0,1]$ is called *absolutely continuous* if, when considered as a joint cdf, it has a joint density given by $\partial C^n(u_1 \dots, u_n)/\partial u_1 \cdots \partial u_n$.

Theorem 3.1 (Sklar [72]). *If $X_1, \dots, X_n$ are random variables defined on a common probability space, with the one-dimensional cdf's $F_{X_k}(x_k) = P(X_k \le x_k)$ and the joint cdf $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$, then there exists an n-dimensional copula $C_{X_1,\dots,X_n}(u_1, \dots, u_n)$ such that $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = C_{X_1,\dots,X_n}(F_{X_1}(x_1), \dots, F_{X_n}(x_n))$ for all $x_k \in \mathbf{R}$, $k = 1, \dots, n$.*

The following theorems give analogues of the representations in the previous section for copulas. Let $V_1, \dots, V_n$ denote independent r.v.'s uniformly distributed on $[0, 1]$.

Theorem 3.2. *A function $C : [0,1]^n \to [0,1]$ is an absolutely continuous n-dimensional copula if and only if there exist functions $\tilde{g}_{i_1,\dots,i_c} : \mathbf{R}^c \to \mathbf{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, ..., n$, satisfying the conditions*
A4 (integrability)*:*

$$\int_0^1 \cdots \int_0^1 |\tilde{g}_{i_1,\dots,i_c}(t_{i_1}, \dots, t_{i_c})| dt_{i_1} \cdots dt_{i_c} < \infty,$$

A5 (degeneracy)*:*

$$E(\tilde{g}_{i_1,\dots,i_c}(V_{i_1}, \dots, V_{i_{k-1}}, V_{i_k}, V_{i_{k+1}}, \dots, V_{i_c}) | V_{i_1}, \dots, V_{i_{k-1}}, V_{i_{k+1}}, \dots, V_{i_c})$$
$$= \int_0^1 \tilde{g}_{i_1,\dots,i_c}(V_{i_1}, \dots, V_{i_{k-1}}, t_{i_k}, V_{i_{k+1}}, \dots, V_{i_c}) dt_{i_k} = 0 \ (a.s.),$$

$1 \le i_1 < \cdots < i_c \le n$, $k = 1, 2, \dots, c$, $c = 2, \dots, n$,

A6 (positive definiteness):

$$\tilde{U}_n(V_1, \dots, V_n) \equiv \sum_{c=2}^{n} \sum_{1 \le i_1 < \cdots < i_c \le n} \tilde{g}_{i_1,\dots,i_c}(V_{i_1}, \dots, V_{i_c}) \ge -1 \ (a.s.)$$

and such that

$$C(u_1, \dots, u_n) = \int_0^{u_1} \cdots \int_0^{u_n} (1 + \tilde{U}_n(t_1, \dots, t_n)) \prod_{i=1}^{n} dt_i. \tag{3.1}$$

Theorem 3.2 and Sklar's theorem formulated above imply the following representation for a joint distribution of r.v.'s.

Theorem 3.3. *A function $F : \mathbf{R}^n \to [0,1]$ is a joint cdf with one-dimensional marginal cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, \dots, n$, absolutely continuous with respect to the product of marginal cdf's $\prod_{k=1}^n F_k(x_k)$ if and only if there exist functions $\tilde{g}_{i_1,\dots,i_c} : [0,1]^c \to \mathbf{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \dots, n$, satisfying conditions* A4–A6 *and such that the following representation holds for F:*

$$F(x_1, \dots, x_n) = \int_0^{F_1(x_1)} \cdots \int_0^{F_n(x_n)} (1 + \tilde{U}_n(t_1, \dots, t_n)) \prod_{i=1}^{n} dt_i, \tag{3.2}$$

or, equivalently, if and only if the element of joint frequency dF can be expressed in the form

$$dF(x_1, \dots, x_n) = \prod_{i=1}^{n} dF_i(x_i)(1 + \tilde{U}_n(F_1(x_1), \dots, F_n(x_n))).$$

Remark 3.1. The functions g and $\tilde{g}$ in Theorems 2.1-3.3 are related in the following way: $g_{i_1,\ldots,i_c}(x_{i_1},\ldots,x_{i_c}) = \tilde{g}_{i_1,\ldots,i_c}(F_{i_1}(x_{i_1}),\ldots,F_{i_c}(x_{i_c}))$.

Theorems 2.1–3.3 provide a general device for constructing multivariate copulas and distributions. E.g., taking in (3.1) and (3.2) $n = 2$, $\tilde{g}_{1,2}(t_1,t_2) = \alpha(1-2t_1)(1-2t_2)$, $\alpha \in [-1,1]$, we get the family of bivariate Eyraud–Farlie–Gumbel–Morgenstern copulas $C_\alpha(u_1,u_2) = u_1u_2(1+\alpha(1-u_1)(1-u_2))$ and corresponding distributions $F_\alpha(x_1,x_2) = F_1(x_1)F_2(x_2)(1+\alpha(1-F_1(x_1))(1-F_2(x_2))$. More generally, taking $\tilde{g}_{i_1,\ldots,i_c}(t_{i_1},\ldots,t_{i_c}) = 0$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2,\ldots,n-1$, $\tilde{g}_{1,2,\ldots,n}(t_1,t_2,\ldots,t_n) = \alpha(1-2t_1)(1-2t_2)\cdots(1-2t_n)$, we obtain the multivariate Eyraud–Farlie–Gumbel–Morgenstern copulas $C_\alpha(u_1,u_2,\ldots,u_n) = \prod_{i=1}^n u_i(1+\alpha\prod_{i=1}^n(1-u_i))$ and corresponding multivariate cdf's $F_\alpha(x_1,x_2,\ldots,x_n) = \prod_{i=1}^n F_i(x_i)(1+\alpha\prod_{i=1}^n(1-F_i(x_i)))$.

Let $\alpha_{i_1,\ldots,i_c} \in \mathbf{R}$ be constants such that $\sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} \alpha_{i_1,\ldots,i_c} \times \delta_{i_1}\cdots\delta_{i_c} \ge -1$ for all $\delta_i \in \{0,1\}$, $i = 1,\ldots,n$. The choice $\tilde{g}_{i_1,\ldots,i_c}(t_{i_1},\ldots,t_{i_c}) = \alpha_{i_1,\ldots,i_c}(1-2t_{i_1})(1-2t_{i_2})\cdots(1-2t_{i_c})$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2,\ldots,n$, gives the following generalized multivariate Eyraud–Farlie–Gumbel–Morgenstern copulas (see Johnson and Kotz [45] and Cambanis [10]):

$$C(u_1,\ldots,u_n) = \prod_{k=1}^n u_k(1+\sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} \alpha_{i_1,\ldots,i_c}(1-u_{i_k})) \tag{3.3}$$

and the corresponding cdf's

$$F(x_1,\ldots,x_n) = \prod_{i=1}^n F_i(x_i)(1+\sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} \alpha_{i_1,\ldots,i_c}(1-F_{i_k}(x_{i_k}))).$$

The importance of the generalized Eyraud–Farlie–Gumbel–Morgenstern copulas and cdf's stems, in particular, from the fact that, as shown in Sharakhmetov and Ibragimov [70], they completely characterize joint distributions of two-valued r.v.'s.

Taking $n = 2$, $\tilde{g}_{1,2}(t_1,t_2) = \theta c(t_1,t_2)$, where c is a continuous function on the unit square $[0,1]^2$ satisfying the properties $\int_0^1 c(t_1,t_2)dt_1 = \int_0^1 c(t_1,t_2)dt_2 = 0$, $1+\theta c(t_1,t_2) \ge 0$ for all $0 \le t_1,t_2 \le 1$, one obtains the class of bivariate densities studied by Rüschendorf [67] and Long and Krzysztofowicz [53] (see also [56, pp. 73–78]) $f(x_1,x_2) = f_1(x_1)f_2(x_2)(1+\theta c(F_1(x_1),F_2(x_2)))$ with the covariance characteristic c and the covariance scalar θ. Furthermore, from Theorems 2.1–3.3 it follows that this representation in fact holds for an arbitrary density function and the function $\theta c(t_1,t_2)$ is unique.

4. From dependence to independence through U-statistics

Denote by $\mathcal{G}_n$ the class of sums of U-statistics of the form

$$U_n(\xi_1,\ldots,\xi_n) = \sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} g_{i_1,\ldots,i_c}(\xi_{i_1},\ldots,\xi_{i_c}), \tag{4.1}$$

where the functions $g_{i_1,\ldots,i_c}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2,\ldots,n$, satisfy conditions A1–A3, and, as before, $\xi_1,\ldots,\xi_n$ are independent r.v.'s with cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1,\ldots,n$.

The following theorem puts into correspondence to any set of arbitrarily dependent r.v.'s a sum of U-statistics in independent r.v.'s with canonical kernels. This

allows one to reduce problems for dependent r.v.'s to well-studied objects and to transfer results known for independent r.v.'s and U-statistics to the case of arbitrary dependence. In what follows, the joint distributions considered are assumed to be absolutely continuous with respect to the product of the marginal distributions $\prod_{k=1}^{n} F_k(x_k)$.

Theorem 4.1. *The r.v.'s $X_1, \dots, X_n$ have one-dimensional cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, \dots, n$, if and only if there exists $U_n \in \mathcal{G}_n$ such that for any Borel measurable function $f : \mathbf{R}^n \to \mathbf{R}$ for which the expectations exist*

$$Ef(X_1, \dots, X_n) = Ef(\xi_1, \dots, \xi_n)(1 + U_n(\xi_1, \dots, \xi_n)). \tag{4.2}$$

Note that the above Theorem 4.1 holds for complex-valued functions f as well as for real-valued ones. That is, letting $f(x_1, \dots, x_n) = \exp(i \sum_{k=1}^{n} t_k x_k)$, $t_k \in \mathbf{R}$, $k = 1, \dots, n$, one gets the following representation for the joint characteristic function of the r.v.'s $X_1, \dots, X_n$:

$$E \exp\left(i \sum_{k=1}^{n} t_k X_k\right) = E \exp\left(i \sum_{k=1}^{n} t_k \xi_k\right) + E \exp\left(i \sum_{k=1}^{n} t_k \xi_k\right) U_n(\xi_1, \dots, \xi_n).$$

5. Characterizations of classes of dependent random variables

The following Theorems 5.1–5.8 give characterizations of different classes of dependent r.v.'s in terms of functions g that appear in the representations for joint distributions obtained in Section 2. Completely similar results hold for the functions $\tilde{g}$ that enter corresponding representations for copulas in Section 3.

Theorem 5.1. *The r.v.'s $X_1, \dots, X_n$ with one-dimensional cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, \dots, n$, are independent if and only if the functions $g_{i_1,\dots,i_c}$ in representations* (2.1) *and* (2.2) *satisfy the conditions*

$$g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0 \text{ (a.s.)}, \ 1 \le i_1 < \dots < i_c \le n, c = 2, \dots, n.$$

Theorem 5.2. *A sequence of r.v.'s $\{X_n\}$ is strictly stationary if and only if the functions $g_{i_1,\dots,i_c}$ in representations* (2.1) *and* (2.2) *for any finite-dimensional distribution (see Remark* 2.1*) satisfy the conditions*

$$g_{i_1+h,\dots,i_c+h}(\xi_{i_1}, \dots, \xi_{i_c}) = g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) \text{ (a.s.) } 1 \le i_1 < \dots < i_c \le n,$$

$$c = 2, 3, \dots, h = 0, 1, \dots$$

Theorem 5.3. *A sequence of r.v.'s $\{X_n\}$ with $EX_k = 0$, $EX_k^2 < \infty$, $k = 1, 2, \dots$, is weakly stationary if and only if the functions g in representations* (2.1) *and* (2.2) *for any finite-dimensional distribution have the property that the function $h(s,t) = E\xi_s \xi_t g_{st}(\xi_s, \xi_t)$, depends only on $|t - s|$, $t, s = 1, 2, \dots$*

Definition 5.1. The r.v.'s $X_1, \dots, X_n$ with $EX_i = 0$, $i = 1, \dots, n$, are called *orthogonal* if $EX_i X_j = 0$ for all $1 \le i < j \le n$.

Theorem 5.4. *The r.v.'s $X_1, \dots, X_n$ with $EX_k = 0$, $k = 1, \dots, n$, are orthogonal if and only if the functions g in representations* (2.1) *and* (2.2) *satisfy the conditions $E\xi_i \xi_j g_{ij}(\xi_i, \xi_j) = 0$, $1 \le i < j \le n$.*

Definition 5.2. The r.v.'s $X_1, \dots, X_n$ are called exchangeable if all $n!$ permutations $(X_{\pi(1)}, \dots, X_{\pi(n)})$ of the r.v.'s have the same joint distributions.

Theorem 5.5. *The identically distributed r.v.'s $X_1, \dots, X_n$ are exchangeable if and only if the functions $g_{i_1,\dots,i_c}$ in representations* (2.1) *and* (2.2) *satisfy the conditions $g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = g_{i_{\pi(1)},\dots,i_{\pi(c)}}(\xi_{i_{\pi(1)}}, \dots, \xi_{i_{\pi(i_c)}})$ (a.s.) for all $1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, n$, and all permutations π of the set $\{1, \dots, n\}$.*

Definition 5.3. The r.v.'s $X_1, \dots, X_n$ are called *m-dependent* ($1 \le m \le n$) if any two vectors $(X_{j_1}, X_{j_2}, \dots, X_{j_{a-1}}, X_{j_a})$ and $(X_{j_{a+1}}, X_{j_{a+2}}, \dots, X_{j_{l-1}}, X_{j_l})$, where $1 \le j_1 < \dots < j_a < \dots < j_l \le n$, $a = 1, 2, \dots, l-1$, $l = 2, \dots, n$, $j_{a+1} - j_a \ge m$, are independent.

Theorem 5.6. *The r.v.'s $X_1, \dots, X_n$ are m-dependent if and only if the functions g in representations* (2.1) *and* (2.2) *satisfy the conditions $g_{i_1,\dots,i_k,i_{k+1},\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_k}, \xi_{i_{k+1}}, \dots, \xi_{i_c}) = g_{i_1,\dots,i_k}(\xi_{i_1}, \dots, \xi_{i_k}) g_{i_{k+1},\dots,i_c}(\xi_{i_{k+1}}, \dots, \xi_{i_c})$ for all $1 \le i_1 < \dots < i_k < i_{k+1} < \dots < i_c \le n$, $i_{k+1} - i_k \ge m$, $k = 1, \dots, c-1$, $c = 2, \dots, n$.*

Definition 5.4. The r.v.'s $X_1, \dots, X_n$ form a *multiplicative system* of order $\alpha \in \mathbf{N}$ (shortly, $MS(\alpha)$) if $E|X_j|^\alpha < \infty$, $j = 1, \dots, n$, and $E\prod_{j=1}^n X_j^{\alpha_j} = \prod_{j=1}^n EX_j^{\alpha_j}$ for any $\alpha_j \in \{0, 1, \dots, \alpha\}$, $j = 1, \dots, n$.

The systems $MS(1)$ and $MS(2)$ under the names multiplicative and strongly multiplicative systems, respectively, were introduced by Alexits [2]. Multiplicative systems of an arbitrary order were considered, e.g., by Kwapień [48] and Sharakhmetov [68]. Examples of the multiplicative systems $MS(1)$ are given, besides independent r.v.'s, by the lacunary trigonometric systems $\{\cos 2\pi n_k x, \sin 2\pi n_k x, k = 1, 2, \dots\}$ on the interval $[0, 1]$ with the Lebesgue measure for $n_{k+1}/n_k \ge 2$ important in Fourier analysis of time series and also by such important classes of dependent r.v.'s as martingale-difference sequences. Examples of strongly multiplicative systems (that is, the systems $MS(2)$) are given by the lacunary trigonometric systems for $n_{k+1}/n_k \ge 3$ and martingale-difference sequences $X_1, \dots, X_n$ satisfying the conditions $E(X_n^2 | X_1, \dots, X_{n-1}) = b_n^2 \in \mathbf{R}$, $n = 1, 2, \dots$. Examples of the systems $MS(\alpha)$ include, for instance, the lacunary trigonometric systems with large lacunas, that is, with $n_{k+1}/n_k \ge \alpha + 1$ and also ϵ-independent and asymptotically independent r.v.'s introduced by Zolotarev [78].

Theorem 5.7. *The r.v.'s $X_1, \dots, X_n$ form a multiplicative system of order α if and only if the functions $g_{i_1,\dots,i_c}$ in representations* (2.1) *and* (2.2) *satisfy the conditions $E\xi_{i_1}^{\alpha_{i_1}} \cdots \xi_{i_c}^{\alpha_{i_c}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0$, $1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, n$, $\alpha_j \in \{0, 1, \dots, \alpha\}$, $j = 1, \dots, n$.*

Definition 5.5. The r.v.'s $X_1, \dots, X_n$ are called *r-independent* ($2 \le r < n$) if any r of them of are jointly independent.

Theorem 5.8. *The r.v.'s $X_1, \dots, X_n$ are r-independent if and only if the functions $g_{i_1,\dots,i_c}$ in representations* (2.1) *and* (2.2) *satisfy the conditions $g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0$ (a.s.), $1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, r$.*

Remark 5.1. Let $F_1(x), \dots, F_n(x)$ be arbitrary one-dimensional distribution functions, $\alpha_1, \dots, \alpha_n \in (-1, 1) \setminus \{0\}$, $\sum_{i=1}^n |\alpha_i| \le 1$. Taking $g_{i_1,\dots,i_c}(t_{i_1}, \dots, t_{i_c}) = 0$, $1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, n$, $c \ne r+1$, $g_{i_1,\dots,i_{r+1}}(t_{i_1}, \dots, t_{i_{r+1}}) = \frac{\alpha_1 \cdots \alpha_n}{\alpha_{i_1} \cdots \alpha_{i_{r+1}}}((k+1)t_{i_1}^k - (k+2)t_{i_1}^{k+1}) \cdots ((k+1)t_{i_c}^k - (k+2)t_{i_c}^{k+1})$, $k = 0, 1, 2, \dots$, in Theorem 3.3, we obtain the following extensions of the examples of r-independent

r.v.'s obtained by Wang [77]: For $k = 0, 1, 2, \ldots$,

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i) \Bigg(1 + \sum_{1 \le i_1 < \ldots < i_{r+1} \le n} \frac{\alpha_1 \ldots \alpha_n}{\alpha_{i_1} \ldots \alpha_{i_r+1}} (F_{i_1}^k(x_{i_1}) - F_{i_1}^{k+1}(x_{i_1})) \cdots \times (F_{i_{r+1}}^k(x_{i_{r+1}}) - F_{i_{r+1}}^{k+1}(x_{i_{r+1}})) \Bigg),$$

(Wang's examples are with $k = 0$).

6. Further applications: a structural property of multiplicative systems

The following theorem shows that r.v.'s forming a multiplicative system of order α and taking not more than $\alpha+1$ values are jointly independent. Let $card(A_i)$ denote the number of elements in (finite) sets A_i.

Theorem 6.1. *Let $\alpha \in \mathbf{N}$, and let A_i, $i = 1, \ldots, n$, be sets of real numbers such that $card(A_i) \le \alpha + 1$, $i = 1, \ldots, n$. The r.v.'s $X_1, \ldots, X_n$ taking values in $A_1, \ldots, A_n$, respectively, form a multiplicative system of order α if and only if they are jointly independent.*

Remark 6.1. From Theorem 6.1 with $\alpha = 1$ the following result obtained in [70] follows: A sequence of r.v.'s $\{X_n\}$ on a probability space $(\Omega, \Im, P)$ assuming two values is a martingale-difference with respect to an increasing sequence of σ-algebras $\Im_0 = (\Omega, \emptyset) \subseteq \Im_1 \subseteq \cdots \subseteq \Im$ if and only if the r.v.'s $\{X_n\}$ are jointly independent. In addition, we obtain that if a sequence of r.v.'s $\{X_n\}$ assuming three values is a martingale-difference with respect to $(\Im_n)$ such that $E(X_n^2 | \Im_{n-1}) = b_n^2 \in \mathbf{R}$, then the r.v.'s are jointly independent.

7. Measures of dependence and sharp moment and probability inequalities for dependent random variables

In this section, we apply the results from Section 2 to study properties of different measures of dependence and convergence of multidimensional statistics of time series. We obtain results that allow one to reduce the analysis of convergence of statistics of time series to the study of convergence of the measures of intertemporal dependence in the time series and limit behavior of the statistics in the case of independence. We also prove new sharp complete decoupling inequalities for dependent r.v.'s in terms of their dependence characteristics. The theory of complete decoupling inequalities has experienced an impetus in recent years. The interested reader should consult de la Peña [14], de la Peña and Giné [15] and de la Peña and Lai [18] (a survey) for more on the subject. Let $X_1, \ldots, X_n$ be r.v.'s with one-dimensional cdf's $F_k(x_k)$, $k = 1, \ldots, n$, and joint cdf $F(x_1, \ldots, x_n)$. Recall that $G(x_1, \ldots, x_n) = dF(x_1, \ldots, x_n) / \prod_{i=1}^n dF_i$ and consider the following measures of dependence for the r.v.'s $X_1, \ldots, X_n$:

$$\begin{aligned} \phi^2_{X_1,\ldots,X_n} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G(x_1, \ldots, x_n) dF(x_1, \ldots, x_n) - 1 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G^2(x_1, \ldots, x_n) \prod_{i=1}^{n} dF_i(x_i) - 1 \end{aligned}$$

(multivariate analog of Pearson's ϕ^2 coefficient), and

$$\delta_{X_1,\dots,X_n} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log(G(x_1,\dots,x_n))dF(x_1,...,x_n)$$

(relative entropy), where the integral signs are in the sense of Lebesgue–Stieltjes and $G(x_1,\dots,x_n)$ is taken to be 1 if $(x_1,\dots,x_n)$ is not in the support of $dF_1\cdots dF_n$. In the case of absolutely continuous r.v.'s $X_1,\dots,X_n$ the measures $\delta_{X_1,\dots,X_n}$ and $\phi^2_{X_1,\dots,X_n}$ were introduced by Joe [42, 43]. In the case of two r.v.'s X_1 and X_2 the measure $\phi^2_{X_1,X_2}$ was introduced by Pearson [63] and was studied, among others, by Lancaster [49–51]. In the bivariate case, the measure δ_{X_1,X_2} is commonly known as Shannon or Kullback–Leibler mutual information between X_1 and X_2. It should be noted (see [43]) that if $(X_1,\dots,X_n)' \sim N(\mu,\Sigma)$, then $\phi^2_{X_1,\dots,X_n} = |R(2I_n - R)|^{-1/2} - 1$, where I_n is the $n\times n$ identity matrix, provided that the correlation matrix R corresponding to Σ has the maximum eigenvalue of less than 2 and is infinite otherwise ($|A|$ denotes the determinant of a matrix A). In addition to that, if in the above case $\operatorname{diag}(\Sigma) = (\sigma_1^2,\dots,\sigma_n^2)$, then $\delta_{X_1,\dots,X_n} = -.5\log(|\Sigma|/\prod_{i=1}^n \sigma_i^2)$. In the case of two normal r.v.'s X_1 and X_2 with the correlation coefficient ρ, $(\phi^2_{X_1,X_2}/(1+\phi^2_{X_1,X_2}))^{1/2} = (1-\exp(-2\delta_{X_1,X_2}))^{1/2} = |\rho|$.

The multivariate Pearson's ϕ^2 coefficient and the relative entropy are particular cases of multivariate divergence measures $D^{\psi}_{X_1,\dots,X_n} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \psi(G(x_1,\dots,x_n))\prod_{i=1}^n dF_i(x_i)$, where ψ is a strictly convex function on $\mathbf{R}$ satisfying $\psi(1)=0$ and $G(x_1,\dots,x_n)$ is to be taken to be 1 if at least one $x_1,\dots,x_n$ is not a point of increase of the corresponding $F_1,\dots,F_n$. Bivariate divergence measures were considered, e.g., by Ali and Silvey [3] and Joe [43]. The multivariate Pearson's ϕ^2 corresponds to $\psi(x) = x^2 - 1$ and the relative entropy is obtained with $\psi(x) = x\log x$.

A class of measures of dependence closely related to the multivariate divergence measures is the class of generalized entropies introduced by Tsallis [75] in the study of multifractals and generalizations of Boltzmann–Gibbs statistics (see also [24, 26, 27])

$$\rho^{(q)}_{X_1,\dots,X_n} = \frac{1}{1-q}\Bigl(1 - \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} G^{1-q}(x_1,\dots,x_n)\Bigr)\prod_{i=1}^n dF_i(x_i),$$

where q is the entropic index. In the limiting case $q\to 1$, the discrepancy measure $\rho^{(q)}$ becomes the relative entropy $\delta_{X_1,\dots,X_n}$ and in the case $q\to 1/2$ it becomes the scaled squared Hellinger distance between dF and $dF_1\cdots dF_n$

$$\rho^{(1/2)}_{X_1,\dots,X_n} = \frac{1}{2}\Bigl(1 - \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} G^{1/2}(x_1,\dots,x_n)\Bigr)\prod_{i=1}^n dF_i(x_i)\Bigr) = 2\mathcal{H}^2_{X_1,\dots,X_n}$$

($\mathcal{H}_{X_1,\dots,X_n}$ stands for the Hellinger distance). The generalized entropy has the form of the multivariate divergence measures $D^{\psi}_{X_1,\dots,X_n}$ with $\psi(x) = (1/(1-q))(1-x^{1-q})$.

In the terminology of information theory (see, e.g., Akaike [1]) the multivariate analog of Pearson coefficient, the relative entropy and, more generally, the multivariate divergence measures represent the mean amount of information for discrimination between the density f of dependent sample and the density of the sample of independent r.v.'s with the same marginals $f_0 = \prod_{k=1}^n f_k(x_k)$ when the actual distribution is dependent $I(f_0, f; \Phi) = \int \Phi(f(x)/f_0(x))f(x)dx$, where Φ is a properly chosen function. The multivariate analog of Pearson coefficient is characterized by the relation (below, f_0 denotes the density of independent sample and f denotes

the density of a dependent sample) $\phi^2 = I(f_0, f; \Phi_1)$, where $\Phi_1(x) = x$; the relative entropy satisfies $\delta = I(f_0, f; \Phi_2)$, where $\Phi_2(x) = \log(x)$; and the multivariate divergence measures satisfy $D^{\psi}_{X_1,\dots,X_n} = I(f_0, f, \Phi_3)$, where $\Phi_3(x) = \psi(x)/x$.

If $g_{i_1,\dots,i_c}(x_{i_1}, \dots, x_{i_c})$ are functions corresponding to Theorem 2.1 and Remark 2.2, then from Theorem 4.1 it follows that the measures $\delta_{X_1,\dots,X_n}$, $\phi^2_{X_1,\dots,X_n}$, $D^{\psi}_{X_1,\dots,X_n}$, $\rho^{(q)}_{X_1,\dots,X_n}$ (in particular, $2\mathcal{H}^2_{X_1,\dots,X_n}$ for $q = 1/2$) and $I(f_0, f; \Phi)$ can be written as

$$\begin{aligned} \delta_{X_1,\dots,X_n} &= E \log\left(1 + U_n(X_1, \dots, X_n)\right) \\ &= E\left(1 + U_n(\xi_1, \dots, \xi_n)\right) \log(1 + U_n(\xi_1, \dots, \xi_n)), \end{aligned} \tag{7.1}$$

$$\begin{aligned} \phi^2_{X_1,\dots,X_n} &= E\left(1+U_n(\xi_1, \dots, \xi_n)\right)^2 - 1 \\ &= EU_n^2(\xi_1, \dots, \xi_n) = EU_n(X_1, \dots, X_n), \end{aligned} \tag{7.2}$$

$$D^{\psi}_{X_1,\dots,X_n} = E\psi\left(1 + U_n(\xi_1, \dots, \xi_n)\right), \tag{7.3}$$

$$\rho^{(q)}_{X_1,\dots,X_n} = (1/(1-q))(1 - E(1 + U_n(\xi_1, \dots, \xi_n))^q), \tag{7.4}$$

$$2\mathcal{H}^2_{X_1,\dots,X_n} = 1/2(1 - E(1 + U_n(\xi_1, \dots, \xi_n))^{1/2}), \tag{7.5}$$

$$I(f_0, f; \Phi) = E\Phi\left(1 + U_n(\xi_1, \dots, \xi_n)\right)\left(1 + U_n(\xi_1, \dots, \xi_n)\right), \tag{7.6}$$

where $U_n(x_1, \dots, x_n)$ is as defined by (4.1).

From (7.2) it follows that the following formula that gives an expansion for $\phi^2_{X_1,\dots,X_n}$ in terms of the "canonical" functions g holds: $\phi^2_{X_1,\dots,X_n} = \sum_{c=2}^{n} \sum_{1 \le i_1 < \dots < i_c \le n} E g^2_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c})$. In particular, in the case of the r.v.'s $X_1, \dots, X_n$ with the generalized multivariate Eyraud–Farlie–Gumbel–Morgenstern copulas (3.3) the measure of dependence $\phi^2_{X_1,\dots,X_n}$ is given by $\phi^2_{X_1,\dots,X_n} = \sum_{c=2}^{n} \sum_{1 \le i_1 < \dots < i_c \le n} \alpha^2_{i_1,\dots,i_c}$.

It is well known that the mutual information between two r.v.'s X_1 and X_2 is nonnegative (see [12, p. 27]). The multivariate analog of this property for $\delta_{X_1,\dots,X_n}$ follows from the results obtained by Joe [43]. It is interesting to note that nonnegativity of $\delta_{X_1,\dots,X_n}$ can be easily obtained from (7.1): since the function $(1 + x)\ln(1 + x)$ is convex in $x \ge 0$, by Jensen inequality we get $\delta_{X_1,\dots,X_n} \ge (1 + EU_n(\xi_1, \dots, \xi_n)) \log(1 + EU_n(\xi_1, \dots, \xi_n)) = 0$. The following theorem gives an inequality between the measures $\delta_{X_1,\dots,X_n}$ and $\phi^2_{X_1,\dots,X_n}$ that generalizes and improves the results obtained, in the bivariate case, by Dragomir [20] (see also Mond and Pečarič [59]).

Theorem 7.1. *The following inequalities hold*

$$\delta_{X_1,\dots,X_n} \le \log(1 + \phi^2_{X_1,\dots,X_n}) \le \phi^2_{X_1,\dots,X_n}.$$

The results obtained in the previous sections, in particular, Theorem 4.1, provide a device for reducing the analysis of convergence in distribution of double arrays of dependent variables to the study of the convergence in distribution of a decoupled counterpart plus a measure of dependence in the time series. We apply this idea in reduction of the analysis of convergence of multidimensional statistics of time series

to the study of convergence of the measures of intertemporal dependence of the time series, including the above multivariate Pearson coefficient ϕ, the relative entropy δ, the divergence measures D^ψ and the mean information for discrimination between the dependence and independence $I(f_0, f; \Phi)$. We obtain the following Theorem 7.2 which deals with the convergence in distribution of m-dimensional statistics of time series.

Let $h : \mathbf{R}^m \to \mathbf{R}$ be an arbitrary function of m arguments, Y be some r.v. and let ψ be a convex function increasing on $[1, \infty)$ and decreasing on $(-\infty, 1)$ with $\psi(1) = 0$. In what follows, $\xrightarrow{\mathcal{D}}$ represents convergence in distribution. In addition, $\{\xi_i^n\}$ and $\{\xi_t\}$ stand for dependent copies of $\{X_i^n\}$ and $\{X_t\}$.

Theorem 7.2. *For the double array $\{X_i^n\}, i = 1, \dots, n, n = 0, 1, \dots$ let functionals $\phi^2_{n,n} = \phi^2_{X_1^n, X_2^n, \dots, X_n^n}$, $\delta_{n,n} = \delta_{X_1^n, X_2^n, \dots, X_n^n}$, $D^\psi_{n,n} = D^\psi_{X_1^n, X_2^n, \dots, X_n^n}$, $\rho^{(q)}_{n,n} = \rho^{(q)}_{X_1^n, X_2^n, \dots, X_n^n}$, $q \in (0,1)$, $\mathcal{H}_{n,n} = (1/2\rho^{(q)}_{n,n})^{1/2}$, $n = 0, 1, 2, \dots$ denote the corresponding distances. Then, as $n \to \infty$, if*

$$\sum_{i=1}^n \xi_i^n \xrightarrow{\mathcal{D}} Y$$

and either $\phi^2_{n,n} \to 0$, $\delta_{n,n} \to 0$, $D^\psi_{n,n} \to 0$, $\rho^{(q)}_{n,n} \to 0$ or $\mathcal{H}_{n,n} \to 0$ as $n \to \infty$, then as $n \to \infty$,

$$\sum_{i=1}^n X_i^n \xrightarrow{\mathcal{D}} Y.$$

For a time series $\{X_t\}_{t=0}^\infty$ let the functionals $\phi_t^2 = \phi^2_{X_t, X_{t+1}, \dots, X_{t+m-1}}$, $\delta_t = \delta_{X_t, X_{t+1}, \dots, X_{t+m-1}}$, $D_t^\psi = D^\psi_{X_t, X_{t+1}, \dots, X_{t+m-1}}$, $\rho_t^{(q)} = \rho^{(q)}_{X_t, X_{t+1}, \dots, X_{t+m-1}}$, $q \in (0,1)$, $\mathcal{H}_t = (1/2\rho_t^{(q)})^{1/2}$, $t = 0, 1, 2, \dots$ denote the m-variate Pearson coefficient, the relative entropy, the multivariate divergence measure associated with the function ψ, the generalized Tsallis entropy and the Hellinger distance for the time series, respectively.

Then, if, as $t \to \infty$,

$$h(\xi_t, \xi_{t+1}, \dots, \xi_{t+m-1}) \xrightarrow{\mathcal{D}} Y$$

and either $\phi_t^2 \to 0$, $\delta_t \to 0$, $D_t^\psi \to 0$, $\rho_t^{(q)} \to 0$ or $\mathcal{H}_t \to 0$ as $t \to \infty$, then, as $t \to \infty$,

$$h(X_t, X_{t+1}, \dots, X_{t+m-1}) \xrightarrow{\mathcal{D}} Y.$$

From the discussion in the beginning of the present section it follows that in the case of Gaussian processes $\{X_t\}_{t=0}^\infty$ with $(X_t, X_{t+1}, \dots, X_{t+m-1}) \sim N(\mu_{t,m}, \Sigma_{t,m})$, the conditions of Theorem 7.2 are satisfied if, for example, $|R_{t,m}(2I_m - R_{t,m})| \to 1$ or $|\Sigma_{t,m}| / \sum_{i=0}^{m-1} \sigma_{t+i}^2 \to 1$, as $t \to \infty$, where $R_{t,m}$ denote correlation matrices corresponding to $\Sigma_{t,m}$ and $(\sigma_t^2, \dots, \sigma_{t+m-1}^2) = \mathrm{diag}(\Sigma_{t,m})$. In the case of processes $\{X_t\}_{t=1}^\infty$ with distributions of r.v.'s $X_1, \dots, X_n$, $n \geq 1$, having generalized Eyraud–Farlie–Gumbel–Morgenstern copulas (3.3) (according to [70], this is the case for any time series of r.v.'s assuming two values), the conditions of the theorem are satisfied if, for example, $\phi_t^2 = \sum_{c=2}^m \sum_{i_1 < \dots < i_c \in \{t, t+1, \dots, t+m-1\}} \alpha^2_{i_1, \dots, i_c} \to 0$ as $t \to \infty$.

It is important to emphasize here that since Theorem 7.2 holds for Tsallis entropy and multivariate divergence measures, it allows one to study convergence of statistics of time series in the case when only lower moments of the above-mentioned U-statistics underlying the dependence generating process for the time series exist.

Therefore, they provide a unifying approach to studying convergence in "heavy-tailed" situations and "standard" cases connected with the convergence of Pearson coefficient and the mutual information and entropy (corresponding, respectively, to the cases of second moments of the U-statistics and the first moments multiplied by logarithm).

The following theorem provides an estimate for the distance between the distribution function of an arbitrary statistic in dependent r.v.'s and the distribution function of the statistic in independent copies of the r.v.'s. The inequality complements (and can be better than) the well-known Pinsker's inequality for total variation between the densities of dependent and independent r.v.'s in terms of the relative entropy (see, e.g., [58]).

Theorem 7.3. *The following inequality holds for an arbitrary statistic* $h(X_1, \dots, X_n)$*:*

$$|P(h(X_1,\dots,X_n) \le x) - P(h(\xi_1,\dots,\xi_n) \le x)| \\ \le \phi_{X_1,\dots,X_n} \max\left[\left(P\left(h(\xi_1,\dots,\xi_n) \le x\right)\right)^{1/2}, \left(P\left(h(\xi_1,\dots,\xi_n) > x\right)\right)^{1/2}\right],$$

$x \in \mathbf{R}$.

The following theorems allow one to reduce the problems of evaluating expectations of general statistics in dependent r.v.'s $X_1, \dots, X_n$ to the case of independence. The theorems contain complete decoupling results for statistics in dependent r.v.'s using the relative entropy and the multivariate Pearson's ϕ^2 coefficient. The results provide generalizations of earlier known results on complete decoupling of r.v.'s from particular dependence classes, such as martingales and adapted sequences of r.v.'s to the case of arbitrary dependence.

Theorem 7.4. *If* $f : \mathbf{R}^n \to \mathbf{R}$ *is a nonnegative function, then the following sharp inequalities hold:*

$$Ef(X_1,\dots,X_n) \le Ef(\xi_1,\dots,\xi_n) + \phi_{X_1,\dots,X_n}(Ef^2(\xi_1,\dots,\xi_n))^{1/2}, \tag{7.7}$$

$$Ef(X_1,\dots,X_n) \le (1+\phi^2_{X_1,\dots,X_n})^{1/q}(Ef^q(\xi_1,\dots,\xi_n))^{1/q},\ q \ge 2, \tag{7.8}$$

$$Ef(X_1,\dots,X_n) \le E\exp(f(\xi_1,\dots,\xi_n)) - 1 + \delta_{X_1,\dots,X_n}, \tag{7.9}$$

$$Ef(X_1,\dots,X_n) \le (1+D^{\psi}_{X_1,\dots,X_n})^{(1-\frac{1}{q})}(Ef^q(\xi_1,\dots,\xi_n))^{1/q},\ q > 1, \tag{7.10}$$

where $\psi(x) = |x|^{q/(q-1)} - 1$.

Remark 7.1. It is interesting to note that from relation (7.2) and inequality (7.7) it follows that the following representation holds for the multivariate Pearson coefficient $\phi_{X_1,\dots,X_n}$:

$$\phi_{X_1,\dots,X_n} = \max_{\substack{f: Ef(\xi_1,\dots,\xi_n)=0,\\ Ef^2(\xi_1,\dots,\xi_n)<\infty}} \frac{(Ef(X_1,\dots,X_n) - Ef(\xi_1,\dots,\xi_n))}{(Ef^2(\xi_1,\dots,\xi_n))^{1/2}}. \tag{7.11}$$

The following result gives complete decoupling inequalities for the tail probabilities of arbitrary statistics in dependent r.v.'s.

Theorem 7.5. *The following inequalities hold:*

$$P\left(h(X_1,\dots,X_n)>x\right)\le P(h(\xi_1,\dots,\xi_n)>x)+\phi_{X_1,\dots,X_n}\left(P(h(\xi_1,\dots,\xi_n)>x)\right)^{\frac{1}{2}},$$

$$P\left(h(X_1,\dots,X_n)>x\right)\le\left(1+\phi^2_{X_1,\dots,X_n}\right)^{1/2}\left(P(h(\xi_1,\dots,\xi_n)>x)\right)^{1/2},$$

$$P(h(X_1,\dots,X_n)>x)\le(e-1)P(h(\xi_1,\dots,\xi_n)>x)+\delta_{X_1,\dots,X_n},$$

$$P(h(X_1,\dots,X_n)>x)\le\left(1+D^{\psi}_{X_1,\dots,X_n}\right)^{(1-\frac{1}{q})}\left(P(h(\xi_1,\dots,\xi_n)>x)\right)^{\frac{1}{q}},\ q>1,$$

$x\in\mathbf{R}$, *where* $\psi(x)=|x|^{q/(q-1)}-1$.

8. Appendix: Proofs

Proof of Theorem 2.1. Let us first prove the necessity part of the theorem. Denote

$$T(x_1,\dots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}(1+U_n(t_1,\dots,t_n))\prod_{i=1}^{n}dF_i(t_i).$$

Let $k\in\{1,\dots,n\}$, $x_k\in\mathbf{R}$. Let us show that

$$T(\infty,\dots,\infty,x_k,\infty,\dots,\infty)=F_k(x_k), \tag{8.1}$$

$x_k\in\mathbf{R}$, $k=1,\dots,n$. It suffices to consider the case $k=1$. We have

$$\begin{aligned}
&T(x_1,\infty,\dots,\infty)\\
&\quad=\underbrace{\int_{-\infty}^{x_1}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}}_{n}(1+U_n(t_1,\dots,t_n))\prod_{i=1}^{n}dF_i(t_i)\\
&\quad=F_1(x_1)+\sum_{c=2}^{n}\sum_{1\le i_1<\dots<i_c\le n}\underbrace{\int_{-\infty}^{x_1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n}g_{i_1,\dots,i_c}(t_{i_1},\dots,t_{i_c})\prod_{i=1}^{n}dF_i(t_i)\\
&\quad=F_1(x_1)+\Sigma''.
\end{aligned}$$

It is easy to see that there is at least one t_s of $t_2,\dots,t_n$ among the arguments of each of the functions $g_{i_1,\dots,i_c}(t_{i_1},\dots,t_{i_c})$, $1\le i_1<\dots<i_c\le n$, $c=2,\dots,n$, in the latter summand. By A2 we get, therefore, that $\Sigma''=0$. Consequently, $T(x_1,\infty,\dots,\infty)=F_1(x_1)$, $x_1\in\mathbf{R}$, and (8.1) holds. It is evident that

$$\lim_{x_k\to-\infty}T(x_1,\dots,x_k,\dots,x_n)=0 \tag{8.2}$$

for all $x_j\in\mathbf{R}$, $j=1,\dots,n$, $j\ne k$, $k=1,\dots,n$. Since

$$T(x_1,\dots,x_n)=\prod_{i=1}^{n}F_i(x_i)+E\left[U_n(\xi_1,\dots,\xi_n))\prod_{i=1}^{n}I(\xi_i\le x_i)\right],$$

from the monotone convergence theorem we obtain that $T(x_1,\dots,x_n)$ is right-continuous in $(x_1,\dots,x_n)\in\mathbf{R}^n$. Let $\delta^k_{[a,b)}T(x_1,\dots,x_n)=T(x_1,\dots,x_{k-1},b,x_{k+1},$

$\dots, x_n) - T(x_1, \dots, x_{k-1}, a, x_{k+1}, \dots, x_n)$, $a < b$. By integrability of the functions $g_{i_1,\dots,i_c}$ and condition A3 we obtain ($I(\cdot)$ denotes the indicator function)

$$\begin{aligned}(8.3)\qquad &\delta^1_{(a_1,b_1]}\delta^2_{(a_2,b_2]}\cdots\delta^n_{(a_n,b_n]}T(x_1,\dots,x_n)\\ &\quad= \prod_{i=1}^n P(a_i < \xi_i \le b_i) + E\left[U_n(\xi_{i_1},\dots,\xi_{i_n})\prod_{i=1}^n I(a_i < \xi_i \le b_i)\right] \ge 0\end{aligned}$$

for all $a_i < b_i$, $i = 1, \dots, n$.[1] Right-continuity of $T(x_1, \dots, x_n)$ and (8.1)–(8.3) imply that $T(x_1, \dots, x_n)$ is a joint cdf of some r.v.'s $X_1, \dots, X_n$ with one-dimensional cdf's $F_k(x_k)$, and the joint cdf $T(x_1, \dots, x_n)$ satisfies (2.1).

Let us now prove the sufficiency part. Consider the functions

$$f_{i_1,\dots,i_c}(x_{i_1},\dots,x_{i_c}) = \sum_{s=2}^{c}(-1)^{c-s}\sum_{j_1<\dots<j_s\in\{i_1,\dots,i_c\}}\left(\frac{dF(x_{j_1},\dots,x_{j_s})}{dF_{j_1}\cdots dF_{j_s}} - 1\right),$$

$1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, n$. Obviously, the functions $f_{i_1,\dots,i_c}$ satisfy condition A1. Let us show that they satisfy condition A2. It suffices to consider the case $i_1 = 1, i_2 = 2, \dots, i_c = c, k = 1$. We have

$$\begin{aligned}&Eg_{1,2,\dots,c}(\xi_1, x_2, \dots, x_c)\\ &\quad= \int_{-\infty}^{\infty} g_{1,2,\dots,c}(x_1, x_2, \dots, x_c)dF_1(x_1)\\ &\quad= \int_{-\infty}^{\infty}\sum_{s=2}^{c}(-1)^{c-s}\left[\sum_{2\le i_2<\dots<i_s\le c}\left(\frac{dF(x_1,x_{i_2},\dots,x_{i_s})}{dF_1dF_{i_2}\cdots dF_{i_s}} - 1\right)\right.\\ &\qquad\qquad\left.+\sum_{2\le i_1<\dots<i_s\le c}\left(\frac{dF(x_{i_1},x_{i_2},\dots,x_{i_s})}{dF_{i_1}dF_{i_2}\cdots dF_{i_s}} - 1\right)\right]dF_1(x_1)\\ &\quad= \sum_{s=2}^{c}(-1)^{c-s}\left\{\sum_{2\le i_2<\dots<i_s\le c}\left(\frac{dF(x_{i_2},\dots,x_{i_s})}{dF_{i_2}\cdots dF_{i_s}} - 1\right)\right.\\ &\qquad\qquad\left.+\sum_{2\le i_1<\dots<i_s\le c}\left(\frac{dF(x_{i_1},\dots,x_{i_s})}{dF_{i_1}\cdots dF_{i_s}} - 1\right)\right\} = 0.\end{aligned}$$

By the inversion formula (see, e.g., [8, pp. 177–178]) it follows that if $a_{i_1,\dots,i_c}$, $b_{i_1,\dots,i_c}$, $1 \le i_1 < \dots < i_c \le n$, $c = 2, \dots, n$, are arbitrary numbers then the relations

$$b_{i_1,\dots,i_c} = \sum_{c=2}^{n}\sum_{j_1<\dots<j_s\in\{i_1,\dots,i_c\}} a_{j_1,\dots,j_s}, 1 \le i_1 < \dots < i_c \le n, c = 2, \dots, n,$$

and

$$a_{i_1,\dots,i_c} = \sum_{s=2}^{c}(-1)^{c-s}\sum_{j_1<\dots<j_s\in\{i_1,\dots,i_c\}} b_{i_1,\dots,i_c}, 1 \le i_1 < \dots < i_c \le n, c = 2, \dots, n,$$

[1] Note that (8.2) and (8.3) are immediate if the probability space and the random variables are defined in the canonical way with $\Omega = \mathbf{R}^n$ and $X_i(\omega) = \omega_i$ for $\omega = (\omega_1, \dots, \omega_n)$ and $P(A) = \int_A (1 + \sum_{1\le i_1<\dots<i_c\le n} g_{i_1,\dots,i_c}(t_{i_1}, \dots, t_{i_c}))\prod_{i=1}^n dF_i(t_i)$.

are equivalent. Taking $a_{i_1,\dots,i_c} = g_{i_1,\dots,i_c}(x_{i_1},\dots,x_{i_c})$, $b_{i_1,\dots,i_c} = dF(x_{i_1},\dots,x_{i_c})/\prod_{j=1}^{c} dF_{i_j} - 1$, for $1 \le i_1 < \cdots < i_c \le n, c = 2,\dots,n$, we obtain, in particular, that

$$G(x_1,\dots,x_n) = 1 + U_n(x_1,\dots,x_n). \tag{8.4}$$

Therefore, representation (2.1) holds and the functions $f_{i_1,\dots,i_c}$ satisfy condition A3. Suppose now that there exists another set of functions $g_{i_1,\dots,i_c}$ satisfying conditions A1–A3 and such that (2.1) holds and, equivalently, (2.2) holds. Let B_s be a set of s integers $j_1,\dots,j_s$ with $1 \le j_1 < \cdots < j_s \le n$, for $s = 2,\dots,n$. By Remark 1 we have

$$\sum_{c=2}^{s} \sum_{\{i_1<\cdots<i_c\}\in B_s} f_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}) = \sum_{c=2}^{s} \sum_{\{i_1<\cdots<i_c\}\in B_s} g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}) \tag{8.5}$$

(a.s.). From (8.5) we subsequently obtain that $f_{i_1,i_2}(\xi_{i_1},\xi_{i_2}) = g_{i_1,i_2}(\xi_{i_1},\xi_{i_2})$ (a.s.), $1 \le i_1 < i_2 \le n$; $f_{i_1,i_2,i_3}(\xi_{i_1},\xi_{i_2},\xi_{i_3}) = g_{i_1,i_2,i_3}(\xi_{i_1},\xi_{i_2},\xi_{i_3})$ (a.s.), $1 \le i_1 < i_2 < i_3 \le n;\dots, f_{1,2,\dots,n}(\xi_1,\xi_2,\dots,\xi_n) = g_{1,2,\dots,n}(\xi_1,\xi_2,\dots,\xi_n)$ (a.s.), that is $g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}) = f_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c})$ (a.s.), $1 \le i_1 < \cdots < i_c \le n, c = 2,\dots,n$. This completes the proof. □

Proof of Theorems 3.2–4.1. By definition, a function $C : [0,1]^n \to [0,1]$ is a n-dimensional copula if and only if it is a joint cdf of n r.v.'s each of which is uniformly distributed on $[0,1]$. Let, as in Section 3, $V_1,\dots,V_n$ denote independent r.v.'s with uniform distribution on $[0,1]$. From Theorem 2.1 we obtain that $C : [0,1]^n \to [0,1]$ is an absolutely continuous copula if and only if there exist functions $\tilde{g}_{i_1,\dots,i_c} : \mathbf{R}^c \to \mathbf{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2,\dots,n$, satisfying the conditions

$$E|\tilde{g}_{i_1,\dots,i_c}(V_{i_1},\dots,V_{i_c})| < \infty,$$

$$E(\tilde{g}_{i_1,\dots,i_c}(V_{i_1},\dots,V_{i_{k-1}},V_{i_k},V_{i_{k+1}},\dots,V_{i_c})|V_{i_1},\dots,V_{i_{k-1}},V_{i_{k+1}},\dots,V_{i_c}) = 0 \ (a.s.),$$

$1 \le i_1 < \cdots < i_c \le n$, $k = 1,2,\dots,c$, $c = 2,\dots,n$,

$$\sum_{c=2}^{n} \sum_{1\le i_1<\dots<i_c\le n} \tilde{g}_{i_1,\dots,i_c}(V_{i_1},\dots,V_{i_c}) \ge -1 \ (a.s.),$$

and such that representation (3.1) holds. This proves Theorem 3.2. Theorem 3.3 follows from Theorem 3.2 and the relation (given by Sklar's Theorem 3.1) $F_{X_1,\dots,X_n}(x_1,\dots,x_n) = C_{X_1,\dots,X_n}(F_{X_1}(x_1),\dots,F_{X_n}(x_n))$, $x_k \in \mathbf{R}$, $k = 1,\dots,n$, between the joint distribution functions $F_{X_1,\dots,X_n}(x_1,\dots,x_n)$ and the corresponding copulas $C_{X_1,\dots,X_n}(u_1,\dots,u_n)$. If U_n is the U-statistic corresponding to the r.v.'s $X_1,\dots,X_n$, then for any Borel measurable function f for which the expectations exist, one has

$$\begin{aligned} Ef(X_1,\dots,X_n) &= E\{f(\xi_1,\dots,\xi_n)(1 + \sum_{c=2}^{n} \sum_{1\le i_1<\cdots<i_c\le n} g_{i_1,\dots,i_c}(\xi_{i_1},\dots,\xi_{i_c}))\} \\ &= E\left[f(\xi_1,\dots,\xi_n)(1 + U_n(\xi_1,\dots,\xi_n))\right]. \end{aligned}$$

This and Theorem 2.1 implies Theorem 4.1. □

Proof of Theorem 5.1. It is evident that $g_{i_1,\dots,i_c}(x_{i_1},\dots,x_{i_c}) = 0$ satisfy conditions A1–A3. If $g_{i_1,\dots,i_c}(x_{i_1},\dots,x_{i_c}) = 0$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2,\dots,n$, then representation (2.1) takes the form

$$F(x_1,\dots,x_n) = \prod_{i=1}^{n} F_i(x_i). \tag{8.6}$$

R.v's with the joint distribution function (8.6) are independent. Let now $X_1, \dots, X_n$ be independent r.v.'s with one-dimensional distribution functions $F_i(x_i), i = 1, \dots, n$. Then their joint distribution function has form (8.6). This and the uniqueness of the functions $g_{i_1,\dots,i_c}$ given by Theorem 2.1 completes the proof of the theorem. □

Proof of Theorems 5.2–5.8. Below, we give proofs of Theorems 5.7 and 5.8. The rest of the theorems can be proven in a similar way. Let $X_1, \dots, X_n$ be r.v.'s with the joint distribution function satisfying representation (2.2) with functions $g_{i_1,\dots,i_c}$ such that $E\xi_{i_1}^{\alpha_{i_1}} \cdots \xi_{i_c}^{\alpha_{i_c}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \dots, n$, where $\xi_1, \dots, \xi_n$ are independent copies of X_k, $k = 1, 2, \dots, n$. Let $\alpha_j \in \{0, 1, \dots, \alpha\}$, $j = 1, \dots, n$. Taking in (4.2) $f(x_1, \dots, x_n) = \prod_{j=1}^n x_j^{\alpha_j}$ and using independence of the r.v.'s $\xi_1, \dots, \xi_n$ we obtain

$$\begin{aligned} E\prod_{j=1}^n X_j^{\alpha_j} &= E\prod_{j=1}^n \xi_j^{\alpha_j} + \sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} E\prod_{j=1}^n \xi_j^{\alpha_j} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) \\ &= \prod_{j=1}^n E\xi_j^{\alpha_j} + \sum_{c=2}^n \sum_{1\le i_1<\cdots<i_c\le n} E\prod_{k=1}^c \xi_{i_k}^{\alpha_{i_k}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) \\ &\qquad \times \prod_{\substack{k=1,\dots,n \\ k\ne i_1,\dots,i_c}} E\xi_k^{\alpha_k}. \end{aligned} \tag{8.7}$$

Using $E\xi_{i_1}^{\alpha_{i_1}} \cdots \xi_{i_c}^{\alpha_{i_c}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0$, we get from (8.7) $E\prod_{j=1}^n X_j^{\alpha_j} = \prod_{j=1}^n E\xi_j^{\alpha_j} = \prod_{j=1}^n EX_j^{\alpha_j}$, that is the r.v.'s $X_1, \dots, X_n$ form a multiplicative system of order α. Let us suppose now that r.v.'s $X_1, \dots, X_n$ form a multiplicative system of order α, that is, $E|X_j|^\alpha < \infty$, $j = 1, \dots, n$, and for all $\alpha_j \in \{0, 1, \dots, \alpha\}$, $j = 1, \dots, n$, $E\prod_{j=1}^n X_j^{\alpha_j} = \prod_{j=1}^n EX_j^{\alpha_j}$. From Remark 2.1 and Theorem 4.1 it follows that

$$E\prod_{r=1}^k X_{j_r}^{\alpha_{j_r}} = \prod_{r=1}^k EX_{j_r}^{\alpha_{j_r}} + \sum_{c=2}^k \sum_{i_1<\cdots<i_c\in\{j_1,\dots,j_k\}} \prod_{r=1}^k \xi_{j_r}^{\alpha_{j_r}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}),$$

$\alpha_{j_r} \in \{0, 1, \dots, \alpha\}$, $1 \le j_1 < \cdots < j_k \le n$, $k = 2, \dots, n$. Therefore, for all $\alpha_{j_r} \in \{0, 1, \dots, \alpha\}$, $1 \le j_1 < \cdots < j_k \le n$, $k = 2, \dots, n$,

$$\sum_{c=2}^k \sum_{i_1<\cdots<i_c\in\{j_1,\dots,j_k\}} E\prod_{r=1}^c \xi_{i_r}^{\alpha_{i_r}} g_{i_1,\dots,i_c} \prod_{\substack{r=1,\dots,k \\ j_r\ne i_1,\dots,i_c}} E\xi_{j_r}^{\alpha_{j_r}} = 0. \tag{8.8}$$

From (8.8) we subsequently obtain that

$$E\xi_{i_1}^{\alpha_{i_1}} \xi_{i_2}^{\alpha_{i_2}} g_{i_1,i_2}(\xi_{i_1}, \xi_{i_2}) = 0, \alpha_{i_1}, \alpha_{i_2} \in \{0, 1, \dots, \alpha\}, 1 \le i_1 < i_2 \le n;$$

$$E\xi_{i_1}^{\alpha_{i_1}} \xi_{i_2}^{\alpha_{i_2}} \xi_{i_3}^{\alpha_{i_3}} g_{i_1,i_2,i_3}(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) = 0, \alpha_{i_1}, \alpha_{i_2}, \alpha_{i_3} \in \{0, 1, \dots, \alpha\}, 1 \le i_1 < i_2 < i_3 \le n \text{ and } E\xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n} g_{1,2,\dots,n}(\xi_1, \xi_2, \dots, \xi_n), \alpha_k \in \{0, 1, \dots, \alpha\}, k = 1, \dots, n.$$

Therefore, $E\xi_{i_1}^{\alpha_{i_1}} \cdots \xi_{i_c}^{\alpha_{i_c}} g_{i_1,\dots,i_c}(\xi_{i_1}, \dots, \xi_{i_c}) = 0$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \dots, n$. Let $X_1, \dots, X_n$ be r-independent r.v.'s. From Remark 2.1 it follows that

$$\begin{aligned} &F(x_{j_1}, \dots, x_{j_k}) \\ &\quad = \int_{-\infty}^{x_{j_1}} \cdots \int_{-\infty}^{x_{j_k}} \Big(1 + \sum_{c=2}^n \sum_{i_1<\dots<i_c\in B_k} g_{i_1,\dots,i_c}(t_{i_1}, \dots, t_{i_c})\Big) \prod_{i=1}^k dF_{j_i}(t_{j_i}) \end{aligned}$$

for all $B_k = \{1 \le j_1 < \cdots < j_k \le n\}$, $k = 1, \ldots, r$. Using r-independence of X_i's we subsequently obtain from here that $g_{i_1,i_2}(\xi_{i_1}, \xi_{i_2}) = 0$ (a.s.), $1 \le i_1 < i_2 \le n$; $g_{i_1,i_2,i_3}(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) = 0$ (a.s.), $1 \le i_1 < i_2 < i_3 \le n; \ldots, g_{i_1,\ldots,i_r}(\xi_{i_1}, \ldots, \xi_{i_r}) = 0$ (a.s.), $1 \le i_1 < \cdots < i_r \le n$. Therefore, $g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c}) = 0$ (a.s.), $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \ldots, r$. Let now $X_1, \ldots, X_n$ be r.v.'s such that the functions $g_{i_1,\ldots,i_c}$ in representations (2.1) and (2.2) satisfy the conditions $g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c}) = 0$ (a.s.), $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \ldots, r$, that is their joint distribution function has the form

$$\begin{aligned} &F(x_1, \ldots, x_n) \\ &\quad = P(X_1 \le x_1, \ldots, X_n \le x_n) \\ &\quad = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} (1 + \sum_{c=r+1}^{n} \sum_{1 \le i_1 < \cdots < i_c \le n} g_{i_1,\ldots,i_c}(t_{i_1}, \ldots, t_{i_c})) \prod_{i=1}^{n} F_i(t_i). \end{aligned}$$

Let $1 \le j_1 < \cdots < j_r \le n$. Let us show that the r.v.'s $X_{j_1}, \ldots, X_{j_r}$ are jointly independent. Without loss of generality, it suffices to consider the case $j_1 = 1, \ldots, j_r = r$. We have, similar to the proof of Theorem 2.1,

$$\begin{aligned} &F(x_1, \ldots, x_r) \\ &\quad = P(X_1 \le x_1, \ldots, X_r \le x_r) \\ &\quad = \underbrace{\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_r} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} \left(1 + \sum_{c=r+1}^{n} \sum_{1 \le i_1 < \cdots < i_c \le n} g_{i_1,\ldots,i_c}(t_{i_1}, \ldots, t_{i_c}) \right) \\ &\qquad\qquad \times \prod_{i=1}^{n} dF_i(t_i) \\ &\quad = \prod_{i=1}^{r} F_i(x_i) + \sum_{c=r+1}^{n} \sum_{1 \le i_1 < \cdots < i_c \le n} g_{i_1,\ldots,i_c}(t_{i_1}, \ldots, t_{i_c}) \prod_{i=1}^{n} F_i(t_i) \\ &\quad = \prod_{i=1}^{r} F_i(x_i) + \Sigma''. \end{aligned}$$

It is easy to see that there is at least one t_s of $t_{r+1}, \ldots, t_n$ among the arguments of each of the functions $g_{i_1,\ldots,i_c}(t_{i_1}, \ldots, t_{i_c})$, $1 \le i_1 < \cdots < i_c \le n$, $c = r+1, \ldots, n$, in the latter summand and, therefore, by A2, $\Sigma'' = 0$. Consequently, $F(x_1, \ldots, x_r) = \prod_{i=1}^{r} F_i(x_i)$ that establishes joint independence of the r.v.'s $X_1, \ldots, X_r$. The proof is complete. □

Proof of Theorem 6.1. Evidently, if the r.v.'s $X_1, \ldots, X_n$ are jointly independent, then they form a multiplicative system of order α. Let us show that if $card(A_i) \le \alpha + 1$, $i = 1, \ldots, n$, and r.v.'s $X_1, \ldots, X_n$ form a multiplicative system of order α, then they are jointly independent. It suffices to show that

$$E \prod_{i=1}^{n} f_i(X_i) = \prod_{i=1}^{n} E f_i(X_i) \tag{8.9}$$

for all continuous functions $f_i : \mathbf{R} \to \mathbf{R}$, $i = 1, \ldots, n$, vanishing outside a finite interval. Let $1 \le i \le n$. It is easy to see that if $f_i(x)$, $x \in \mathbf{R}$, is an arbitrary function, then there exists a polynomial $r_i(x)$, $x \in \mathbf{R}$, of degree not greater than α, such that $f_i(x) = r_i(x)$, $x \in A_i$, that is $f_i(\xi_i) = r_i(\xi_i)$ (a.s.). Using Theorems

4.1 and 5.7, we get that for all continuous functions $f_i : \mathbf{R} \to \mathbf{R}$ (below, $r_i(x)$ are polynomials corresponding to $f_i(x)$)

$$\begin{aligned} E\prod_{i=1}^{n} f_i(X_i) &= E\prod_{i=1}^{n} f_i(\xi_i) + \sum_{c=2}^{n} \sum_{1\le i_1<\cdots<i_c\le n} E\prod_{i=1}^{n} f_i(\xi_i) g_{i_1,\ldots,i_c}(\xi_{i_1},\ldots,\xi_{i_c}) \\ &= E\prod_{i=1}^{n} f_i(\xi_i) + \sum_{c=2}^{n} \sum_{1\le i_1<\cdots<i_c\le n} E\prod_{i=1}^{n} r_i(\xi_i) g_{i_1,\ldots,i_c}(\xi_{i_1},\ldots,\xi_{i_c}) \\ &= E\prod_{i=1}^{n} f_i(\xi_i). \end{aligned}$$

The proof is complete. □

Proof of Theorem 7.1. From concavity of the functions $\log(1+x)$, relation (7.2) and the inequality $\log(1+x) \le x$, $x \ge 0$, we get $\delta_{X_1,\ldots,X_n} = E\log(1 + U_n(X_1,\ldots,X_n)) \le \log(1 + EU_n(X_1,\ldots,X_n)) = \log(1+\phi^2_{X_1,\ldots,X_n}) \le \phi^2_{X_1,\ldots,X_n}$. The proof is complete. □

Proof of Theorem 7.2. Let $h(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1}) \xrightarrow{\mathcal{D}} Y$. By Theorem 4.1 we have that for any continuous bounded function $g : \mathbf{R} \to \mathbf{R}$

$$Eg(h(X_t,\ldots,X_{t+m-1})) = Eg(h(\xi_t,\ldots,\xi_{t+m-1}))(1 + U_m(\xi_t,\ldots,\xi_{t+m-1})).$$

By Chebyshev's inequality and (7.2) we have that for all $\epsilon > 0$,

$$P(|U_m(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1})| > \epsilon) \le \phi_t^2/\epsilon^2.$$

Since the function $w(x) = (1+x)\ln(1+x) - x$ is increasing in $x \in [0,\infty)$ and decreasing in $x \in (-\infty, 0)$, we have that if $U_m(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1}) > \epsilon$ or $U_m(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1}) < -\epsilon$, then

$$w(U_m(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1})) > (w(\epsilon) \wedge w(-\epsilon)),$$

where $a \wedge b = \min(a,b)$. Therefore, by Chebyshev's inequality, (7.1) and since $EU_m(\xi_t,\ldots,\xi_{t+m-1}) = 0$ (by condition A2) we get, for $0 < \epsilon < 1$,

$$\begin{aligned} &P(|U_m(\xi_t,\ldots,\xi_{t+m-1})| > \epsilon) \\ &\quad\le P(w(U_m(\xi_t,\ldots,\xi_{t+m-1})) > (w(\epsilon) \wedge w(-\epsilon))) \\ &\quad\le Ew(U_m(\xi_t,\xi_{t+1},\ldots,\xi_{t+m-1}))/(w(\epsilon) \wedge w(-\epsilon)) \\ &\quad= E\,(1 + U_m(\xi_t,\ldots,\xi_{t+m-1})) \\ &\qquad\times \log(1 + U_m(\xi_t,\ldots,\xi_{t+m-1}))\,/(w(\epsilon) \wedge w(-\epsilon)) \\ &\quad= \delta_t/(w(\epsilon) \wedge w(-\epsilon)). \end{aligned} \tag{8.10}$$

If $\epsilon \ge 1$, Chebyshev's inequality and $U_m(\xi_t,\ldots,\xi_{t+m-1}) \ge -1$ yield

$$P(|U_m(\xi_t,\ldots,\xi_{t+m-1})| > \epsilon) \le \frac{Ew(U_m(\xi_t,\ldots,\xi_{t+m-1}))}{w(\epsilon)} = \delta_t/w(\epsilon). \tag{8.11}$$

Similar to the above, by Chebyshev's inequality and (7.3), for $0 < \epsilon < 1$,

$$\begin{aligned} P(|U_m(\xi_t,\ldots,\xi_{t+m-1})| > \epsilon) &\le P(\psi(1+U_m(\xi_t,\ldots,\xi_{t+m-1})) > (\psi(1+\epsilon) \wedge \psi(1-\epsilon)) \\ &\le E\psi(1 + U_m(\xi_t,\ldots,\xi_{t+m-1}))/(\psi(1+\epsilon) \wedge \psi(1-\epsilon)) \\ &= D_t^{\psi}/(\psi(1+\epsilon) \wedge \psi(1-\epsilon)). \end{aligned} \tag{8.12}$$

For $\epsilon \geq 1$,

$$P(|U_m(\xi_t,\dots,\xi_{t+m-1})| > \epsilon) \leq P(\psi(1+U_m(\xi_t,\dots,\xi_{t+m-1})) > \psi(1+\epsilon)) \leq D_t^{\psi}/\psi(1+\epsilon). \tag{8.13}$$

Inequalities (8.10)–(8.13) imply that $U_m(\xi_t,\xi_{t+1},\dots,\xi_{t+m-1}) \to 0$ (in probability) as $t \to \infty$, if $\phi_t^2 \to 0$, or $\delta_t \to 0$, or $D_t^{\psi} \to 0$ as $t \to \infty$. The same argument as in the case of the measure D_t^{ψ}, used with $\psi(x) = x^{1-q}$, establishes that $U_m(\xi_t,\xi_{t+1},\dots,\xi_{t+m-1}) \to 0$ (in probability) as $t \to \infty$, if $\rho_t^{(q)} \to 0$ as $t \to \infty$ for $q \in (0,1)$. In particular, the latter holds for the case $q = 1/2$, and, consequently, for the Hellinger distance $\mathcal{H}_t$. The above implies, by Slutsky theorem, that $Eg(h(X_t,X_{t+1},\dots,X_{t+m-1})) \to Eg(Y)$ as $t \to \infty$. Since this holds for any continuous bounded function g, we get $h(X_t,X_{t+1},\dots,X_{t+m-1}) \to Y$ (in distribution) as $t \to \infty$. The proof is complete. The case of double arrays requires only minor notational modifications. □

Proof of Theorem 7.3. From Theorem 4.1, relation (7.2) and Hölder inequality we obtain that for any $x \in \mathbf{R}$ and r.v.'s $X_1,\dots,X_n$

$$\begin{aligned} &P(h(X_1,\dots,X_n)\leq x)-P(h(\xi_1,\dots,\xi_n)\leq x) \\ &\quad = EI(h(\xi_1,\dots,\xi_n)\leq x)U_n(\xi_1,\dots,\xi_n) \\ &\quad \leq \phi_{X_1,\dots X_n}(P(h(\xi_1,\dots,\xi_n)\leq x))^{1/2} \end{aligned} \tag{8.14}$$

$$\begin{aligned} &P(h(X_1,\dots,X_n)>x)-P(h(\xi_1,\dots,\xi_n)>x) \\ &\quad = EI(h(\xi_1,\dots,\xi_n)>x)U_n(\xi_1,\dots,\xi_n) \\ &\quad \leq \phi_{X_1,\dots,X_n}(P(h(\xi_1,\dots,\xi_n)> x))^{1/2} \end{aligned} \tag{8.15}$$

The latter inequalities imply that for any $x \in \mathbf{R}$

$$\begin{aligned} &|P(h(X_1,\dots,X_n)\leq x) - P(h(\xi_1,\dots,\xi_n)\leq x)| \\ &\quad \leq \phi_{X_1,\dots,X_n} \max\left[(P(h(\xi_1,\dots,\xi_n)\leq x))^{1/2}, (P(h(\xi_1,\dots,\xi_n)> x))^{1/2}\right]. \end{aligned}$$

The proof is complete. □

Proof of Theorem 7.4. By Theorem 4.1 we have $Ef(X_1,\dots,X_n) = Ef(\xi_1,\dots,\xi_n) + EU_n(\xi_1,\dots,\xi_n)f(\xi_1,\dots,\xi_n)$. By Cauchy–Schwarz inequality and relation (7.2) we get

$$EU_n(\xi_1,\dots,\xi_n)f(\xi_1,\dots,\xi_n) \leq \left(EU_n^2(\xi_1,\dots,\xi_n)\right)^{1/2}\left(Ef^2(\xi_1,\dots,\xi_n)\right)^{1/2}.$$

Therefore, (7.7) holds. Sharpness of (7.7) follows from the choice of independent $X_1,\dots,X_n$. Similarly, from Hölder inequality it follows that if $q > 1$, $1/p+1/q = 1$, then

$$Ef(X_1,\dots,X_n) \leq (E(1+U_n(\xi_1,\dots,\xi_n))^p)^{1/p}(Ef(\xi_1,\dots,\xi_n))^q)^{1/q}. \tag{8.16}$$

This implies (7.10). If in estimate (8.16) $q \geq 2$ and, therefore, $p \in (1,2]$, by Theorem 4.1, Jensen inequality and relation (7.2) we have

$$\begin{aligned} E(1+U_n(\xi_1,\dots,\xi_n))^p &= E(1+U_n(X_1,\dots,X_n))^{p-1} \\ &\leq (1+EU_n(X_1,\dots,X_n))^{p-1} \\ &= (1+\phi^2_{X_1,\dots,X_n})^{p/q}. \end{aligned}$$

Therefore, (7.8) holds. Sharpness of (7.8) and (7.10) follows from the choice of $X_i = const$ (a.s.), $i = 1, \ldots, n$. According to Young's inequality (see [19, p. 512]), if $p : [0, \infty) \to [0, \infty)$ is a non-decreasing right-continuous function satisfying $p(0) = \lim_{t \to 0+} p(t) = 0$ and $p(\infty) = \lim_{t \to \infty} p(t) = \infty$, and $q(t) = \sup\{u : p(u) \le t\}$ is a right-continuous inverse of p, then

$$st \le \phi(s) + \psi(t), \tag{8.17}$$

where $\phi(t) = \int_0^t p(s)ds$ and $\psi(t) = \int_0^t q(s)ds$. Using (8.17) with $p(t) = ln(1+t)$ and (7.1), we get that

$$\begin{aligned} EU_n(\xi_1, \ldots, \xi_n) f(\xi_1, \ldots, \xi_n) \;&\le\; E(e^{f(\xi_1, \ldots, \xi_n)}) - 1 - Ef(\xi_1, \ldots, \xi_n) \\ &\quad + E(1 + U_n(\xi_1, \ldots, \xi_n)) \log(1 + U_n(\xi_1, \ldots, \xi_n)) \\ &= \; E(e^{f(\xi_1, \ldots, \xi_n)}) - 1 - Ef(\xi_1, \ldots, \xi_n) \\ &\quad + \delta_{X_1, \ldots, X_n}. \end{aligned}$$

This establishes (7.9). Sharpness of (7.9) follows, e.g., from the choice of independent X_i's and $f \equiv 0$. □

Proof of Theorem 7.5. The theorem follows from inequalities (7.7)–(7.10) applied to $f(x_1, \ldots, x_n) = I(h(x_1, \ldots, x_n) > x)$. □

Acknowledgements

The authors are grateful to Peter Phillips, two anonymous referees, the editor, and the participants at the Prospectus Workshop at the Department of Economics, Yale University, in 2002-2003 for helpful comments and suggestions. We also thank the participants at the Third International Conference on High Dimensional Probability, June 2002, and the 28th Conference on Stochastic Processes and Their Applications at the University of Melbourne, July 2002, where some of the results in the paper were presented.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, B. N. Petrov and F. Caski, eds. Akademiai Kiado, Budapest, 267–281 (reprinted in: *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe and G. Kitagawa, eds., Springer Series in Statistics: Perspectives in Statistics. Springer-Verlag, New York, 1998, pp. 199–213).

[2] Alexits, G. (1961). *Convergence Problems of Orthogonal Series*. International Series of Monographs in Pure and Applied Mathematics, Vol. 20, Pergamon Press, New York–Oxford–Paris.

[3] Ali, S. M., and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28**, 131–142.

[4] Ang, A. and Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics* **63**, 443–494.

[5] Barndorff-Nielsen, O. E. and Shephard, N. (2001). Modeling by Lévy processes for financial econometrics. In *Lévy Processes. Theory and Applications* (Barndorff-Nielsen, O. E., Mikosch, T. and Resnick, S. I., eds.). Birkhäuser, Boston, 283–318.

[6] Beneš, V. and Štěpán, J. (eds.) (1997). *Distributions with Given Marginals and Moment Problems*. Kluwer Acad. Publ., Dordrecht.
[7] Blyth, S. (1996). Out of line. *Risk* **9**, 82–84.
[8] Borovskikh, Yu. V. and Korolyuk, V. S. (1997). *Martingale Approximation*. VSP, Utrecht.
[9] Boyer, B. H., Gibson, M. S. and Loretan, M. (1999). Pitfalls in tests for changes in correlations. Federal Reserve Board, IFS Discussion Paper No. 597R.
[10] Cambanis, S. (1977). Some properties and generalizations of multivariate Eyraud–Gumbel–Morgenstern distributions. *J. Multivariate Anal.* **7**, 551–559.
[11] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* **1**, 223–236.
[12] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.
[13] Dall'Aglio, G., Kotz, S. and Salinetti, G. (eds.) (1991). *Advances in Probability Distributions with Given Marginals*. Kluwer Acad. Publ., Dordrecht.
[14] de la Peña, V. H. (1990). Bounds on the expectation of functions of martingales and sums of positive RVs in terms of norms of sums of independent random variables. *Proc. Amer. Math. Soc.* **108**, 233–239.
[15] de la Peña, V. H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Probability and Its Applications. Springer, New York.
[16] de la Peña, V. H., Ibragimov, R. and Sharakhmetov, S. (2002). On sharp Burkholder–Rosenthal-type inequalities for infinite-degree U-statistics. *Ann. Inst. H. Poincaré Probab. Statist.* **38**, 973–990.
[17] de la Peña, V. H., Ibragimov, R. and Sharakhmetov, S. (2003). On extremal distributions and sharp L_p-bounds for sums of multilinear forms. *Ann. Probab.* **31**, 630–675.
[18] de la Peña, V. H. and Lai, T. L. (2001). Theory and applications of decoupling. In *Probability and Statistical Models with Applications* (Ch. A. Charalambides, M. V. Koutras and N. Balakrishnan, eds.), Chapman and Hall/CRC, New York, 117–145.
[19] Dilworth, S. J. (2001). Special Banach lattices and their applications. In *Handbook of the Geometry of Banach Spaces*, Vol. I. North-Holland, Amsterdam, 497–532.
[20] Dragomir, S. S. (2000). An inequality for logarithmic mapping and applications for the relative entropy. *Nihonkai Math. J.* **11**, 151–158.
[21] Embrechts, P., Lindskog, F. and McNeil, A. (2001). Modeling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance* (S. Rachev, ed.). Elsevier, 329–384, Chapter 8.
[22] Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk manage- ment: properties and pitfalls. In *Risk Management: Value at Risk and Beyond* (M. A. H. Dempster, ed.). Cambridge University Press, Cambridge, 176–223.
[23] Fackler, P. (1991). Modeling interdependence: an approach ro simulation and elicitation. *American Journal of Agricultural Economics* **73**, 1091–1097.
[24] Fernandes, M. and Flôres, M. F. (2001). Tests for conditional independence. Working paper, `http://www.vwl.uni-mannheim.de/brownbag/flores.pdf`

[25] Frees, E., Carriere, J. and Valdez, E. (1996). Annuity valuation with dependent mortality. *Journal of Risk and Insurance* **63**, 229–261.

[26] Golan, A. (2002). Information and entropy econometrics – editor's view. *J. Econometrics* **107**, 1–15.

[27] Golan, A. and Perloff, J. M. (2002). Comparison of maximum entropy and higher-order entropy estimators. *J. Econometrics* **107**, 195–211.

[28] Gouriéroux, C. and Monfort, A. (1979). On the characterization of a joint probability distribution by conditional distributions. *J. Econometrics* **10**, 115–118.

[29] Granger, C. W. J. and Lin, J. L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Ser. Anal.* **15**, 371–384.

[30] Granger, C. W. J., Teräsvirta, T. and Patton, A. J. (2002). Common factors in conditional distributions. Univ. Calif., San Diego, Discussion Paper 02-19; Economic Research Institute, Stockholm School of Economics, Working Paper 515.

[31] Hong, Y.-H. and White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* **73**, 873–901.

[32] Hu, L. (2006). Dependence patterns across financial markets: A mixed copula approach. *Appl. Financial Economics* **16** 717–729.

[33] Ibragimov, R. (2004). On the robustness of economic models to heavy-tailedness assumptions. Mimeo, Yale University. Available at `http://post.economics.harvard.edu/faculty/ibragimov/Papers/HeavyTails.pdf`.

[34] Ibragimov, R. (2005). New majorization theory in economics and martingale convergence results in econometrics. Ph.D. dissertation, Yale University.

[35] Ibragimov, R. and Phillips, P. C. B. (2004). Regression asymptotics using martingale convergence methods. Cowles Foundation Discussion Paper 1473, Yale University. Available at `http://cowles.econ.yale.edu/P/cd/d14b/d1473.pdf`

[36] Ibragimov, R. and Sharakhmetov, S. (1997). On an exact constant for the Rosenthal inequality. *Theory Probab. Appl.* **42**, 294–302.

[37] Ibragimov, R. and Sharakhmetov, S. (1999). Analogues of Khintchine, Marcinkiewicz–Zygmund and Rosenthal inequalities for symmetric statistics. *Scand. J. Statist.* **26**, 621–623.

[38] Ibragimov, R. and Sharakhmetov, S. (2001a). The best constant in the Rosenthal inequality for nonnegative random variables. *Statist. Probab. Lett.* **55**, 367–376.

[39] Ibragimov R. and Sharakhmetov, S. (2001b). The exact constant in the Rosenthal inequality for random variables with mean zero. *Theory Probab. Appl.* **46**, 127–132.

[40] Ibragimov, R. and Sharakhmetov, S. (2002). Bounds on moments of symmetric statistics. *Studia Sci. Math. Hungar.* **39**, 251–275.

[41] Ibragimov R., Sharakhmetov S. and Cecen A. (2001). Exact estimates for moments of random bilinear forms. *J. Theoret. Probab.* **14**, 21–37.

[42] Joe, H. (1987). Majorization, randomness and dependence for multivariate distributions. *Ann. Probab.* **15**, 1217–1225.

[43] Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.* **84**, 157–164.

[44] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Monographs on Statistics and Applied Probability, Vol. **73**. Chapman & Hall, London.
[45] Johnson, N. L. and Kotz, S. (1975). On some generalized Farlie–Gumbel–Morgenstern distributions. *Comm. Statist.* **4**, 415–424.
[46] Klugman, S., and Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance Math. Econom.* **24**, 139–148.
[47] Kotz, S. and Seeger, J. P. (1991). A new approach to dependence in multivariate distributions. In: *Advances in Probability Distributions with Given Marginals* (Rome, 1990). Mathematics and Its Applications, Vol. **67**. Kluwer Acad. Publ., Dordrecht, 113–127.
[48] Kwapień, S. (1987). Decoupling inequalities for polynomial chaos. *Ann. Probab.* **15**, 1062–1071.
[49] Lancaster, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* **29**, 719–736. Corrig. **35** (1964) 1388.
[50] Lancaster, H. O. (1963). Correlations and canonical forms of bivariate distributions. *Ann. Math. Statist.* **34**, 532–538.
[51] Lancaster, H. O. (1969). *The chi-Squared Distribution*. Wiley, New York.
[52] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37**, 1137–1153.
[53] Long, D. and Krzysztofowicz, R. (1995) A family of bivariate densities constructed from marginals. *J. Amer. Statist. Assoc.* **90**, 739–746.
[54] Longin, F. and Solnik, B. (2001). Extreme Correlation of International Equity Markets. *J. Finance* **56**, 649–676.
[55] Loretan, M. and Phillips, P. C. B. (1994). Testing the covariance stationarity of heavy-tailed time series. *Journal of Empirical Finance* **3**, 211–248.
[56] Mari, D. D. and Kotz, S. (2001). *Correlation and Dependence*. Imp. Coll. Press, London.
[57] Massoumi, E. and Racine, J. (2002). Entropy and predictability of stock market returns. *J. Econometrics* **107**, 291–312.
[58] Miller, D. J., and Liu, W.-H. (2002). On the recovery of joint distributions from limited information. *J. Econometrics* **107**, 259–274.
[59] Mond, B. and Pečarić, J. (2001). On some applications of the AG inequality in information theory. *JIPAM. J. Inequal. Pure Appl. Math.* **2**, Article 11.
[60] Nelsen, R. B. (1999). *An introduction to copulas*. Lecture Notes in Statistics, Vol. **139**. Springer-Verlag, New York.
[61] Patton, A. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *J. Financial Econometrics* **2**, 130–168.
[62] Patton, A. (2006). Modelling asymmetric exchange rate dependence. *Internat. Economic Rev.* **47**, 527–556.
[63] Pearson, K. (1904). Mathematical contributions in the theory of evolution, XIII: On the theory of contingency and its relation to association and normal correlation. In *Drapers' Company Research Memoirs* (Biometric Series I), London: University College (reprinted in *Early Statistical Papers* (1948) by the Cambridge University Press, Cambridge, U.K.).
[64] Reiss, R. and Thomas, M. (2001). *Statistical Analysis of Extreme Values. From Insurance, Finance, Hydrology and Other Fields.* Birkhäuser, Basel.
[65] Richardson, J., Klose, S. and Gray, A. (2000). An applied procedure for estimating and simulating multivariate empirical (MVE) probability distributions in farm-level risk assessment and policy analysis. *Journal of Agricultural and Applied Economics* **32**, 299–315.

[66] Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Rev. Econom. Stud.* **58**, 437–453.
[67] Rüschendorf, L. (1985). Construction of multivariate distributions with given marginals. *Ann. Inst. Statist. Math.* **37**, Part A, 225–233.
[68] Sharakhmetov, S. (1993). r-independent random variables and multiplicative systems (in Russian). *Dopov. Dokl. Akad. Nauk Ukraïni*, 43–45.
[69] Sharakhmetov, S. (2001). On a problem of N. N. Leonenko and M. I. Yadrenko (in Russian) *Dopov. Nats. Akad. Nauk Ukr. Mat. Prirodozn. Tekh. Nauki*, 23–27.
[70] Sharakhmetov, S. and Ibragimov, R. (2002). A characterization of joint distribution of two-valued random variables and its applications. *J. Multivariate Anal.* **83**, 389–408.
[71] Shaw, J. (1997). Beyod VAR and stress testing. In *VAR: Understanding and Applying Value at Risk*. Risk Publications, London, 211–224.
[72] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–231.
[73] Soofi, E. S. and Retzer, J. J. (2002). Information indices: unification and applications. *J. Econometrics* **107**, 17–40.
[74] Taylor, C. R. (1990). Two practical procedures for estimating multivariate nonnormal probability density functions. *American Journal of Agricultural Economics* **72**, 210–217.
[75] Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *J. Statist. Phys.* **52**, 479–487.
[76] Ullah, A. (2002). Uses of entropy and divergence measures for evaluating econometric approximations and inference. *J. Econometrics* **107**, 313–326.
[77] Wang, Y. H. (1990). Dependent random variables with independent subsets II. *Canad. Math. Bull.* **33**, 22–27.
[78] Zolotarev, V. M. (1991). Reflection on the classical theory of limit theorems. I. *Theory Probab. Appl.* **36**, 124–137.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 210–228
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000464

Regression tree models for designed experiments*

Wei-Yin Loh[1]

University of Wisconsin, Madison

Abstract: Although regression trees were originally designed for large datasets, they can profitably be used on small datasets as well, including those from replicated or unreplicated complete factorial experiments. We show that in the latter situations, regression tree models can provide simpler and more intuitive interpretations of interaction effects as differences between conditional main effects. We present simulation results to verify that the models can yield lower prediction mean squared errors than the traditional techniques. The tree models span a wide range of sophistication, from piecewise constant to piecewise simple and multiple linear, and from least squares to Poisson and logistic regression.

1. Introduction

Experiments are often conducted to determine if changing the values of certain variables leads to worthwhile improvements in the mean yield of a process or system. Another common goal is estimation of the mean yield at given experimental conditions. In practice, both goals can be attained by fitting an accurate and interpretable model to the data. Accuracy may be measured, for example, in terms of prediction mean squared error, PMSE $= \sum_i E(\hat{\mu}_i - \mu_i)^2$, where μ_i and $\hat{\mu}_i$ denote the true mean yield and its estimated value, respectively, at the ith design point.

We will restrict our discussion here to complete factorial designs that are unreplicated or are equally replicated. For a replicated experiment, the standard analysis approach based on significance tests goes as follows. (i) Fit a full ANOVA model containing all main effects and interactions. (ii) Estimate the error variance σ^2 and use t-intervals to identify the statistically significant effects. (iii) Select as the "best" model the one containing only the significant effects.

There are two ways to control a given level of significance α: the individual error rate (IER) and the experimentwise error rate (EER) (Wu and Hamda [22, p. 132]). Under IER, each t-interval is constructed to have individual confidence level $1 - \alpha$. As a result, if all the effects are null (i.e., their true values are zero), the probability of concluding at least one effect to be non-null tends to exceed α. Under EER, this probability is at most α. It is achieved by increasing the lengths of the t-intervals so that their simultaneous probability of a Type I error is bounded by α. The appropriate interval lengths can be determined from the studentized maximum modulus distribution if an estimate of σ is available. Because EER is more conservative than IER, the former has a higher probability of discovering the

*This material is based upon work partially supported by the National Science Foundation under grant DMS-0402470 and by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-05-1-0047.

[1]Department of Statistics, 1300 University Avenue, University of Wisconsin, Madison, WI 53706, USA, e-mail: loh@stat.wisc.edu

AMS 2000 subject classifications: primary 62K15; secondary 62G08.

Keywords and phrases: AIC, ANOVA, factorial, interaction, logistic, Poisson.

right model in the null situation where no variable has any effect on the yield. On the other hand, if there are one or more non-null effects, the IER method has a higher probability of finding them. To render the two methods more comparable in the examples to follow, we will use $\alpha = 0.05$ for IER and $\alpha = 0.1$ for EER.

Another standard approach is AIC, which selects the model that minimizes the criterion $\text{AIC} = n \log(\tilde{\sigma}^2) + 2\nu$. Here $\tilde{\sigma}$ is the maximum likelihood estimate of σ for the model under consideration, ν is the number of estimated parameters, and n is the number of observations. Unlike IER and EER, which focus on statistical significance, AIC aims to minimize PMSE. This is because $\tilde{\sigma}^2$ is an estimate of the residual mean squared error. The term 2ν discourages over-fitting by penalizing model complexity. Although AIC can be used on any given collection of models, it is typically applied in a stepwise fashion to a set of *hierarchical* ANOVA models. Such models contain an interaction term only if all its lower-order effects are also included. We use the R implementation of stepwise AIC [14] in our examples, with initial model the one containing all the main effects.

We propose a new approach that uses a recursive partitioning algorithm to produce a set of nested piecewise linear models and then employs cross-validation to select a parsimonious one. For maximum interpretability, the linear model in each partition is constrained to contain main effect terms at most. Curvature and interaction effects are captured by the partitioning conditions. This forces interaction effects to be expressed and interpreted naturally—as contrasts of conditional main effects.

Our approach applies to unreplicated complete factorial experiments too. Quite often, two-level factorials are performed without replications to save time or to reduce cost. But because there is no unbiased estimate of σ^2, procedures that rely on statistical significance cannot be applied. Current practice typically invokes empirical principles such as *hierarchical ordering, effect sparsity*, and *effect heredity* [22, p. 112] to guide and limit model search. The hierarchical ordering principle states that high-order effects tend to be smaller in magnitude than low-order effects. This allows σ^2 to be estimated by pooling estimates of high-order interactions, but it leaves open the question of how many interactions to pool. The effect sparsity principle states that usually there are only a few significant effects [2]. Therefore the smaller estimated effects can be used to estimate σ^2. The difficulty is that a good guess of the actual number of significant effects is needed. Finally, the effect heredity principle is used to restrict the model search space to hierarchical models.

We will use the GUIDE [18] and LOTUS [5] algorithms to construct our piecewise linear models. Section 2 gives a brief overview of GUIDE in the context of earlier regression tree algorithms. Sections 3 and 4 illustrate its use in replicated and unreplicated two-level experiments, respectively, and present simulation results to demonstrate the effectiveness of the approach. Sections 5 and 6 extend it to Poisson and logistic regression problems, and Section 7 concludes with some suggestions for future research.

2. Overview of regression tree algorithms

GUIDE is an algorithm for constructing piecewise linear regression models. Each piece in such a model corresponds to a partition of the data and the sample space of the form $X \le c$ (if X is numerically ordered) or $X \in A$ (if X is unordered). Partitioning is carried out recursively, beginning with the whole dataset, and the set of partitions is presented as a binary decision tree. The idea of recursive parti-

tioning was first introduced in the AID algorithm [20]. It became popular after the appearance of CART [3] and C4.5 [21], the latter being for classification only.

CART contains several significant improvements over AID, but they both share some undesirable properties. First, the models are piecewise constant. As a result, they tend to have lower prediction accuracy than many other regression models, including ordinary multiple linear regression [3, p. 264]. In addition, the piecewise constant trees tend to be large and hence cumbersome to interpret. More importantly, AID and CART have an inherent bias in the variables they choose to form the partitions. Specifically, variables with more splits are more likely to be chosen than variables with fewer splits. This selection bias, intrinsic to all algorithms based on optimization through greedy search, effectively removes much of the advantage and appeal of a regression tree model, because it casts doubt upon inferences drawn from the tree structure. Finally, the greedy search approach is computationally impractical to extend beyond piecewise constant models, especially for large datasets.

GUIDE was designed to solve both the computational and the selection bias problems of AID and CART. It does this by breaking the task of finding a split into two steps: first find the variable X and then find the split values c or A that most reduces the total residual sum of squares of the two subnodes. The computational savings from this strategy are clear, because the search for c or A is skipped for all except the selected X.

To solve the selection bias problem, GUIDE uses significance tests to assess the fit of each X variable at each node of the tree. Specifically, the values (grouped if necessary) of each X are cross-tabulated with the signs of the linear model residuals and a chi-squared contingency table test is performed. The variable with the smallest chi-squared p-value is chosen to split the node. This is based on the expectation that any effects of X not captured by the fitted linear model would produce a small chi-squared p-value, and hence identify X as a candidate for splitting. On the other hand, if X is independent of the residuals, its chi-squared p-value would be approximately uniformly distributed on the unit interval.

If a constant model is fitted to the node and if all the X variables are independent of the response, each will have the same chance of being selected. Thus there is no selection bias. On the other hand, if the model is linear in some predictors, the latter will have zero correlation with the residuals. This tends to inflate their chi-squared p-values and produce a bias in favor of the non-regressor variables. GUIDE solves this problem by using the bootstrap to shrink the p-values that are so inflated. It also performs additional chi-squared tests to detect local interactions between pairs of variables. After splitting stops, GUIDE employs CART's pruning technique to obtain a nested sequence of piecewise linear models and then chooses the tree with the smallest cross-validation estimate of PMSE. We refer the reader to Loh [18] for the details. Note that the use of residuals for split selection paves the way for extensions of the approach to piecewise nonlinear and non-Gaussian models, such as logistic [5], Poisson [6], and quantile [7] regression trees.

3. Replicated 2^4 experiments

In this and the next section, we adopt the usual convention of letting capital letters A, B, C, etc., denote the names of variables as well as their main effects, and AB, ABC, etc., denote interaction effects. The levels of each factor are indicated in two ways, either by "$-$" and "$+$" signs, or as -1 and $+1$. In the latter notation, the variables A, B, C, ..., are denoted by x_1, x_2, x_3, ..., respectively.

TABLE 1
Estimated coefficients and standard errors for 2^4 experiment

	Estimate	Std. error	t	Pr(>|t|)
Intercept	14.161250	0.049744	284.683	< 2e-16
x1	-0.038729	0.049744	-0.779	0.438529
x2	0.086271	0.049744	1.734	0.086717
x3	-0.038708	0.049744	-0.778	0.438774
x4	0.245021	0.049744	4.926	4.45e-06
x1:x2	0.003708	0.049744	0.075	0.940760
x1:x3	-0.046229	0.049744	-0.929	0.355507
x1:x4	-0.025000	0.049744	-0.503	0.616644
x2:x3	0.028771	0.049744	0.578	0.564633
x2:x4	-0.015042	0.049744	-0.302	0.763145
x3:x4	-0.172521	0.049744	-3.468	0.000846
x1:x2:x3	0.048750	0.049744	0.980	0.330031
x1:x2:x4	0.012521	0.049744	0.252	0.801914
x1:x3:x4	-0.015000	0.049744	-0.302	0.763782
x2:x3:x4	0.054958	0.049744	1.105	0.272547
x1:x2:x3:x4	0.009979	0.049744	0.201	0.841512

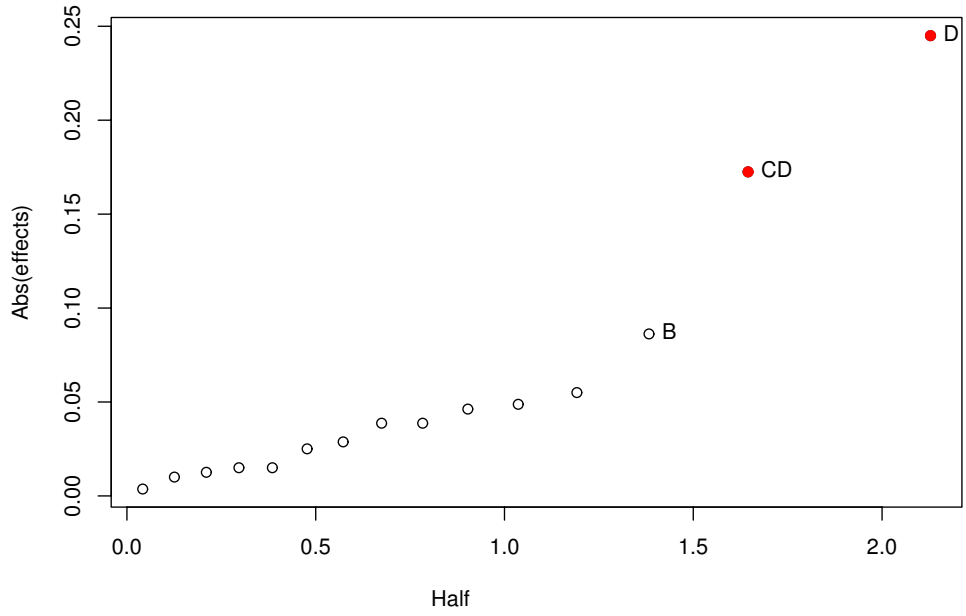


FIG 1. *Half-normal quantile plot of estimated effects from replicated 2^4 silicon wafer experiment.*

We begin with an example from Wu and Hamada [22, p. 97] of a 2^4 experiment on the growth of epitaxial layers on polished silicon wafers during the fabrication of integrated circuit devices. The experiment was replicated six times and a full model fitted to the data yields the results in Table 1.

Clearly, at the 0.05-level, the IER method finds only two statistically significant effects, namely D and CD. This yields the model

$$\hat{y} = 14.16125 + 0.24502x_4 - 0.17252x_3x_4 \tag{3.1}$$

which coincides with that obtained by the EER method at level 0.1.

Figure 1 shows a half-normal quantile plot of the estimated effects. The D and CD effects clearly stand out from the rest. There is a hint of a B main effect, but it is not included in model (3.1) because its p-value is not small enough. The B effect appears, however, in the AIC model

$$\hat{y} = 14.16125 + 0.08627x_2 - 0.03871x_3 + 0.24502x_4 - 0.17252x_3x_4. \tag{3.2}$$

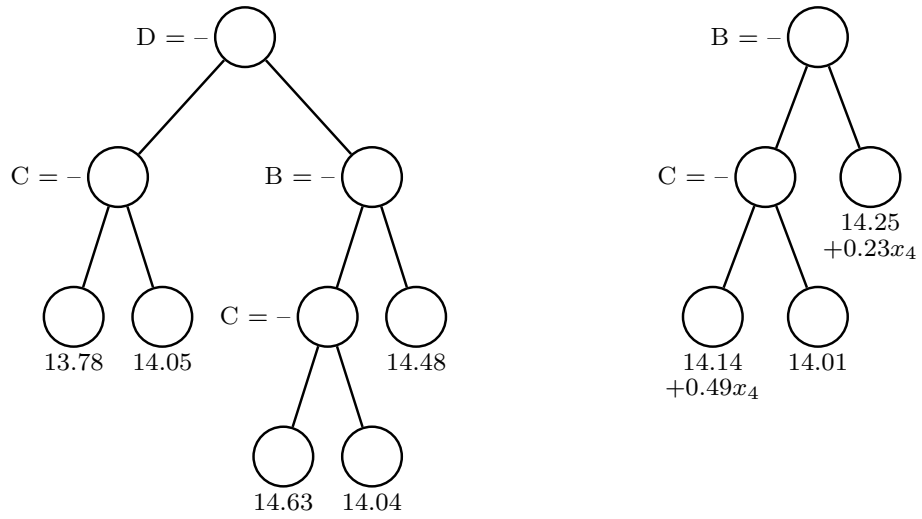


FIG 2. *Piecewise constant (left) and piecewise best simple linear or stepwise linear (right) GUIDE models for silicon wafer experiment. At each intermediate node, an observation goes to the left branch if the stated condition is satisfied; otherwise it goes to the right branch. The fitted model is printed beneath each leaf node.*

Note the presence of the small C main effect. It is due to the presence of the CD effect and to the requirement that the model be hierarchical.

The piecewise constant GUIDE tree is shown on the left side of Figure 2. It has five leaf nodes, splitting first on D, the variable with the largest main effect. If $D = +$, it splits further on B and C. Otherwise, if $D = -$, it splits once on C. We observe from the node sample means that the highest predicted yield occurs when $B = C = -$ and $D = +$. This agrees with the prediction of model (3.1) but not (3.2), which prescribes the condition $B = D = +$ and $C = -$. The difference in the two predicted yields is very small though. For comparison with (3.1) and (3.2), note that the GUIDE model can be expressed algebraically as

$$\begin{aligned}\hat{y} &= 13.78242(1-x_4)(1-x_3)/4 + 14.05(1-x_4)(1+x_3)/4 \\ &\quad + 14.63(1+x_4)(1-x_2)(1-x_3)/8 + 14.4775(1+x_4)(1+x_2)/4 \\ &\quad + 14.0401(1+x_4)(1-x_2)(1+x_3)/8 \\ &= 14.16125 + 0.24502x_4 - 0.14064x_3x_4 - 0.00683x_3 \\ &\quad + 0.03561x_2(x_4+1) + 0.07374x_2x_3(x_4+1).\end{aligned} \tag{3.3}$$

The piecewise best simple linear GUIDE tree is shown on the right side of Figure 2. Here, the data in each node are fitted with a simple linear regression model, using the X variable that yields the smallest residual mean squared error, provided a statistically significant X exists. If there is no significant X, i.e., none with absolute t-statistic greater than 2, a constant model is fitted to the data in the node. In this tree, factor B is selected to split the root node because it has the smallest chi-squared p-value after allowing for the effect of the best linear predictor. Unlike the piecewise constant model, which uses the variable with the largest main effect to split a node, the piecewise linear model tries to keep that variable as a linear predictor. This explains why D is the linear predictor in two of the three leaf nodes

of the tree. The piecewise best simple linear GUIDE model can be expressed as

$$\begin{aligned}\hat{y} &= (14.14246 + 0.4875417x_4)(1 - x_2)(1 - x_3)/4 \\ &\quad + 14.0075(1 - x_2)(1 + x_3)/4 \\ &\quad + (14.24752 + 0.2299792x_4)(1 + x_2)/2 \\ &= 14.16125 + 0.23688x_4 + 0.12189x_3x_4(x_2 - 1) \\ &\quad + 0.08627x_2 + 0.03374x_3(x_2 - 1) - 0.00690x_2x_4.\end{aligned} \tag{3.4}$$

Figure 3, which superimposes the fitted functions from the three leaf nodes, offers a more vivid way to understand the interactions. It shows that changing the level of D from $-$ to $+$ never decreases the predicted mean yield and that the latter varies less if $D = -$ than if $D = +$. The same tree model is obtained if we fit a piecewise multiple linear GUIDE model using forward and backward stepwise regression to select variables in each node.

A simulation experiment was carried out to compare the PMSE of the methods. Four models were employed, as shown in Table 2. Instead of performing the simula-

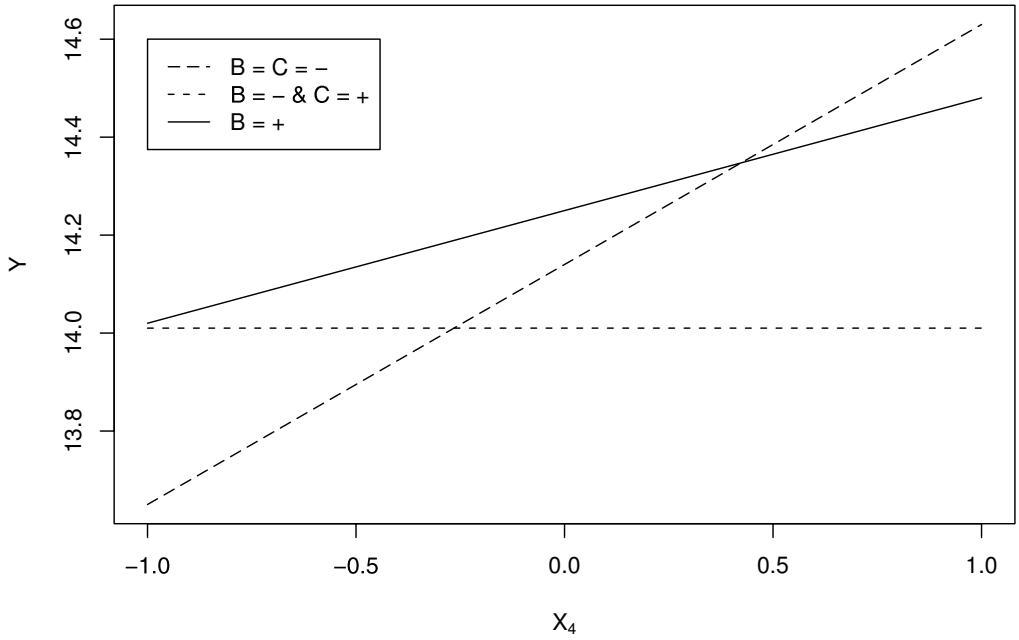


FIG 3. *Fitted values versus* x_4 (D) *for the piecewise simple linear GUIDE model shown on the right side of Figure* 2.

TABLE 2

Simulation models for a 2^4 *design; the* β_i*'s are uniformly distributed and* ε *is normally distributed with mean* 0 *and variance* 0.25*;* $U(a, b)$ *denotes a uniform distribution on the interval* (a, b)*;* ε *and the* β_i*'s are mutually independent*

Name	Simulation model	β distribution
Null	$y = \varepsilon$	
Unif	$y = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1x_4 + \beta_8x_2x_3 + \beta_9x_2x_4 + \beta_{10}x_3x_4 + \beta_{11}x_1x_2x_3 + \beta_{12}x_1x_2x_4 + \beta_{13}x_1x_3x_4 + \beta_{14}x_2x_3x_4 + \beta_{15}x_1x_2x_3x_4 + \varepsilon$	$U(-1/4, 1/4)$
Exp	$y = \exp(\beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon)$	$U(-1, 1)$
Hier	$y = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_1\beta_2x_1x_2 + \beta_1\beta_3x_1x_3 + \beta_1\beta_4x_1x_4 + \beta_2\beta_3x_2x_3 + \beta_2\beta_4x_2x_4 + \beta_3\beta_4x_3x_4 + \beta_1\beta_2\beta_3x_1x_2x_3 + \beta_1\beta_2\beta_4x_1x_2x_4 + \beta_1\beta_3\beta_4x_1x_3x_4 + \beta_2\beta_3\beta_4x_2x_3x_4 + \beta_1\beta_2\beta_3\beta_4x_1x_2x_3x_4 + \varepsilon$	$U(-1, 1)$

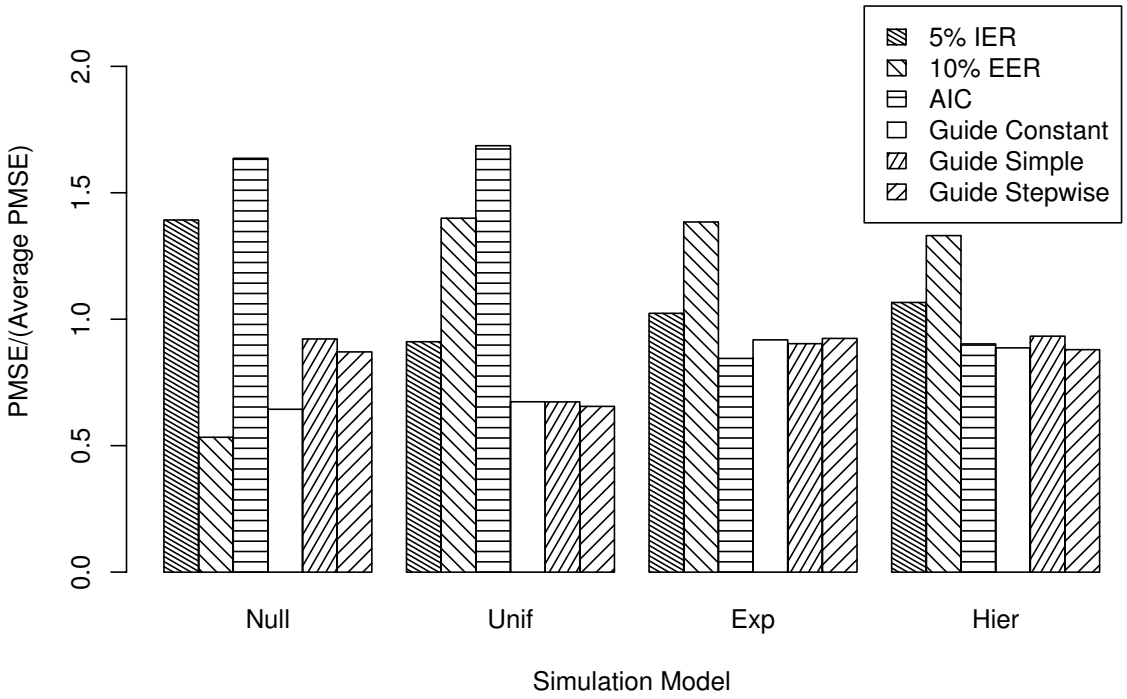


FIG 4. *Barplots of relative PMSE of methods for the four simulation models in Table* 2. *The relative PMSE of a method at a simulation model is defined as its PMSE divided by the average PMSE of the six methods at the same model.*

tions with a fixed set of regression coefficients, we randomly picked the coefficients from a uniform distribution in each simulation trial. The *Null* model serves as a baseline where none of the predictor variables has any effect on the mean yield, i.e., the true model is a constant. The *Unif* model has main and interaction effects independently drawn from a uniform distribution on the interval $(-0.25, 0.25)$. The *Hier* model follows the hierarchical ordering principle—its interaction effects are formed from products of main effects that are bounded by 1 in absolute value. Thus higher-order interaction effects are smaller in magnitude than their lower-order parent effects. Finally, the *Exp* model has non-normal errors and variance heterogeneity, with the variance increasing with the mean.

Ten thousand simulation trials were performed for each model. For each trial, 96 observations were simulated, yielding 6 replicates at each of the 16 factor-level combinations of a 2^4 design. Each method was applied to find estimates, $\hat{\mu}_i$, of the 16 true means, μ_i, and the sum of squared errors $\sum_1^{16}(\hat{\mu}_i - \mu_i)^2$ was computed. The average over the 10,000 simulation trials gives an estimate of the PMSE of the method. Figure 4 shows barplots of the relative PMSEs, where each PMSE is divided by the average PMSE over the methods. This is done to overcome differences in the scale of the PMSEs among simulation models. Except for a couple of bars of almost identical lengths, the differences in length for all the other bars are statistically significant at the 0.1-level according to Tukey HSD simultaneous confidence intervals.

It is clear from the lengths of the bars for the IER and AIC methods under the *Null* model that they tend to overfit the data. Thus they are more likely than the other methods to identify an effect as significant when it is not. As may be expected, the EER method performs best at controlling the probability of false positives. But it has the highest PMSE values under the non-null situations. In contrast, the three

GUIDE methods provide a good compromise; they have relatively low PMSE values across all four simulation models.

4. Unreplicated 2^5 experiments

If an experiment is unreplicated, we cannot get an unbiased estimate of σ^2. Consequently, the IER and ERR approaches to model selection cannot be applied. The AIC method is useless too because it always selects the full model. For two-level factorial experiments, practitioners often use a rather subjective technique, due to Daniel [11], that is based on a half-normal quantile plot of the absolute estimated main and interaction effects. If the true effects are all null, the plotted points would lie approximately on a straight line. Daniel's method calls for fitting a line to a subset of points that appear linear near the origin and labeling as outliers those that fall far from the line. The selected model is the one that contains only the effects associated with the outliers.

For example, consider the data from a 2^5 reactor experiment given in Box, Hunter, and Hunter [1, p. 260]. There are 32 observations on five variables and Figure 5 shows a half-normal plot of the estimated effects. The authors judge that there are only five significant effects, namely, B, D, E, BD, and DE, yielding the model

$$\hat{y} = 65.5 + 9.75x_2 + 5.375x_4 - 3.125x_5 + 6.625x_2x_4 - 5.5x_4x_5. \tag{4.1}$$

Because Daniel did not specify how to draw the straight line and what constitutes an outlier, his method is difficult to apply objectively and hence cannot be evaluated by simulation. Formal algorithmic methods were proposed by Lenth [16], Loh [17], and Dong [12]. Lenth's method is the simplest. Based on the tables in Wu and Hamada [22, p. 620], the 0.05 IER version of Lenth's method gives the same model

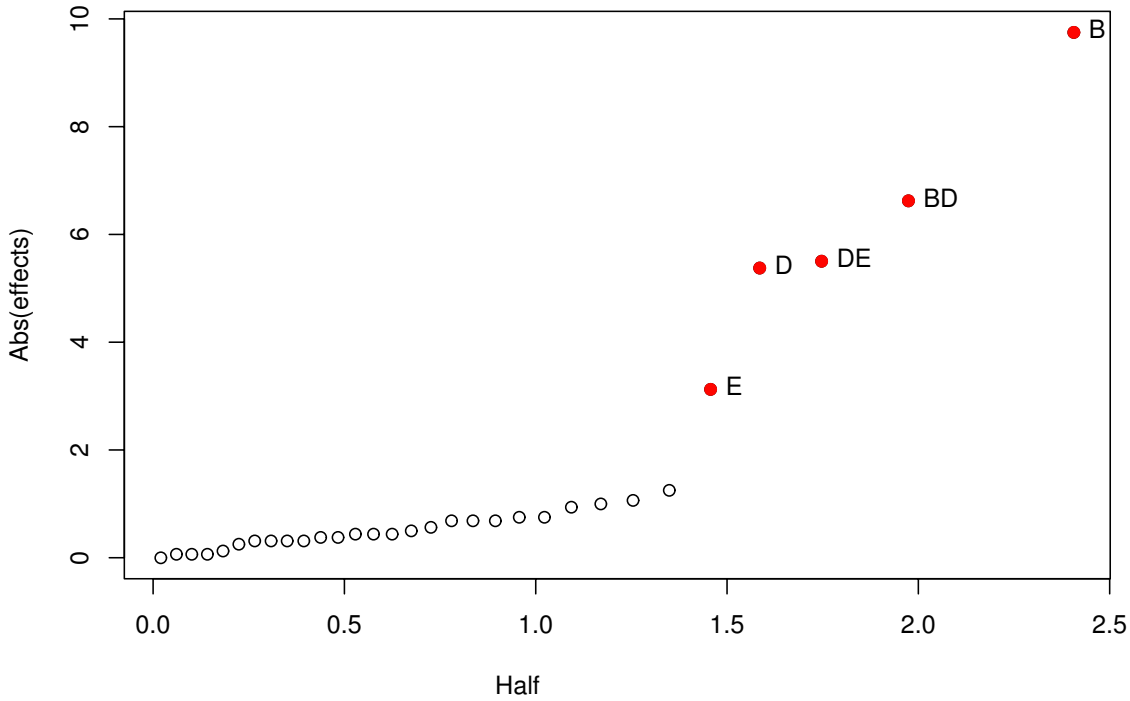


FIG 5. *Half-normal quantile plot of estimated effects from* 2^5 *reactor experiment.*

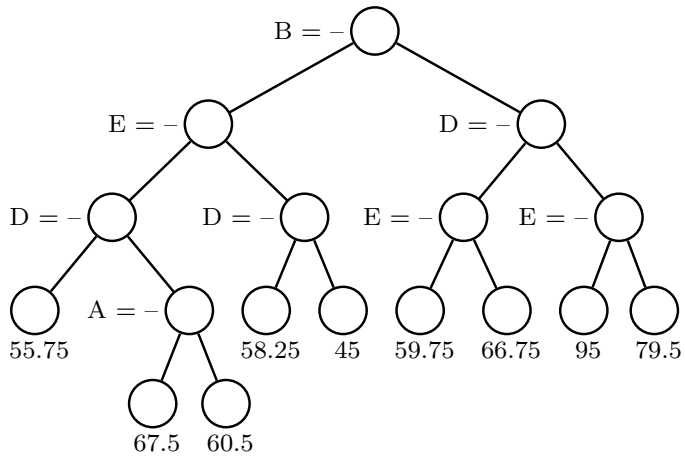


FIG 6. *Piecewise constant GUIDE model for the* 2^5 *reactor experiment. The sample y-mean is given beneath each leaf node.*

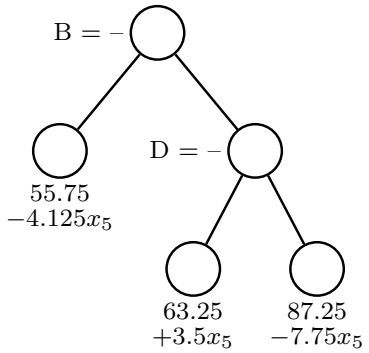


FIG 7. *Piecewise simple linear GUIDE model for the* 2^5 *reactor experiment. The fitted equation is given beneath each leaf node.*

as (4.1). The 0.1 EER version drops the E main effect, giving

$$\hat{y} = 65.5 + 9.75x_2 + 5.375x_4 + 6.625x_2x_4 - 5.5x_4x_5. \tag{4.2}$$

The piecewise constant GUIDE model for this dataset is shown in Figure 6. Besides variables B, D, and E, it finds that variable A also has some influence on the yield, albeit in a small region of the design space. The maximum predicted yield of 95 is attained when $B = D = +$ and $E = -$, and the minimum predicted yield of 45 when $B = -$ and $D = E = +$.

If at each node, instead of fitting a constant we fit a best simple linear regression model, we obtain the tree in Figure 7. Factor E, which was used to split the nodes at the second and third levels of the piecewise constant tree, is now selected as the best linear predictor in all three leaf nodes. We can try to further simplify the tree structure by fitting a multiple linear regression in each node. The result, shown on the left side of Figure 8, is a tree with only one split, on factor D. This model was also found by Cheng and Li [8], who use a method called principal Hessian directions to search for linear functions of the regressor variables; see Filliben and Li [13] for another example of this approach.

We can simplify the model even more by replacing multiple linear regression with stepwise regression at each node. The result is shown by the tree on the right side of Figure 8. It is almost the same as the tree on its left, except that only factors B

and E appear as regressors in the leaf nodes. This coincides with the Box, Hunter, and Hunter model (4.1), as seen by expressing the tree model algebraically as

$$\begin{aligned}\hat{y} &= (60.125 + 3.125x_2 + 2.375x_5)(1 - x_4)/2 \\ &\quad + (70.875 + 16.375x_2 - 8.625x_5)(1 + x_4)/2 \\ &= 65.5 + 9.75x_2 + 5.375x_4 - 3.125x_5 + 6.625x_2x_4 - 5.5x_4x_5.\end{aligned} \tag{4.3}$$

An argument can be made that the tree model on the right side of Figure 8 provides a more intuitive explanation of the BD and DE interactions than equation (4.4). For example, the coefficient for the x_2x_4 term (i.e., BD interaction) in (4.4) is $6.625 = (16.375 - 3.125)/2$, which is half the difference between the coefficients of the x_2 terms (i.e., B main effects) in the two leaf nodes of the tree. Since the root node is split on D, this matches the standard definition of the BD interaction as half the difference between the main effects of B conditional on the levels of D.

How do the five models compare? Their fitted values are very similar, as Figure 9 shows. Note that every GUIDE model satisfies the heredity principle, because by

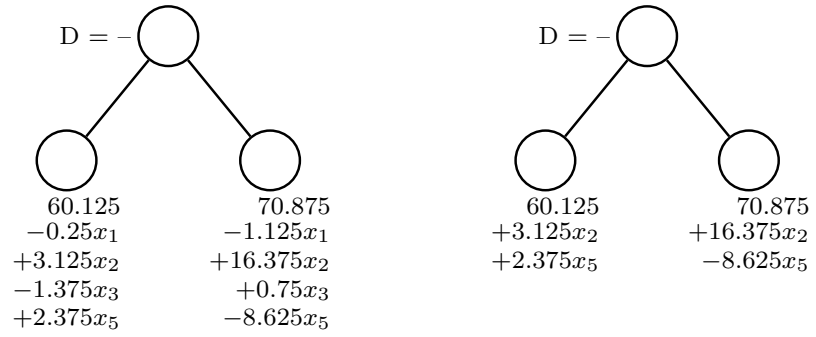


FIG 8. *GUIDE piecewise multiple linear (left) and stepwise linear (right) models.*

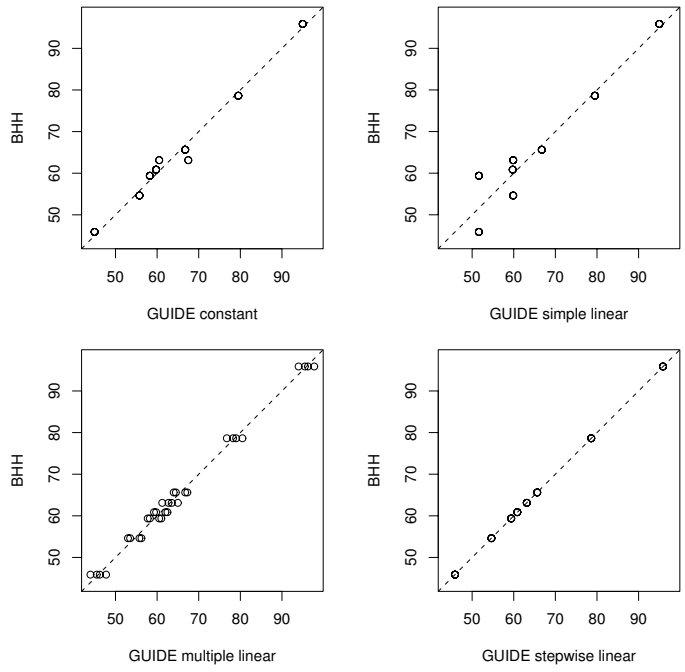


FIG 9. *Plots of fitted values from the Box, Hunter, and Hunter (BHH) model versus fitted values from four GUIDE models for the unreplicated* 2^5 *example.*

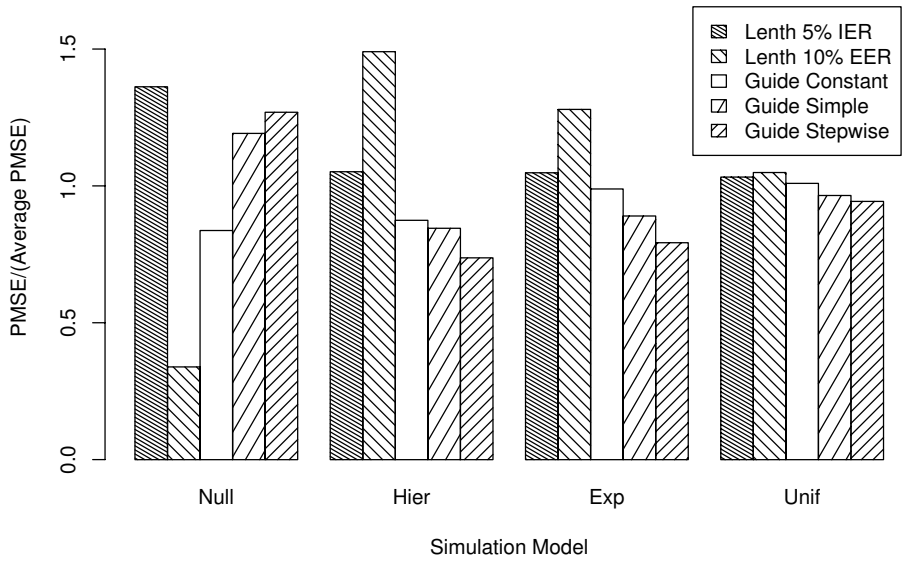


FIG 10. *Barplots of relative PMSEs of Lenth and GUIDE methods for four simulation models. The relative PMSE of a method at a simulation model is defined as its PMSE divided by the average PMSE of the five methods at the same model.*

construction an nth-order interaction effect appears only if the tree has $(n + 1)$ levels of splits. Thus if a model contains a cross-product term, it must also contain cross-products of all subsets of those variables.

Figure 10 shows barplots of the simulated relative PMSEs of the five methods for the four simulation models in Table 2. The methods being compared are: (i) Lenth using 0.05 IER, (ii) Lenth using 0.1 EER, (iii) piecewise constant GUIDE, (iv) piecewise best simple linear GUIDE, and (v) piecewise stepwise linear GUIDE. The results are based on 10,000 simulation trials with each trial consisting of 16 observations from an unreplicated 2^4 factorial. The behavior of the GUIDE models is quite similar to that for replicated experiments in Section 3. Lenth's EER method does an excellent job in controlling the probability of Type I error, but it does so at the cost of under-fitting the non-null models. On the hand, Lenth's IER method tends to over-fit more than any of the GUIDE methods, across all four simulation models.

5. Poisson regression

Model interpretation is much harder if some variables have more than two levels. This is due to the main and interaction effects having more than one degree of freedom. We can try to interpret a main effect by decomposing it into orthogonal contrasts to represent linear, quadratic, cubic, etc., effects, and similarly decompose an interaction effect into products of these contrasts. But because the number of products increases quickly with the order of the interaction, it is not easy to interpret several of them simultaneously. Further, if the experiment is unreplicated, model selection is more difficult because significance test-based and AIC-based methods are inapplicable without some assumptions on the order of the correct model.

To appreciate the difficulties, consider an unreplicated $3\times2\times4\times10\times3$ experiment on wave-soldering of electronic components in a printed circuit board reported in Comizzoli, Landwehr, and Sinclair [10]. There are 720 observations and the variables and their levels are:

TABLE 3
Results from a second-order Poisson loglinear model fitted to solder data

Term	Df	Sum of Sq	Mean Sq	F	Pr(>F)
Opening	2	1587.563	793.7813	568.65	0.00000
Solder	1	515.763	515.7627	369.48	0.00000
Mask	3	1250.526	416.8420	298.62	0.00000
Pad	9	454.624	50.5138	36.19	0.00000
Panel	2	62.918	31.4589	22.54	0.00000
Opening:Solder	2	22.325	11.1625	8.00	0.00037
Opening:Mask	6	66.230	11.0383	7.91	0.00000
Opening:Pad	18	45.769	2.5427	1.82	0.01997
Opening:Panel	4	10.592	2.6479	1.90	0.10940
Solder:Mask	3	50.573	16.8578	12.08	0.00000
Solder:Pad	9	43.646	4.8495	3.47	0.00034
Solder:Panel	2	5.945	2.9726	2.13	0.11978
Mask:Pad	27	59.638	2.2088	1.58	0.03196
Mask:Panel	6	20.758	3.4596	2.48	0.02238
Pad:Panel	18	13.615	0.7564	0.54	0.93814
Residuals	607	847.313	1.3959		

1. `Opening:` amount of clearance around a mounting pad (levels 'small', 'medium', or 'large')
2. `Solder:` amount of solder (levels 'thin' and 'thick')
3. `Mask:` type and thickness of the material for the solder mask (levels A1.5, A3, B3, and B6)
4. `Pad:` geometry and size of the mounting pad (levels D4, D6, D7, L4, L6, L7, L8, L9, W4, and W9)
5. `Panel:` panel position on a board (levels 1, 2, and 3)

The response is the number of solder skips, which ranges from 0 to 48.

Since the response variable takes non-negative integer values, it is natural to fit the data with a Poisson log-linear model. But how do we choose the terms in the model? A straightforward approach would start with an ANOVA-type model containing all main effect and interaction terms and then employ significance tests to find out which terms to exclude. We cannot do this here because fitting a full model to the data leaves no residual degrees of freedom for significance testing. Therefore we have to begin with a smaller model and hope that it contains all the necessary terms.

If we fit a second-order model, we obtain the results in Table 3. The three most significant two-factor interactions are between `Opening`, `Solder`, and `Mask`. These variables also have the most significant main effects. Chambers and Hastie [4, p. 10]—see also Hastie and Pregibon [14, p. 217]—determine that a satisfactory model for these data is one containing all main effect terms and these three two-factor interactions. Using set-to-zero constraints (with the first level in alphabetical order set to 0), this model yields the parameter estimates given in Table 4. The model is quite complicated and is not easy to interpret as it has many interaction terms. In particular, it is hard to explain how the interactions affect the mean response.

Figure 11 shows a piecewise constant Poisson regression GUIDE model. Its size is a reflection of the large number of variable interactions in the data. More interesting, however, is the fact that the tree splits first on `Opening`, `Mask`, and `Solder`—the three variables having the most significant two-factor interactions.

As we saw in the previous section, we can simplify the tree structure by fitting

TABLE 4
A Poisson loglinear model containing all main effects and all two-factor interactions involving `Opening`, `Solder`, *and* `Mask`.

Regressor	Coef	t	Regressor	Coef	t
Constant	-2.668	-9.25			
maskA3	0.396	1.21	openmedium	0.921	2.95
maskB3	2.101	7.54	opensmall	2.919	11.63
maskB6	3.010	11.36	soldthin	2.495	11.44
padD6	-0.369	-5.17	maskA3:openmedium	0.816	2.44
padD7	-0.098	-1.49	maskB3:openmedium	-0.447	-1.44
padL4	0.262	4.32	maskB6:openmedium	-0.032	-0.11
padL6	-0.668	-8.53	maskA3:opensmall	-0.087	-0.32
padL7	-0.490	-6.62	maskB3:opensmall	-0.266	-1.12
padL8	-0.271	-3.91	maskB6:opensmall	-0.610	-2.74
padL9	-0.636	-8.20	maskA3:soldthin	-0.034	-0.16
padW4	-0.110	-1.66	maskB3:soldthin	-0.805	-4.42
padW9	-1.438	-13.80	maskB6:soldthin	-0.850	-4.85
panel2	0.334	7.93	openmedium:soldthin	-0.833	-4.80
panel3	0.254	5.95	opensmall:soldthin	-0.762	-5.13

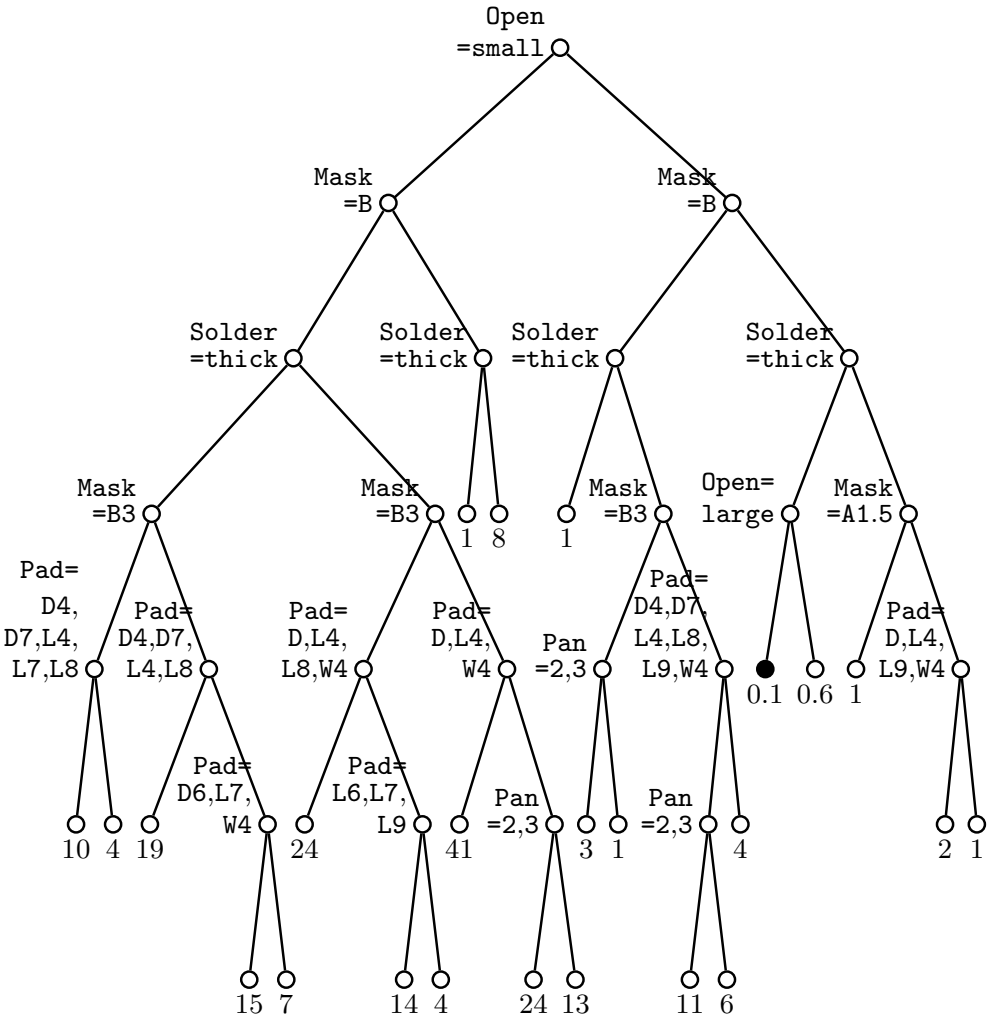


FIG 11. *GUIDE piecewise constant Poisson regression tree for solder data. "Panel" is abbreviated as "Pan". The sample mean yield is given beneath each leaf node. The leaf node with the lowest mean yield is painted black.*

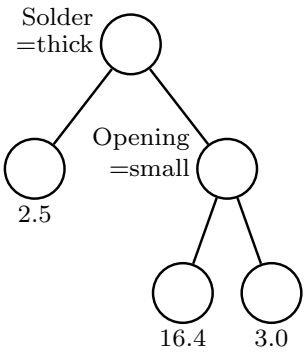


FIG 12. *GUIDE piecewise main effect Poisson regression tree for solder data. The number beneath each leaf node is the sample mean response.*

TABLE 5
Regression coefficients in leaf nodes of Figure 12

	Solder thick		Solder thin			
			Opening small		Opening not small	
Regressor	Coef	t	Coef	t	Coef	t
Constant	-2.43	-10.68	2.08	21.50	-0.37	-1.95
mask=A3	0.47	2.37	0.31	3.33	0.81	4.55
mask=B3	1.83	11.01	1.05	12.84	1.01	5.85
mask=B6	2.52	15.71	1.50	19.34	2.27	14.64
open=medium	0.86	5.57	aliased	-	0.10	1.38
open=small	2.46	18.18	aliased	-	aliased	-
pad=D6	-0.32	-2.03	-0.25	-2.79	-0.80	-4.65
pad=D7	0.12	0.85	-0.15	-1.67	-0.19	-1.35
pad=L4	0.70	5.53	0.08	1.00	0.21	1.60
pad=L6	-0.40	-2.46	-0.72	-6.85	-0.82	-4.74
pad=L7	0.04	0.29	-0.65	-6.32	-0.76	-4.48
pad=L8	0.15	1.05	-0.43	-4.45	-0.36	-2.41
pad=L9	-0.59	-3.43	-0.64	-6.26	-0.67	-4.05
pad=W4	-0.05	-0.37	-0.09	-1.00	-0.23	-1.57
pad=W9	-1.32	-5.89	-1.38	-10.28	-1.75	-7.03
panel=2	0.22	2.72	0.31	5.47	0.58	5.73
panel=3	0.07	0.81	0.19	3.21	0.69	6.93

a main effects model to each node instead of a constant. This yields the much smaller piecewise main effect GUIDE tree in Figure 12. It has only two splits, first on `Solder` and then, if the latter is thin, on `Opening`. Table 5 gives the regression coefficients in the leaf nodes and Figure 13 graphs them for each level of `Mask` and `Pad` by leaf node.

Because the regression coefficients in Table 5 pertain to conditional main effects only, they are simple to interpret. In particular, all the coefficients except for the constants and the coefficients for `Pad` have positive values. Since negative coefficients are desirable for minimizing the response, the best levels for all variables except `Pad` are thus those not in the table (i.e, whose levels are set to zero). Further, `W9` has the largest negative coefficient among `Pad` levels in every leaf node. Hence, irrespective of `Solder`, the best levels to minimize mean yield are A1.5 `Mask`, large `Opening`, W9 `Pad`, and `Panel` position 1. Finally, since the largest negative constant term occurs when `Solder` is thick, the latter is the best choice for minimizing mean yield. Conversely, it is similarly observed that the worst combination (i.e., one giving the highest predicted mean number of solder skips) is thin `Solder`, small `Opening`, B6 `Mask`, L4 `Pad`, and `Panel` position 2.

Given that the tree has only two levels of splits, it is safe to conclude that

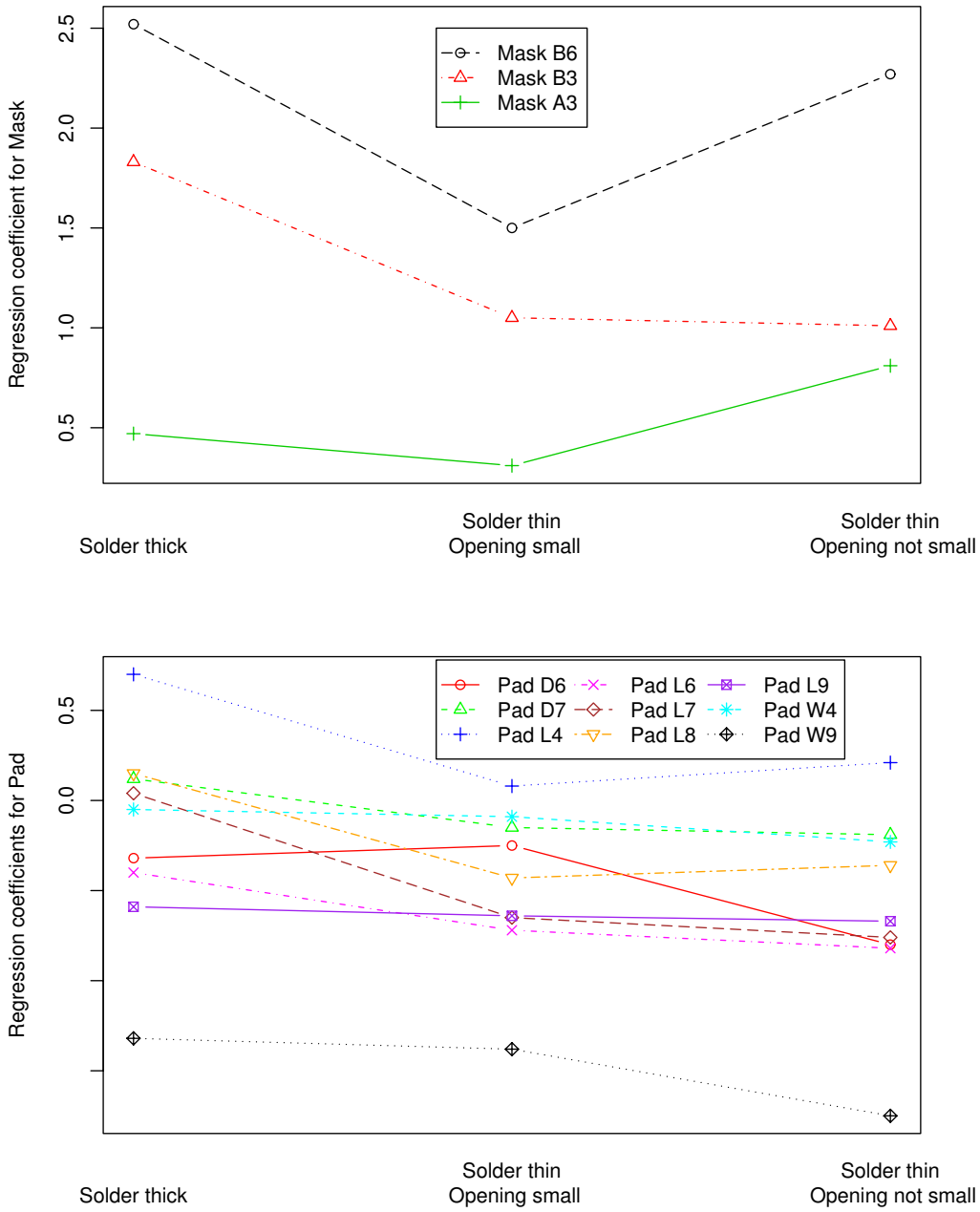


FIG 13. *Plots of regression coefficients for* `Mask` *and* `Pad` *from Table* 5.

four-factor and higher interactions are negligible. On the other hand, the graphs in Figure 13 suggest that there may exist some weak three-factor interactions, such as between `Solder`, `Opening`, and `Pad`. Figure 14, which compares the fits of this model with those of the Chambers-Hastie model, shows that the former fits slightly better.

6. Logistic regression

The same ideas can be applied to fit logistic regression models when the response variable is a sample proportion. For example, Table 6 shows data reported in Collett [9, p. 127] on the number of seeds germinating, out of 100, at two germination temperatures. The seeds had been stored at three moisture levels and three storage temperatures. Thus the experiment is a $2 \times 3 \times 3$ design.

Treating all the factors as nominal, Collett [9, p. 128] finds that a linear logistic

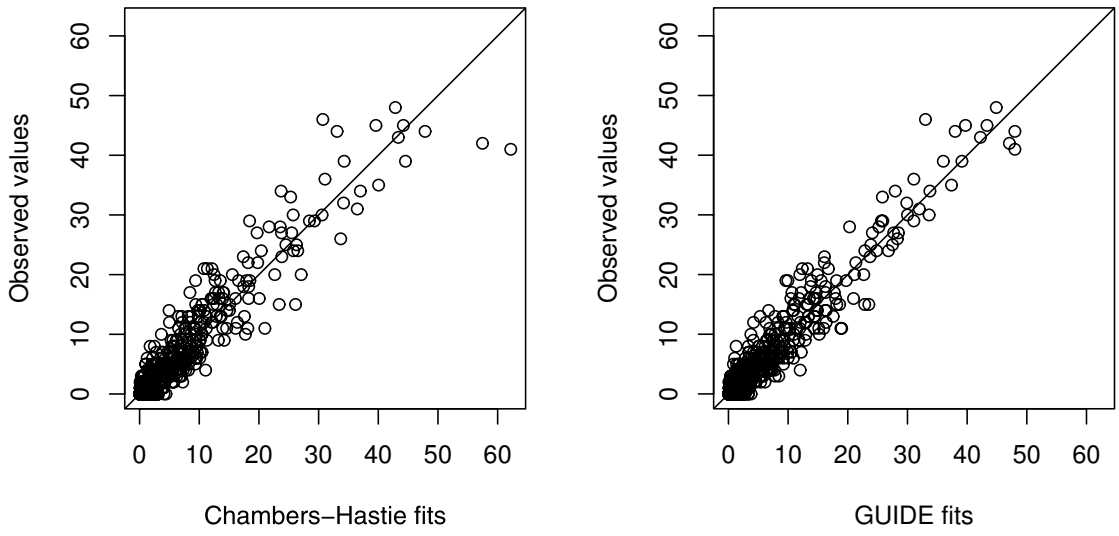


FIG 14. *Plots of observed versus fitted values for the Chambers–Hastie model in Table* 4 *(left) and the GUIDE piecewise main effects model in Table* 5 *(right).*

TABLE 6
Number of seeds, out of 100, that germinate

Germination temp. (°C)	Moisture level	Storage temp. (°C) 21	42	62
11	low	98	96	62
11	medium	94	79	3
11	high	92	41	1
21	low	94	93	65
21	medium	94	71	2
21	high	91	30	1

TABLE 7
Logistic regression fit to seed germination data using set-to-zero constraints

	Coef	SE	z	$\Pr(>\|z\|)$
(Intercept)	2.5224	0.2670	9.447	< 2e-16
germ21	-0.2765	0.1492	-1.853	0.06385
store42	-2.9841	0.2940	-10.149	< 2e-16
store62	-6.9886	0.7549	-9.258	< 2e-16
moistlow	0.8026	0.4412	1.819	0.06890
moistmed	0.3757	0.3913	0.960	0.33696
store42:moistlow	2.6496	0.5595	4.736	2.18e-06
store62:moistlow	4.3581	0.8495	5.130	2.89e-07
store42:moistmed	1.3276	0.4493	2.955	0.00313
store62:moistmed	0.5561	0.9292	0.598	0.54954

regression model with all three main effects and the interaction between moisture level and storage temperature fits the sample proportions reasonably well. The parameter estimates in Table 7 show that only the main effect of storage temperature and its interaction with moisture level are significant at the 0.05 level. Since the storage temperature main effect has two terms and the interaction has four, it takes some effort to fully understand the model.

A simple linear logistic regression model, on the other hand, is completely and intuitively explained by its graph. Therefore we will fit a piecewise simple linear logistic model to the data, treating the three-valued storage temperature variable as a continuous linear predictor. We accomplish this with the LOTUS [5] algorithm,

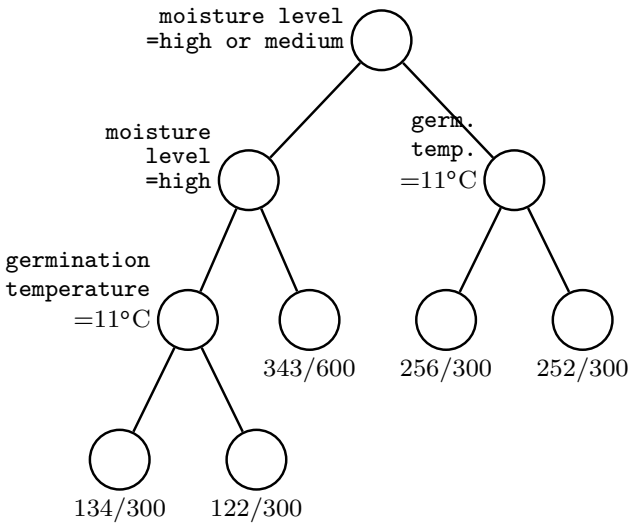


FIG 15. *Piecewise simple linear LOTUS logistic regression tree for seed germination experiment. The fraction beneath each leaf node is the sample proportion of germinated seeds.*

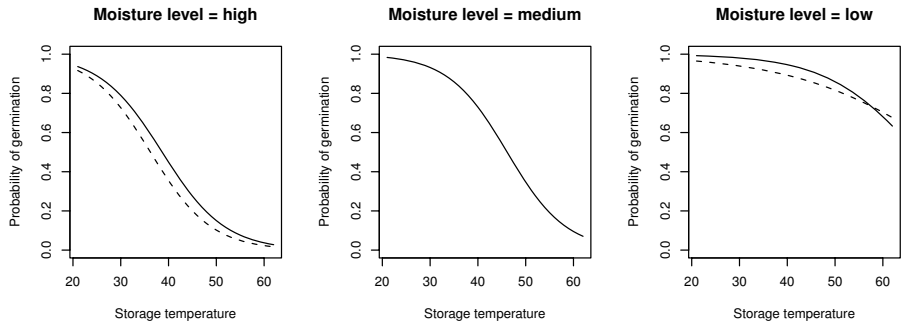


FIG 16. *Fitted probability functions for seed germination data. The solid and dashed lines pertain to fits at germination temperatures of* 11 *and* 21 *degrees, respectively. The two lines coincide in the middle graph.*

which extends the GUIDE algorithm to logistic regression. It yields the logistic regression tree in Figure 15. Since there is only one linear predictor in each node of the tree, the LOTUS model can be visualized through the fitted probability functions shown in Figure 16. Note that although the tree has five leaf nodes, and hence five fitted probability functions, we can display the five functions in three graphs, using solid and dashed lines to differentiate between the two germination temperature levels. Note also that the solid and dashed lines coincide in the middle graph because the fitted probabilities there are independent of germination temperature.

The graphs show clearly the large negative effect of storage temperature, especially when moisture level is medium or high. Further, the shapes of the fitted functions for low moisture level are quite different from those for medium and high moisture levels. This explains the strong interaction between storage temperature and moisture level found by Collett [9].

7. Conclusion

We have shown by means of examples that a regression tree model can be a useful supplement to a traditional analysis. At a minimum, the former can serve as a check on the latter. If the results agree, the tree offers another way to interpret the main

effects and interactions beyond their representations as single degree of freedom contrasts. This is especially important when variables have more than two levels because their interactions cannot be fully represented by low-order contrasts. On the other hand, if the results disagree, the experimenter may be advised to reconsider the assumptions of the traditional analysis. Following are some problems for future study.

1. A tree structure is good for uncovering interactions. If interactions exist, we can expect the tree to have multiple levels of splits. What if there are no interactions? In order for a tree structure to represent main effects, it needs one level of splits for each variable. Hence the complexity of a tree is a sufficient but not necessary condition for the presence of interactions. One way to distinguish between the two situations is to examine the algebraic equation associated with the tree. If there are no interaction effects, the coefficients of the cross-product terms can be expected to be small relative to the main effect terms. A way to formalize this idea would be useful.
2. Instead of using empirical principles to exclude all high-order effects from the start, a tree model can tell us which effects might be important and which unimportant. Here "importance" is in terms of prediction error, which is a more meaningful criterion than statistical significance in many applications. High-order effects that are found this way can be included in a traditional stepwise regression analysis.
3. How well do the tree models estimate the true response surface? The only way to find out is through computer simulation where the true response function is known. We have given some simulation results to demonstrate that the tree models can be competitive in terms of prediction mean squared error, but more results are needed.
4. Data analysis techniques for designed experiments have traditionally focused on normally distributed response variables. If the data are not normally distributed, many methods are either inapplicable or become poor approximations. Wu and Hamada [22, Chap. 13] suggest using generalized linear models for count and ordinal data. The same ideas can be extended to tree models. GUIDE can fit piecewise normal or Poisson regression models and LOTUS can fit piecewise simple or multiple linear logistic models. But what if the response variable takes unordered nominal values? There is very little statistics literature on this topic. Classification tree methods such as CRUISE [15] and QUEST [19] may provide solutions here.
5. Being applicable to balanced as well as unbalanced designs, tree methods can be useful in experiments where it is impossible or impractical to obtain observations from particular combinations of variable levels. For the same reason, they are also useful in response surface experiments where observations are taken sequentially at locations prescribed by the shape of the surface fitted up to that time. Since a tree algorithm fits the data piecewise and hence locally, all the observations can be used for model fitting even if the experimenter is most interested in modeling the surface in a particular region of the design space.

References

[1] Box, G. E. P., Hunter, W. G. and Hunter, J. S. (2005). *Statistics for Experimenters*, 2nd ed. Wiley, New York.

[2] Box, G. E. P. and Meyer, R. D. (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology* **25**, 94–105.
[3] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.
[4] Chambers, J. M. and Hastie, T. J. (1992). An appetizer. In *Statistical Models in S*, J. M. Chambers and T. J. Hastie, eds. Wadsworth & Brooks/Cole. Pacific Grove, 1–12.
[5] Chan, K.-Y. and Loh, W.-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics* **13**, 826–852.
[6] Chaudhuri, P., Lo, W.-D., Loh, W.-Y. and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica* **5**, 641–666.
[7] Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli* **8**, 561–576.
[8] Cheng, C.-S. and Li, K.-C. (1995). A study of the method of principal Hessian direction for analysis of data from designed experiments. *Statistica Sinica* **5**, 617–640.
[9] Collett, D. (1991). *Modelling Binary Data*. Chapman and Hall, London.
[10] Comizzoli, R. B., Landwehr, J. M. and Sinclair, J. D. (1990). Robust materials and processes: Key to reliability. *AT&T Technical Journal* **69**, 113–128.
[11] Daniel, C. (1971). *Applications of Statistics to Industrial Experimentation*. Wiley, New York.
[12] Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica* **3**, 209–217.
[13] Filliben, J. J. and Li, K.-C. (1997). A systematic approach to the analysis of complex interaction patterns in two-level factorial designs. *Technometrics* **39**, 286–297.
[14] Hastie, T. J. and Pregibon, D. (1992). Generalized linear models. In *Statistical Models in S*, J. M. Chambers and T. J. Hastie, eds. Wadsworth & Brooks/Cole. Pacific Grove, 1–12.
[15] Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association* **96**, 589–604.
[16] Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31**, 469–473.
[17] Loh, W.-Y. (1992). Identification of active contrasts in unreplicated factorial experiments. *Computational Statistics and Data Analysis* **14**, 135–148.
[18] Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* **12**, 361–386.
[19] Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica* **7**, 815–840.
[20] Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58**, 415–434.
[21] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.
[22] Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 229–240
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000473

On competing risk and degradation processes

Nozer D. Singpurwalla[1,*]

The George Washington University

Abstract: Lehmann's ideas on concepts of dependence have had a profound effect on mathematical theory of reliability. The aim of this paper is two-fold. The first is to show how the notion of a "hazard potential" can provide an explanation for the cause of dependence between life-times. The second is to propose a general framework under which two currently discussed issues in reliability and in survival analysis involving interdependent stochastic processes, can be meaningfully addressed via the notion of a *hazard potential*. The first issue pertains to the failure of an item in a dynamic setting under multiple interdependent risks. The second pertains to assessing an item's life length in the presence of observable surrogates or markers. Here again the setting is dynamic and the role of the marker is akin to that of a leading indicator in multiple time series.

1. Preamble: Impact of Lehmann's work on reliability

Erich Lehmann's work on non-parametrics has had a conceptual impact on reliability and life-testing. Here two commonly encountered themes, one of which bears his name, encapsulate the essence of the impact. These are: the notion of a ***Lehmann Alternative***, and his exposition on ***Concepts of Dependence***. The former (see Lehmann [4]) comes into play in the context of accelerated life testing, wherein a Lehmann alternative is essentially a model for accelerating failure. The latter (see Lehmann [5]) has spawned a large body of literature pertaining to the reliability of complex systems with interdependent component lifetimes. Lehmann's original ideas on characterizing the nature of dependence has helped us better articulate the effect of failures that are *causal* or *cascading*, and the consequences of lifetimes that exhibit a negative correlation. The aim of this paper is to propose a framework that has been inspired by (though not directly related to) Lehmann's work on dependence. The point of view that we adopt here is "dynamic", in the sense that what is of relevance are dependent stochastic processes. We focus on two scenarios, one pertaining to competing risks, a topic of interest in survival analysis, and the other pertaining to degradation and its markers, a topic of interest to those working in reliability. To set the stage for our development we start with an overview of the notion of a *hazard potential*, an entity which helps us better conceptualize the process of failure and the cause of interdependent lifetimes.

*Research supported by Grant DAAD 19-02-01-0195, The U. S. Army Research Office.
[1]Department of Statistics, The George Washington University, Washington, DC 20052, USA, e-mail: nozer@gwu.edu
AMS 2000 subject classifications: primary 62N05, 62M05; secondary 60J65.
Keywords and phrases: biomarkers, dynamic reliability, hazard potential, interdependence, survival analysis, inference for stochastic processes, Wiener maximum processes.

2. Introduction: The hazard potential

Let T denote the time to failure of a unit that is scheduled to operate in some specified static environment. Let $h(t)$ be the hazard rate function of the survival function of T, namely, $P(T \geq t), t \geq 0$. Let $H(t) = \int_0^t h(u)du$, be the cumulative hazard function at $t; H(t)$ is increasing in t. With $h(t), t \geq 0$ specified, it is well known that

$$\Pr(T \geq t; h(t), t \geq 0) = \exp(-H(t)).$$

Consider now an exponentially distributed random variable X, with scale parameter λ, $\lambda \geq 0$. Then for some $H(t) \geq 0$,

$$\Pr(X \geq H(t)|\lambda = 1) = \exp(-H(t));$$

thus

$$\Pr(T \geq t; h(t), t \geq 0) = \exp(-H(t)) = \Pr(X \geq H(t)|\lambda = 1). \tag{2.1}$$

The right hand side of the above equation says that the item in question will fail when its cumulative hazard $H(t)$ crosses a threshold X, where X has a unit exponential distribution. Singpurwalla [11] calls X the ***Hazard Potential*** of the item, and interprets it as an unknown resource that the item is endowed with at inception. Furthermore, $H(t)$ is interpreted as the amount of resource consumed at time t, and $h(t)$ is the rate at which that resource gets consumed. Looking at the failure process in terms of an endowed and a consumed resource enables us to characterize an environment as being *normal* when $H(t) = t$, and as being *accelerated* (*decelerated*) when $H(t) \geq (\leq)\ t$. More importantly, with X interpreted as an unknown resource, we are able to interpret dependent lifetimes as the consequence of dependent hazard potentials, the later being a manifestation of commonalities of design, manufacture, or genetic make-up. Thus one way to generate dependent lifetimes, say T_1 and T_2 is to start with a bivariate distribution (X_1, X_2) whose marginal distributions are exponential with scale parameter one, and which is not the product of exponential marginals. The details are in Singpurwalla [11].

When the environment is dynamic, the rate at which an item's resource gets consumed is random. Thus $h(t); t \geq 0$ is better described as a stochastic process, and consequently, so is $H(t), t \geq 0$. Since $H(t)$ is increasing in t, the ***cumulative hazard process*** $\{H(t); t \geq 0\}$ is a continuous increasing process, and the item fails when this process hits a random threshold X, the item's hazard potential. Candidate stochastic processes for $\{H(t); t \geq 0\}$ are proposed in the reference given above, and the nature of the resulting lifetimes described therein. Noteworthy are an increasing Lévy process, and the maxima of a Wiener process.

In what follows we show how the notion of a hazard potential serves as a unifying platform for describing the competing risk phenomenon and the phenomenon of failure due to ageing or degradation in the presence of a marker (or a bio marker) such as crack size (or a CD4 cell count).

3. Dependent competing risks and competing risk processes

By "competing risks" one generally means failure due to agents that presumably compete with each other for an item's lifetime. The traditional model that has been used for describing the competing risk phenomenon has been the reliability of a series system whose component lifetimes are independent or dependent. The idea

here is that since the failure of any component of the system leads to the failure of the system, the system experiences multiple risks, each risk leading to failure. Thus if T_i denotes the lifetime of component $i,\ i = 1, \ldots, k$, say, then the cause of system failure is that component whose lifetime is smallest of the k lifetimes. Consequently, if T denotes a system's lifetime, then

$$\Pr(T \geq t) = P(H_1(t) \leq X_1, \ldots, H_k(t) \leq X_k), \tag{3.1}$$

where X_i is the hazard potential of the i-th component, and $H_i(t)$ its cumulative hazard (or the risk to component i) at time t. If the X_i's are assumed to be independent (a simplifying assumption), then (3.1) leads to the result that

$$\Pr(T \geq t) = \exp[-(H_1(t) + \cdots + H_k(t))], \tag{3.2}$$

suggesting an additivity of cumulative hazard functions, or equivalently, an additivity of the risks. Were the X_i's assumed dependent, then the nature of their dependence will dictate the manner in which the risks combine. Thus for example if for some $\theta,\ 0 \leq \theta \leq 1$, we suppose that

$$\Pr(X_1 \geq x_1, X_2 \geq x_2|\theta) = \exp(-x_1 - x_2 - \theta x_1 x_2),$$

namely one of Gumbel's bivariate exponential distributions, then

$$\Pr(T \geq t|\theta) = \exp[-(H_1(t) + H_2(t) + \theta H_1(t) H_2(t))].$$

The cumulative hazards (or equivalently, the risks) are no longer additive.

The series system model discussed above has also been used to describe the failure of a single item that experiences several failure causing agents that compete with each other. However, we question this line of reasoning because a single item posseses only one unknown resource. Thus the $X_1, \ldots, X_k$ of the series system model should be replaced by a single X, where $X_1 = X_2 = \cdots = X_k = X$ (in probability). To set the stage for the single item case, suppose that the item experiences k agents, say $C_1, \ldots, C_k$, where an agent is seen as a cause of failure; for example, the consumption of fatty foods. Let $H_i(t)$ be the consequence of agent C_i, were C_i be the only agent acting on the item. Then under the simultaneous action by all of the k agents the item's survival function

$$\begin{aligned} &\Pr(T \geq t; h_1(t), \ldots, h_k(t)) \\ &\quad = P(H_1(t) \leq X, \ldots, H_k(t) \leq X) \\ &\quad = \exp(-\max(H_1(t), \ldots, H_k(t))). \end{aligned} \tag{3.3}$$

Here again, the cumulative hazards are not additive.

Taking a clue from the fact that dependent hazard potentials lead us to a non-additivity of the cumulative hazard functions, we observe that the condition $X_1 \overset{P}{=} X_2 \overset{P}{=} \cdots \overset{P}{=} X_k \overset{P}{=} X$ (where $X_1 \overset{P}{=} X_2$ denotes that $X_1$and X_2 are equal in probability) implies that $X_1, \ldots, X_k$ are *totally positively dependent*, in the sense of Lehmann (1966). Thus (3.2) and (3.3) can be combined to claim that in general, under the series system model for competing risks, $P(T \geq t)$ can be bounded as

$$\exp(-\sum_{1}^{k} H_i(t)) \leq P(T \geq t) \leq \exp(-\max(H_1(t), \ldots, H_k(t))). \tag{3.4}$$

Whereas (3.4) above may be known, our argument leading up to it could be new.

3.1. Competing risk processes

The prevailing view of what constitutes dependent competing risks entails a consideration of dependent component lifetimes in the series system model mentioned above. By contrast, our position on a proper framework for describing dependent competing risks is different. Since it is the $H_i(t)$'s that encapsulate the notion of risk, dependent competing risks should entail interdependence between $H_i(t)$'s, $i = 1, \ldots, k$. This would require that the $H_i(t)$'s be random, and a way to do so is to assume that each $\{H_i(t); t \geq 0\}$ is a stochastic process; we call this a ***competing risk process***. The item fails when any one of the $\{H_i(t); t \geq 0\}$ processes first hits the items hazard potential X. To incorporate interdependence between the $H_i(t)$'s, we conceptualize a k-variate process $\{H_1(t), \ldots, H_k(t); t \geq 0\}$, that we call a ***dependent competing risk process***. Since $H_i(t)$'s are increasing in t, one possible choice for each $\{H_i(t); t \geq 0\}$ could be a *Brownian Maximum Process*. That is $H_i(t) = \sup_{0<s\leq t}\{W_i(s); s \geq 0\}$, where $\{W_i(s); s \geq 0\}$ is a standard Brownian motion process. Dependence between the $H_i(t)$'s can be induced via a dependence between the $\{W_i(s); s \geq 0\}$ processes. Thus for example, in the bivariate case, if ρ denotes the correlation between two standard Brownian motion processes, then

$$\Pr(T \geq t) = \int_0^\infty P\left(H_1(t) \leq x, H_2(t) \leq x\right) e^{-x} dx$$

and it can be shown (details omitted) that,

$$\Pr(T \geq t) = \frac{\int_0^t \int_0^t \exp\left[-\frac{\left(a^2+b^2-2\rho ab\right)}{2t(1-\rho^2)}\right] dadb}{\int_0^\infty \int_0^\infty \exp\left[-\frac{\left(u^2+v^2-2\rho uv\right)}{2t(1-\rho^2)}\right] dudv}. \tag{3.5}$$

Another possibility, again for the case of $k = 2$, is to assume that $\{H_1(t); t \geq 0\}$ is some non-negative, non-decreasing, right-continuous process, but that $\{H_2(t); t \geq 0\}$ has a sample path which is an impulse function of the form $H_2(t) = 0$ for all $t < t^*$, and that $H_2(t^*) = \infty$ for some $t^* > 0$, where the rate of occurrence of the impulse at time t depends on $H_1(t)$. The process $\{H_2(t); t \geq 0\}$ can be identified with some sort of a traumatic event that competes with the process $\{H_1(t); t \geq 0\}$ for the lifetime of the item. In the absence of trauma the item fails when the process $\{H_1(t); t \geq 0\}$ hits the item's hazard potential. This scenario parallels the one considered by Lemoine and Wenocur [6], albeit in a context that is different from ours. By assuming that the probability of occurrence of an impulse in the time interval $[t, t + h)$, given that $H_1(t) = \omega$, is $1 - \exp(-\omega h)$, Lemoine and Wenocur [6] have shown that for $X = x$, the probability of survival of an item to time t is of the form:

$$\Pr(T \geq t) = E\left[\exp\left(\int_0^t H_1(s)ds\right) I_{[0,x)}\left(H_1(t)\right)\right], \tag{3.6}$$

where $I_A(\bullet)$ is the indicator of a set A, and the expectation is with respect to the distribution of the process $\{H_1(t); t \geq 0\}$. As a special case, when $\{H_1(t); t \geq 0\}$ is a gamma process (see Singpurwalla [10]), and x is infinite, so that $I_{[0,\infty)}\left(H_1(t)\right) = 1$ for $H_1(t) \geq 0$, the above equation takes the form

$$\Pr(T \geq t) = \exp(-(1 + t)\log(1 + t) + t). \tag{3.7}$$

The closed form result of (3.7) suffers from the disadvantage of having the effect of the hazard potential de facto nullified. The more realistic case of (3.6) will call for numerical or simulation based approaches. These remain to be done; our aim here has been to give some flavor of the possibilities.

4. Biomarkers and degradation processes

A topic of current interest in both reliability and survival analysis pertains to assessing lifetimes based on observable surrogates, such as crack length, and biomarkers like CD4 cell counts. Here again the hazard potential provides a unified perspective for looking at the interplay between the unobservable failure causing phenomenon, and an observable surrogate. It is an assumed dependence between the above two processes that makes this interplay possible.

To engineers (cf. Bogdanoff and Kozin [1]) degradation is the irreversible accumulation of damage throughout life that leads to failure. The term "damage" is not defined; however it is claimed that damage manifests itself via surrogates such as cracks, corrosion, measured wear, etc. Similarly, in the biosciences, the notion of "ageing" pertains to a unit's position in a state space wherein the probabilities of failure are greater than in a former position. Ageing manifests itself in terms of biomedical and physical difficulties experienced by individuals and other such biomarkers.

With the above as background, our proposal here is to conceptualize ageing and degradation as unobservable constructs (or latent variables) that serve to describe a process that results in failure. These constructs can be seen as the cause of observable surrogates like cracks, corrosion, and biomarkers such as CD4 cell counts. This modelling viewpoint is not in keeping with the work on degradation modelling by Doksum [3] and the several references therein. The prevailing view is that degradation is an observable phenomenon that reveals itself in the guise of crack length and CD4 cell counts. The item fails when the observable phenomenon hits some threshold whose nature is not specified. Whereas this may be meaningful in some cases, a more general view is to separate the observable and the unobservable and to attribute failure as a consequence of the behavior of the unobservable.

To mathematically describe the cause and effect phenomenon of degradation (or ageing) and the observables that it spawns, we view the (unobservable) cumulative hazard function as degradation, or ageing, and the biomarker as an observable process that is influenced by the former. The item fails when the cumulative hazard function hits the item's hazard potential X, where X has exponential (1) distribution. With the above in mind we introduce the ***degradation process*** as a bivariate stochastic process $\{H(t), Z(t), t \geq 0\}$, with $H(t)$ representing the unobservable degradation, and $Z(t)$ an observable marker. Whereas $H(t)$ is required to be non-decreasing, there is no such requirement on $Z(t)$. For the marker to be useful as a predictor of failure, it is necessary that $H(t)$ and $Z(t)$ be related to each other. One way to achieve this linkage is via a *Markov Additive Process* (cf. Cinlar [2]) wherein $\{Z(t); t \geq 0\}$ is a Markov process and $\{H(t); t \geq 0\}$ is an increasing Lévy process whose parameters depend on the state of the $\{Z(t); t \geq 0\}$ process. The ramifications of this set-up need to be explored.

Another possibility, and one that we are able to develop here in some detail (see Section 5), is to describe $\{Z(t); t \geq 0\}$ by a Wiener process (cracks do heal and CD4 cell counts do fluctuate), and the unobservable degradation process $\{H(t); t \geq 0\}$

by a ***Wiener Maximum Process***, namely,

$$H(t) = \sup_{0<s\leq t} \{Z(s); s \geq 0\}. \tag{4.1}$$

What makes the topic of analyzing degradation processes attractive is not just the modeling part; the statistical and computational issues that the set-up creates are quite challenging. Since $\{Z(t); t \geq 0\}$ is an observable process, how may one use observations on this process until some time, say t^*, to make inferences about the process of interest $H(t)$, for $t > t^*$? In other words, how does one assess $\Pr(T > t|\{Z(s); 0 < s \leq t^* < t\})$, where T is an item's time to failure? Furthermore, as is often the case, the process $\{Z(s); s \geq 0\}$ cannot be monitored continuously. Rather, what one is able to do is observe $\{Z(s); s \geq 0\}$ at k discrete time points and use these as a basis for inference about $\Pr(T > t|\{Z(s); 0 < s \leq t^* < t\})$. These and other matters are discussed next in Section 5, which could be viewed as a prototype of what else is possible using other models for degradation.

5. Inference under a Wiener maximum process for degradation

We start with some preliminaries about a Wiener process and its hitting time to a threshold. The notation used here is adopted from Doksum [3].

5.1. Hitting time of a Wiener maximum process to a random threshold

Let Z_t denote an observable marker process $\{Z(t); t \geq 0\}$, and H_t an unobservable degradation process $\{H(t); t \geq 0\}$. The relationship between these two processes is prescribed by (4.1). Suppose that Z_t is described by a Wiener process with a drift parameter η and a diffusion parameter $\sigma^2 > 0$. That is, $Z(0) = 0$ and Z_t has independent increments. Also, for any $t > 0$, $Z(t)$ has a Gaussian distribution with $E(Z(t)) = \eta t$, and for any $0 \leq t_1 < t_2$, $\mathrm{Var}[Z(t_2) - Z(t_1)] = (t_2 - t_1)\sigma^2$. Let T_x denote the first time at which Z_t crosses a threshold $x > 0$; that is, T_x is the hitting time of Z_t to x. Then, when $\eta = 0$,

$$\begin{aligned} \Pr(Z(t) \geq x) &= \Pr(Z(t) \geq x|T_x \leq t) \Pr(T_x \leq t) \\ &\quad + \Pr(Z(t) \geq x|T_x > t) \Pr(T_x > t), \end{aligned} \tag{5.1}$$

$$\Pr(T_x \leq t) = 2 \Pr(Z(t) \geq x). \tag{5.2}$$

This is because $\Pr(Z(t) \geq x|T_x \leq t)$ can be set to $1/2$, and the second term on the right hand side of (5.1) is zero. When $Z(t)$ has a Gaussian distribution with mean ηt and variance $\sigma^2 t$, $\Pr(Z(t) \geq x)$ can be similarly obtained, and thence $\Pr(T_x \leq t) \stackrel{def}{=} F_x(t|\eta, \sigma)$. Specifically it can be seen that

$$F_x(t|\eta, \sigma) = \Phi\left(\frac{\sqrt{\lambda}}{\mu}\sqrt{t} - \frac{\sqrt{\lambda}}{\sqrt{t}}\right) + \Phi\left(-\frac{\sqrt{\lambda}}{\mu}\sqrt{t} - \frac{\sqrt{\lambda}}{\sqrt{t}}\right) \exp\left(\frac{2\lambda}{\mu}\right), \tag{5.3}$$

where $\mu = x/\eta$ and $\lambda = x^2/\sigma^2$. The distribution $F_x$is the ***Inverse Gaussian Distribution*** (IG-Distribution) with parameters μ and λ, where $\mu = E(T_x)$ and $\lambda\mu^2 = \mathrm{Var}(T_x)$. Observe that when $\eta = 0$, both $E(T_x)$ and $\mathrm{Var}(T_x)$ are infinite, and thus for any meaningful description of a marker process via a Wiener process, the drift parameter η needs to be greater than zero.

The probability density of F_x at t takes the form:

$$f_x(t|\eta,\sigma) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left[-\frac{\lambda}{2\mu^2}\frac{(t-\mu)^2}{t}\right], \tag{5.4}$$

for $t, \mu, \lambda > 0$.

We now turn attention to H_t, the process of interest. We first note that because of (4.1), $H(0) = 0$, and $H(t)$ is non-decreasing in t; this is what was required of H_t. An item experiencing the process H_t fails when H_t first crosses a threshold X, where X is unknown. However, our uncertainty about X is described by an exponential distribution with probability density $f(x) = e^{-x}$. Let T denote the time to failure of the item in question. Then, following the line of reasoning leading to (5.1), we would have, in the case of $\eta = 0$,

$$\Pr(T \le t) = 2\Pr(H(t) \ge x).$$

Furthermore, because of (4.1), the hitting time of H_t to a random threshold X will coincide with T_x, the hitting time of Z_t (with $\eta > 0$) to X. Consequently,

$$\begin{aligned}\Pr(T \le t) = \Pr(T_x \le t) &= \int_0^\infty \Pr(T_x \le t|X = x) f(x)dx \\ &= \int_0^\infty \Pr(T_x \le t)e^{-x}dx = \int_0^\infty F_x(t|\eta,\sigma)e^{-x}dx.\end{aligned}$$

Rewriting $F_x(t|\eta,\sigma)$ in terms of the marker process parameters η and σ, and treating these parameters as known, we have

$$\begin{aligned}\Pr(T \le t|\eta,\sigma) &\overset{def}{=} F(t|\eta,\sigma) \\ &= \int_0^\infty \left[\Phi\left(\frac{\eta}{\sigma}\sqrt{t} - \frac{x}{\sigma\sqrt{t}}\right) + \Phi\left(-\frac{\eta}{\sigma}\sqrt{t} - \frac{x}{\sigma\sqrt{t}}\right)\right] \\ &\quad \times \exp\left(x\left(\frac{2\eta}{\sigma^2} - 1\right)\right)dx,\end{aligned} \tag{5.5}$$

as our assessment of an item's time to failure with η and σ assumed known. It is convenient to summarize the above development as follows

Theorem 5.1. *The time to failure T of an item experiencing failure due to ageing or degradation described by a* ***Wiener Maximum Process*** *with a drift parameter $\eta > 0$, and a diffusion parameter $\sigma^2 > 0$, has the distribution function $F(t|\eta,\sigma)$ which is a location mixture of Inverse Gaussian Distributions. This distribution function, which is also the hitting time of the process to an exponential* (1) *random threshold, is given by* (5.5).

In Figure 1 we illustrate the behavior of the *IG*-Distribution function $F_x(t)$, for $x = 1, 2, 3, 4$, and 5, when $\eta = \sigma = 1$, and superimpose on these a plot of $F(t|\eta = \sigma = 1)$ to show the effect of averaging the threshold x. As can be expected, averaging makes the S-shapedness of the distribution functions less pronounced.

5.2. Assessing lifetimes using surrogate (biomarker) data

The material leading up to Theorem 5.1 is based on the thesis that η and σ^2 are known. In actuality, they are of course unknown. Thus, besides the hazard potential

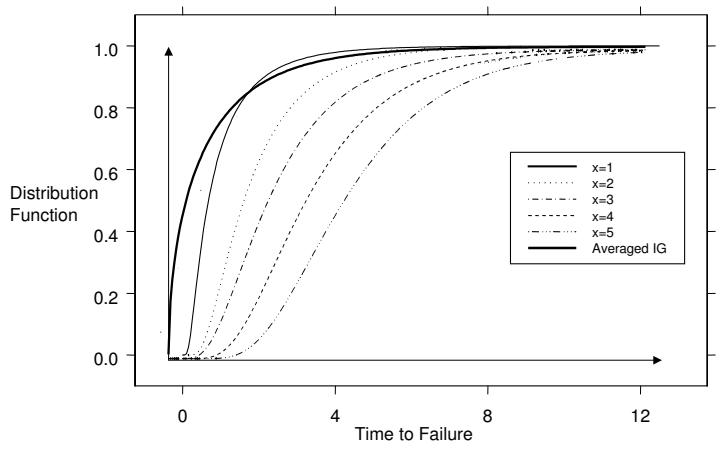


FIG 1. *The IG-Distribution with thresholds* $x = 1, \ldots, 5$ *and the averaged IG-Distribution.*

X, the η and σ^2 constitute the unknowns in our set-up. To assess η and σ^2 we may use prior information, and when available, data on the underlying processes Z_t and H_t. The prior on X is an exponential distribution with scale one, and this prior can also be updated using process data. In the remainder of this section, we focus attention on the case of a single item and describe the nature of the data that can be collected on it. We then outline an overall plan for incorporating these data into our analyses.

In Section 5.3 we give details about the inferential steps. The scenario of observing several items to failure in order to predict the lifetime of a future item will not be discussed.

In principle, we have assumed that H_t is an unobservable process. This is certainly true in our particular case when the observable marker process Z_t cannot be continuously monitored. Thus it is not possible to collect data on H_t. Contrast our scenario to that of Doksum [3], Lu and Meeker [7], and Lu, Meeker and Escobar [8], who assume that degradation is an observable process and who use data on degradation to predict an item's lifetime. We assume that it is the surrogate (or the biomarker) process Z_t that is observable, but only prior to T, the item's failure time. In some cases we may be able to observe Z_t at $t=T$, but doing so in the case of a single item would be futile, since our aim is to assess an unobserved T. Data on Z_t will certainly provide information about η and σ^2, but also about X; this is because for any $t < T$, we know that $X > Z(t)$. Thus, as claimed by Nair [9], data on (the observable surrogates of) degradation helps sharpen lifetime assessments, because a knowledge of η, σ^2 and X translates to a knowledge of T.

It is often the case – at least we assume so – that Z_t cannot be continuously monitored, so that observations on Z_t could be had only at times $0 < t_1 < t_2 < \cdots < t_k < T$, yielding $\mathbf{Z} = (Z(t_1), \ldots, Z(t_k))$ as data. Furthermore, based on $Z(t_k)$, we are able to assert that $X > Z(t_k)$. This means that our updated uncertainty about X will be encapsulated by a shifted exponential distribution with scale parameter one, and a location (or shift) parameter $Z(t_k)$.

Thus for an item experiencing failure due to degradation, whose marker process yields $\mathbf{Z}$ as data, our aim will be to assess the item's *residual life* $(T - t_k)$. That is, for any $u > 0$, we need to know $\Pr(T > t_k + u; \mathbf{Z}) = \Pr(T > t_k + u; T > t_k)$, and this under a certain assumption (cf. Singpurwalla [12]) is tantamount to knowing

$$\frac{\Pr(T > t_k + u)}{\Pr(T > t_k)}, \tag{5.6}$$

for $0 < u < \infty$. To assess the two quantities in the above ratio, we need to consider the quantity $\Pr(T > t; \mathbf{Z})$, for some $t > 0$. Let $\pi(\eta, \sigma^2, x; \mathbf{Z})$ encapsulate our uncertainty about η, σ^2 and X in the light of the data $\mathbf{Z}$. In Section 5.3 we describe our approach for assessing $\pi(\eta, \sigma^2, x; \mathbf{Z})$. Now

$$\begin{aligned}(5.7)\quad \Pr(T > t; \mathbf{Z}) &= \int_{\eta,\sigma^2,x} \Pr(T > t|\eta, \sigma^2, x; \mathbf{Z})\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx) \\ &= \int_{\eta,\sigma^2,x} \Pr(T_x > t|\eta, \sigma^2)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx) \\ (5.8)\qquad &= \int_{\eta,\sigma^2,x} F_x(t|\eta, \sigma)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx),\end{aligned}$$

where $F_x(t|\eta, \sigma)$ is the IG-Distribution of (5.3).

Implicit to going from (5.7) to (5.8) is the assumption that the event $(T > t)$ is independent of $\mathbf{Z}$ given η, σ^2 and X. In Section 5.3 we will propose that η be allowed to vary between a and b; also, $\sigma^2 > 0$, and having observed $Z(t_k)$, it is clear that x must be greater than $Z(t_k)$. Consequently, (5.8) gets written as

$$(5.9)\qquad \Pr(T > t; \mathbf{Z}) = \int_a^b \int_0^\infty \int_{Z(t_k)}^\infty F_x(t|\eta, \sigma)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx),$$

and the above can be used to obtain $\Pr(T > t_k + u; \mathbf{Z})$ and $\Pr(T > t_k; \mathbf{Z})$. Once these are obtained, we are able to assess the residual life $\Pr(T > t_k + u|T > t_k)$, for $u > 0$.

We now turn our attention to describing a Bayesian approach specifying $\pi(\eta, \sigma^2, x; \mathbf{Z})$.

5.3. Assessing the posterior distribution of η, σ^2 and X

The purpose of this section is to describe an approach for assessing $\pi(\eta, \sigma^2, x; \mathbf{Z})$, the posterior distribution of the unknowns in our set-up. For this, we start by supposing that $\mathbf{Z}$ is an unknown and consider the quantity $\pi(\eta, \sigma^2, x|\ \mathbf{Z})$. This is done to legitimize the ensuing simplifications. By the multiplication rule, and using obvious notation

$$\pi(\eta, \sigma^2, x|\mathbf{Z}) = \pi_1(\eta, \sigma^2|X, \mathbf{Z})\pi_2(X|\mathbf{Z}).$$

It makes sense to suppose that η and σ^2 do not depend on X; thus

$$(5.10)\qquad \pi(\eta, \sigma^2, x|\mathbf{Z}) = \pi_1(\eta, \sigma^2|\mathbf{Z})\pi_2(X|\mathbf{Z}).$$

However, $\mathbf{Z}$ is an observed quantity. Thus (5.10) needs to be recast as:

$$(5.11)\qquad \pi(\eta, \sigma^2, x; \mathbf{Z}) = \pi_1(\eta, \sigma^2; \mathbf{Z})\pi_2(X; \mathbf{Z}).$$

Regarding the quantity $\pi_2(X; \mathbf{Z})$, the only information that $\mathbf{Z}$ provides about X is that $X > Z(t_k)$. Thus $\pi_2(X; \mathbf{Z})$ becomes $\pi_2(X; Z(t_k))$. We may now invoke Bayes' law on $\pi_2(X; Z(t_k))$ and using the facts that the prior on X is an exponential (1) distribution on $(0, \infty)$, obtain the result that the posterior of X is also an exponential (1) distribution, but on $(Z(t_k), \infty)$. That is, $\pi_2(X; Z(t_k))$ is a shifted exponential distribution of the form $\exp(-(x - Z(t_k)))$, for $x > Z(t_k)$.

Turning attention to the quantity $\pi_1(\eta, \sigma^2; \mathbf{Z})$ we note, invoking Bayes' law, that

$$(5.12)\qquad \pi_1(\eta, \sigma^2; \mathbf{Z}) \propto \mathcal{L}(\eta, \sigma^2; \mathbf{Z})\pi^*(\eta, \sigma^2),$$

where $\mathcal{L}(\eta, \sigma^2; \mathbf{Z})$ is the likelihood of η and σ^2 with $\mathbf{Z}$ fixed, and $\pi^*(\eta, \sigma^2)$ our prior on η and σ^2. In what follows we discuss the nature of the likelihood and the prior.

The Likelihood of η and σ^2

Let $Y_1 = Z(t_1),\ Y_2 = (Z(t_2) - Z(t_1)), \ldots, Y_k = (Z(t_k) - Z(t_{k-1}))$, and $s_1 = t_1,\ s_2 = t_2 - t_1, \ldots, s_k = t_k - t_{k-1}$. Because the Wiener process has independent increments, the y_i's are independent. Also, $y_i \sim N(\eta s_i, \sigma^2 s_i),\ i = 1, \ldots, k$, where $N(\mu, \xi^2)$ denotes a Gaussian distribution with mean μ and variance ξ^2. Thus, the joint density of the y_i's, $i = 1, \ldots, k$, which is useful for writing out a likelihood of η and σ^2, will be of the form

$$\prod_{i=1}^{k} \phi\left(\frac{y_i - \eta s_i}{\sigma^2 s_i}\right);$$

where ϕ denotes a standard Gaussian probability density function. As a consequence of the above, the likelihood of η and σ^2 with $\mathbf{y} = (y_1, \ldots, y_k)$ fixed, can be written as:

$$\mathcal{L}(\eta, \sigma^2; \mathbf{y}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi s_i}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - \eta s_i}{\sigma^2 s_i}\right)^2\right]. \tag{5.13}$$

The Prior on η and σ^2

Turning attention to $\pi^*(\eta, \sigma^2)$, the prior on η and σ^2, it seems reasonable to suppose that η and σ^2 are not independent. It makes sense to suppose that the fluctuations of Z_t depend on the trend η. The larger the η, the bigger the σ^2, so long as there is a constraint on the value of η. If η is not constrained the marker will take negative values. Thus, we need to consider, in obvious notation

$$\pi^*(\eta, \sigma^2) = \pi^*(\sigma^2|\eta)\pi^*(\eta). \tag{5.14}$$

Since η can take values in $(0, \infty)$, and since $\eta = \tan\theta$ – see Figure 2 – θ must take values in $(0, \pi/2)$.

To impose a constraint on η, we may suppose that θ has a *translated beta* density on (a, b), where $0 < a < b < \pi/2$. That is, $\theta = a + (b - a)W$, where W has a beta distribution on $(0, 1)$. For example, a could be $\pi/8$ and b could be $3\pi/8$. Note that were θ assumed to be uniform over $(0, \pi/2)$, then η will have a density of the form $2/[\pi(1 + \eta^2)]$ – which is a *folded Cauchy*.

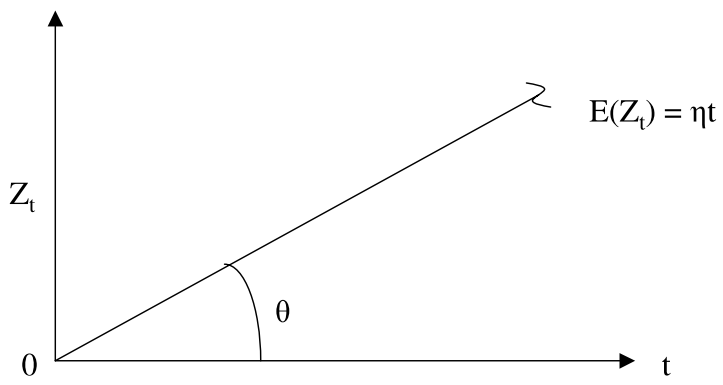


FIG 2. *Relationship between Z_t and η.*

The choice of $\pi^*(\sigma^2|\eta)$ is trickier. The usual approach in such situations is to opt for natural conjugacy. Accordingly, we suppose that $\psi \stackrel{def}{=} \sigma^2$ has the prior

$$\pi^*(\psi|\eta) \propto \psi^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{\eta}{2\psi}\right), \tag{5.15}$$

where ν is a parameter of the prior.

Note that $E(\psi|\eta,\nu) = \eta/(\nu-2)$, and so $\psi = \sigma^2$ increases with η, and η is constrained over a and b. Thus a constraint on σ^2 as well.

To pin down the parameter ν, we anchor on time $t = 1$, and note that since $E(Z_1) = \eta$ and $\mathrm{Var}(Z_1) = \sigma^2 = \psi$, σ should be such that $\Delta\sigma$ should not exceed η for some $\Delta = 1, 2, 3, \ldots$; otherwise Z_1 will become negative. With $\Delta = 3$, $\eta = 3\sigma$ and so $\psi = \sigma^2 = \eta^2/9$. Thus ν should be such that $E(\sigma^2|\eta,\nu) \approx \eta^2/9$. But $E(\sigma^2|\eta,\nu) = \eta/(\nu-2)$, and therefore by setting $\eta/(\nu-2) = \eta^2/9$, we would have $\nu = 9/\eta + 2$. In general, were we to set $\eta = \Delta\sigma$, $\nu = \Delta^2/\eta + 2$, for $\Delta = 1, 2, \ldots$. Consequently, $\nu/2 + 1 = (\Delta^2/\eta + 2)/2 + 1 = \Delta^2/2\eta + 2$, and thus

$$\pi^*(\psi|\eta;\Delta) = \psi^{-\left(\frac{\Delta^2}{2\eta}+2\right)} \exp\left(-\frac{\eta}{2\psi}\right), \tag{5.16}$$

would be our prior of σ^2, conditioned on ψ, and $\Delta = 1, 2, \ldots$, serving as a prior parameter. Values of Δ can be used to explore sensitivity to the prior.

This completes our discussion on choosing priors for the parameters of a Wiener process model for Z_t. All the necessary ingredients for implementing (5.9) are now at hand. This will have to be done numerically; it does not appear to pose major obstacles. We are currently working on this matter using both simulated and real data.

6. Conclusion

Our aim here was to describe how Lehmann's original ideas on (positive) dependence framed in the context of non-parametrics have been germane to reliability and survival analysis, and even so in the context of survival dynamics. The notion of a hazard potential has been the "hook" via which we can attribute the cause of dependence, and also to develop a framework for an appreciation of competing risks and degradation. The hazard potential provides a platform through which the above can be discussed in a unified manner. Our platform pertains to the hitting times of stochastic processes to a random threshold. With degradation modeling, the unobservable cumulative hazard function is seen as the metric of degradation (as opposed to an observable, like crack growth) and when modeling competing risks, the cumulative hazard is interpreted as a risk. Our goal here was not to solve any definitive problem with real data; rather, it was to propose a way of looking at two commonly encountered problems in reliability and survival analysis, problems that have been well discussed, but which have not as yet been recognized as having a common framework. The material of Section 5 is purely illustrative; it shows what is possible when one has access to real data. We are currently persuing the details underlying the several avenues and possibilities that have been outlined here.

Acknowledgements

The author acknowledges the input of Josh Landon regarding the hitting time of a Brownian maximum process, and Bijit Roy in connection with the material of

Section 5. The idea of using Wiener Maximum Processes for the cumulative hazard was the result of a conversation with Tom Kurtz.

References

[1] Bogdanoff, J. L. and Kozin, F. (1985). *Probabilistic Models of Cumulative Damage*. John Wiley and Sons, New York.

[2] Cinlar, E. (1972). Markov additive processes. II. *Z. Wahrsch. Verw. Gebiete* **24**, 94–121.

[3] Doksum, K. A. (1991). Degradation models for failure time and survival data. *CWI Quarterly, Amsterdam* **4**, 195–203.

[4] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Stat.* **24**, 23–43.

[5] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Stat.* **37**, 1137–1135.

[6] Lemoine, A. J. and Wenocur, M. L. (1989). On failure modeling. *Naval Research Logistics Quarterly* **32**, 497–508.

[7] Lu, C. J. and Meeker, W. Q. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics* **35**, 161–174.

[8] Lu, C. J., Meeker, W. Q. and Escobar, L. A. (1996). A comparison of degradation and failure-time analysis methods for estimating a time-to-failure distribution. *Statist. Sinica* **6**, 531–546.

[9] Nair, V. N. (1988). Discussion of "Estimation of reliability in field-performance studies" by J. D. Kalbfleisch and J. F. Lawless. *Technometrics* **30**, 379–383.

[10] Singpurwalla, N. D. (1997). Gamma processes and their generalizations: An overview. In *Engineering Probabilistic Design and Maintenance for Flood Protection* (R. Cook, M. Mendel and H. Vrijling, eds.). Kluwer Acad. Publishers, 67–73.

[11] Singpurwalla, N. D. (2005). Betting on residual life. Technical report. The George Washington University.

[12] Singpurwalla, N. D. (2006). The hazard potential: Introduction and overview. *J. Amer. Statist. Assoc.*, to appear.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 241–252
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000482

Restricted estimation of the cumulative incidence functions corresponding to competing risks

Hammou El Barmi[1] **and Hari Mukerjee**[2]

Baruch College, City University of New York and Wichita State University

Abstract: In the competing risks problem, an important role is played by the cumulative incidence function (CIF), whose value at time t is the probability of failure by time t from a particular type of failure in the presence of other risks. In some cases there are reasons to believe that the CIFs due to various types of failure are linearly ordered. El Barmi et al. [3] studied the estimation and inference procedures under this ordering when there are only two causes of failure. In this paper we extend the results to the case of k CIFs, where $k \geq 3$. Although the analyses are more challenging, we show that most of the results in the 2-sample case carry over to this k-sample case.

1. Introduction

In the competing risks model, a unit or subject is exposed to several risks at the same time, but the actual failure (or death) is attributed to exactly one cause. Suppose that there are $k \geq 3$ risks and we observe (T, δ), where T is the time of failure and $\{\delta = j\}$ is the event that the failure was due to cause j, $j = 1, 2, \ldots, k$. Let F be the distribution function (DF) of T, assumed to be continuous, and let $S = 1 - F$ be its survival function (SF).

The cumulative incidence function (CIF) due to cause j is a sub-distribution function (SDF), defined by

$$F_j(t) = P[T \leq t,\ \delta = j],\ j = 1, 2, \ldots, k, \tag{1.1}$$

with $F(t) = \sum_j F_j(t)$. The cause specific hazard rate due to cause j is defined by

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}[t \leq T < t + \Delta t,\ \delta = j \mid T \geq t],\ j = 1, 2, \ldots, k,$$

and the overall hazard rate is $\lambda(t) = \sum_j \lambda_j(t)$. The CIF, $F_j(t)$, may be written as

$$F_j(t) = \int_0^t \lambda_j(u) S(u)\, du. \tag{1.2}$$

Experience and empirical evidence indicate that in some cases the cause specific hazard rates or the CIFs are ordered, i.e.,

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \ \text{ or } \ F_1 \leq F_2 \leq \cdots \leq F_k.$$

[1]Department of Statistics and Computer Information Systems, Baruch College, City University of New York, New York, NY 10010, e-mail: hammou_elbarmi@baruch.cuny.edu

[2]Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033.

AMS 2000 subject classifications: primary 62G05; secondary 60F17, 62G30.

Keywords and phrases: competing risks, cumulative incidence functions, estimation, hypothesis test, k-sample problems, order restriction, weak convergence.

The hazard rate ordering implies the stochastic ordering of the CIFs, but not *vice versa*. Thus, the stochastic ordering of the CIFs is a milder assumption. El Barmi et al. [3] discussed the motivation for studying the restricted estimation using several real life examples and developed statistical inference procedures under this stochastic ordering, but only for $k = 2$. They also discussed the literature on this subject extensively. They found that there were substantial improvements by using the restricted estimators. In particular, the asymptotic mean squared error (AMSE) is reduced at points where two CIFs cross. For two stochastically ordered DFs with (small) independent samples, Rojo and Ma [17] showed essentially a uniform reduction of MSE when an estimator similar to ours is used in place of the nonparametric maximum likelihood estimator (NPMLE) using simulations. Rojo and Ma [17] also proved that the estimator is better in risk for many loss functions than the NPMLE in the one-sample problem and a simulation study suggests that this result extends to the 2-sample case. The purpose of this paper is to extend the results of El Barmi et al. [3] to the case where $k \geq 3$. The NPMLEs for k continuous DFs or SDFs under stochastic ordering are not known. Hogg [7] proposed a pointwise isotonic estimator that was used by El Barmi and Mukerjee [4] for k stochastically ordered continuous DFs. We use the same estimator for our problem. As far as we are aware, there are no other estimators in the literature for these problems. In Section 2 we describe our estimators and show that they are strongly uniformly consistent. In Section 3 we study the weak convergence of the resulting processes. In Section 4 we show that confidence intervals using the restricted estimators instead of the empiricals could possibly increase the coverage probability. In Section 5 we compare asymptotic bias and mean squared error of the restricted estimators with those of the unrestricted ones, and develop procedures for computing confidence intervals. In Section 6 we provide a test for testing equality of the CIFs against the alternative that they are ordered. In Section 7 we extend our results to the censoring case. Here, the results essentially parallel those in the uncensored case using the Kaplan-Meier [9] estimators for the survival functions instead of the empiricals. In Section 8 we present an example to illustrate our results. We make some concluding remarks in Section 9.

2. Estimators and consistency

Suppose that we have n items exposed to k risks and we observe (T_i, δ_i), the time and cause of failure of the ith item, $1 \leq i \leq n$. On the basis of this data, we wish to estimate the CIFs, $F_1, F_2, \ldots, F_k$, defined by (1.1) or (1.2), subject to the order restriction

$$F_1 \leq F_2 \leq \cdots \leq F_k. \tag{2.1}$$

It is well known that the NPMLE in the unrestricted case when $k = 2$ is given by (see Peterson, [12])

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \leq t,\, \delta_i = j),\ j = 1, 2, \tag{2.2}$$

and this result extends easily to $k > 2$. Unfortunately, these estimators are not guaranteed to satisfy the order constraint (2.1). Thus, it is desirable to have estimators that satisfy this order restriction. Our estimation procedure is as follows.

For each t, define the vector $\hat{\mathbf{F}}(t) = (\hat{F}_1(t), \hat{F}_2(t), \ldots, \hat{F}_k(t))^T$ and let $\mathcal{I} = \{\mathbf{x} \in R^k : x_1 \leq x_2 \leq \cdots \leq x_k\}$, a closed, convex cone in R^k. Let $E(\mathbf{x}|\mathcal{I})$ denote the least

squares projection of $\mathbf{x}$ onto $\mathcal{I}$ with equal weights, and let

$$Av[\hat{\mathbf{F}}; r, s] = \frac{\sum_{j=r}^{s} \hat{F}_j}{s - r + 1}.$$

Our restricted estimator of F_i is

$$\hat{F}_i^* = \max_{r \le i} \min_{s \ge i} Av[\hat{\mathbf{F}}; r, s] = E((\hat{F}_1, \ldots, \hat{F}_k)^T | \mathcal{I})_i, \quad 1 \le i \le k. \tag{2.3}$$

Note that for each t, equation (2.3) defines the isotonic regression of $\{\hat{F}_i(t)\}_{i=1}^k$ with respect to the simple order with equal weights. Robertson et al. [13] has a comprehensive treatment of the properties of isotonic regression. It can be easily verified that the $\hat{F}_i^*$s are CIFs for all i, and that $\sum_{i=1}^k \hat{F}_i^*(t) = \hat{F}(t)$, where $\hat{F}$ is the empirical distribution function of T, given by $\hat{F}(t) = \sum_{i=1}^n I(T_i \le t)/n$ for all t.

Corollary B, page 42, of Robertson et al. [13] implies that

$$\max_{1 \le j \le k} |\hat{F}_j^*(t) - F_j(t)| \le \max_{1 \le j \le k} |\hat{F}_j(t) - F_j(t)| \quad \text{for each } t.$$

Therefore $\| \hat{F}_i^* - F_i \| \le \max_{1 \le j \le k} ||\hat{F}_j - F_j||$ for all i where $||.||$ is used to denote the sup norm. Since $||\hat{F}_i - F_i|| \to 0$ *a.s.* for all i, we have

Theorem 2.1. $P[||\hat{F}_i^* - F_i|| \to 0 \text{ as } n \to \infty, \ i = 1, 2, \ldots, k] = 1.$

If $k = 2$, the restricted estimators of F_1 and F_2 are $\hat{F}_1^* = \hat{F}_1 \wedge \hat{F}/2$ and $\hat{F}_2^* = \hat{F}_1 \vee \hat{F}/2$, respectively. Here $\wedge$ ($\vee$) is used to denote max (min). This case has been studied in detail in El Barmi et al. [3].

3. Weak convergence

Weak convergence of the process resulting from an estimator similar to (2.3) when estimating two stochastically ordered distributions with independent samples was studied by Rojo [15]. Rojo [16] also studied the same problem using the estimator in (2.3). Praestgaard and Huang [14] derived the weak convergence of the NPMLE. El Barmi et al. [3] studied the weak convergence of two CIFs using (2.3). Here we extend their results to the k-sample case. Define

$$Z_{in} = \sqrt{n}[\hat{F}_i - F_i] \quad \text{and} \quad Z_{in}^* = \sqrt{n}[\hat{F}_i^* - F_i], \ i = 1, 2, \ldots, k.$$

It is well known that

$$(Z_{1n}, Z_{2n}, \ldots, Z_{kn})^T \stackrel{w}{\Longrightarrow} (Z_1, Z_2, \ldots, Z_k)^T, \tag{3.1}$$

a k-variate Gaussian process with the covariance function given by

$$Cov(Z_i(s), Z_j(t)) = F_i(s)[\delta_{ij} - F_j(t)], \quad 1 \le i, j \le k, \quad \text{for } s \le t,$$

where δ_{ij} is the Kronecker delta. Therefore, $Z_i \stackrel{d}{=} B_i^0(F_i)$ for all i, the B_i^0s being dependent standard Brownian bridges.

Weak convergence of the starred processes is a direct consequence of this and the continuous mapping theorem. First, we consider the convergence in distribution at a fixed point, t. Let

$$\mathcal{S}_{it} = \{j : F_j(t) = F_i(t)\}, \ i = 1, 2, \ldots, k. \tag{3.2}$$

Note that $\mathcal{S}_{it}$ is an interval of consecutive integers from $\{1, 2, \ldots, k\}$, $F_j(t) - F_i(t) = 0$ for $j \in \mathcal{S}_{it}$, and, as $n \to \infty$,

$$\sqrt{n}[F_j(t) - F_i(t)] \to \infty, \text{ and } \sqrt{n}[F_j(t) - F_i(t)] \to -\infty, \tag{3.3}$$

for $j > i^*(t)$ and $j < i_*(t)$, respectively, where $i_*(t) = \min\{j : j \in S_{it}\}$ and $i^*(t) = \max\{j : j \in S_{it}\}$.

Theorem 3.1. *Assume that* (2.1) *holds and t is fixed. Then*

$$(Z^*_{1n}(t), Z^*_{2n}(t), \ldots, Z^*_{kn}(t))^T \xrightarrow{d} (Z^*_1(t), Z^*_2(t), \ldots, Z^*_k(t))^T,$$

where

$$Z^*_i(t) = \max_{i_*(t) \le r \le i} \min_{i \le s \le i^*(t)} \frac{\sum_{\{r \le j \le s\}} Z_j(t)}{s - r + 1}. \tag{3.4}$$

Except for the order restriction, there are no restrictions on the F_is for the convergence in distribution at a point in Theorem 2. For $k = 2$, if the F_is are *distribution functions* and the $\hat{F}_i$s are the empiricals based on *independent* random samples of sizes n_1 and n_2, then, using restricted estimators $\hat{F}^*_i$s that are slightly different from those in (2.3), Rojo [15] showed that the weak convergence of $(Z^*_{1n_1}, Z^*_{2n_2})$ fails if $F_1(b) = F_2(b)$ and $F_1 < F_2$ on $(b, c]$ for some $b < c$ with $0 < F_2(b) < F_2(c) < 1$. El Barmi et al. [3] showed that the same is true for two CIFs. They also showed that, if $F_1 < F_2$ on $(0, b)$ and $F_1 = F_2$ on $[b, \infty)$, with $F_1(b) > 0$, then weak convergence holds, but the limiting process is discontinuous at b with positive probability. Thus, some restrictions are needed for weak convergence of the starred processes.

Let c_i (d_i) be the left (right) endpoint of the support of F_i, and let $\mathcal{S}_i = \{j : F_j \equiv F_i\}$ for $i = 1, 2, \ldots, k$. In most applications $c_i \equiv 0$. Letting $i^* = \max\{j : j \in S_i\}$, we assume that, for $i = 1, 2, \ldots, k-1$,

$$\inf_{c_i + \eta \le t \le d_i - \eta} [F_j(t) - F_i(t)] > 0 \text{ for all } \eta > 0 \text{ and } j > i^*. \tag{3.5}$$

Note that $i \in \mathcal{S}_i$ for all i. Assumption (3.5) guarantees that, if $F_j \ge F_i$, then, either $F_j \equiv F_i$ or $F_j(t) > F_i(t)$, except possibly at the endpoints of their supports. This guarantees that the pathology of nonconvergence described in Rojo [15] does not occur. Also, from the results in El Barmi et al. [3] discussed above, if $d_i = d_j$ for some $i \ne j \notin \mathcal{S}_i$, then weak convergence will hold, but the paths will have jumps at d_i with positive probability. We now state these results in the following theorem.

Theorem 3.2. *Assume that condition* (2.1) *and assumption* (3.5) *hold. Then*

$$(Z^*_{1n}, Z^*_{2n}, \ldots, Z^*_{kn})^T \xRightarrow{w} (Z^*_1, Z^*_2, \ldots, Z^*_k)^T,$$

where

$$Z^*_i = \max_{i_* \le r \le i} \min_{i \le s \le i^*} \frac{\sum_{\{r \le j \le s\}} Z_j}{s - r + 1}.$$

Note that, if $\mathcal{S}_i = \{i\}$, then $Z^*_{in} \xRightarrow{w} Z_i$ under the conditions of the theorem.

4. A stochastic dominance result

In the 2-sample case, El Barmi et al. [3] showed that $|Z^*_j|$ is stochastically dominated by $|Z_j|$ in the sense that

$$P[|Z^*_j(t)| \le u] > P[|Z_j(t)| \le u], \ j = 1, 2, \text{ for all } u > 0 \text{ and for all } t,$$

if $0 < F_1(t) = F_2(t) < 1$. This is an extension of Kelly's [10] result for *independent* samples case, but restricted to $k = 2$; Kelly called this result a reduction of stochastic loss by isotonization. Kelly's [10] proof was inductive. For the 2-sample case, El Barmi et al. [3] gave a constructive proof that showed the fact that the stochastic dominance result given above holds even when the order restriction is violated along some contiguous alternatives. We have been unable to provide such a constructive proof for the k-sample case; however, we have been able to extend Kelly's [10] result to our (special) dependent case.

Theorem 4.1. *Suppose that for some* $1 \le i \le k$, S_{it}, *as defined in* (3.2), *contains more than one element for some* t *with* $0 < F_i(t) < 1$. *Then, under the conditions of Theorem* 3,

$$P[|Z_i^*(t)| \le u] > P[|Z_i(t)| \le u] \quad \textit{for all } u > 0.$$

Without loss of generality, assume that $S_{it} = \{j : F_j(t) = F_i(t)\} = \{1, 2, \ldots, l\}$ for some $2 \le l \le k$. Note that $\{Z_i(t)\}$ is a multivariate normal with mean $\mathbf{0}$, and

$$Cov(Z_i(t), Z_j(t)) = F_1(t)[\delta_{ij} - F_1(t)], \quad 1 \le i, j \le l. \tag{4.1}$$

Also note that $\{Z_j^*(t); 1 \le j \le k\}$ is the isotonic regression of $\{Z_j(t) : 1 \le j \le k\}$ with equal weights from its form in (3.4). Define

$$\mathbf{X}^{(i)}(t) = (Z_1(t) - Z_i(t), Z_2(t) - Z_i(t), \ldots, Z_l(t) - Z_i(t))^T. \tag{4.2}$$

Kelly [10] shows that, on the set $\{Z_i(t) \ne Z_i^*(t)\}$,

$$P[|Z_i^*(t)| \le u \,|\, \mathbf{X}^{(i)}] > P[|Z_i(t)| \le u \,|\, \mathbf{X}^{(i)}] \ a.s. \quad \forall u > 0, \tag{4.3}$$

using the key result that $\mathbf{X}^{(i)}(t)$ and $Av(\mathbf{Z}(t); 1, k)$ are independent when the $Z_i(t)$'s are independent. Although the $Z_i(t)$s are not independent in our case, they are exchangeable random variables from (4.1). Computing the covariances, it easy to see that $\mathbf{X}^{(i)}(t)$ and $Av(\mathbf{Z}(t); 1, k)$ are independent in our case also. The rest of Kelly's [10] proof consists of showing that the left hand side of (4.3) is of the form $\Phi(a + v) - \Phi(a - v)$, while the right hand side of (4.3) is $\Phi(b + v) - \Phi(b - v)$ using (4.2), where Φ is the standard normal DF, and b is further away from 0 than a. This part of the argument depends only on properties of isotonic regression, and it is identical in our case. This concludes the proof of the theorem.

5. Asymptotic bias, MSE, confidence intervals

If $S_{it} = \{i\}$ for some i and t, then $Z_i^*(t) = Z_i(t)$ from Theorem 2, and they have the same asymptotic bias and AMSE. If S_{it} has more than one element, then, for $k = 2$, El Barmi et al. [3] computed the exact asymptotic bias and AMSE of $Z_i^*(t)$, $i = 1, 2$, using the representations, $Z_1^* = Z_1 + 0 \wedge (Z_2 - Z_1)/2$ and $Z_2^* = Z_2 - 0 \wedge (Z_2 - Z_1)/2$. The form of Z_i^* in (3.4) makes these computations intractable. However, from Theorem 4, we can conclude that $E[Z_i^*(t)]^2 < E[Z_i(t)]^2$, implying an improvement in AMSE when the restricted estimators are used.

From Theorem 4.1 it is clear that confidence intervals using the restricted estimators will be more conservative than those using the empiricals. Although we believe that the same will be true for confidence bands, we have not been able to prove it.

The confidence bands could always be improved by the following consideration. The $100(1-\alpha)\%$ simultaneous confidence bands, $[L_i, U_i]$, for $F_i,\ 1 \le i \le k$, in the unrestricted case obey the following probability inequality

$$P(F_i \in [L_i, U_i] : 1 \le i \le k) \ge 1 - \alpha.$$

Under our model, $F_1 \le F_2 \le \cdots \le F_k$, this probability is not reduced if we replace $[L_i, U_i]$ by $[L_i^*, U_i^*]$, where

$$L_i^* = \max\{L_j : 1 \le j \le i\} \quad \text{and} \quad U_i^* = \min\{U_j : i \le j \le k\}, 1 \le i \le k.$$

6. Hypotheses testing

Let $H_0 : F_1 = F_2 = \cdots = F_k$ and $H_a : F_1 \le F_2 \le \cdots \le F_k$. In this section we propose an asymptotic test of H_0 against $H_a - H_0$. This problem has already been considered by El Barmi et al. [3] when $k = 2$, and the test statistic they proposed is $T_n = \sqrt{n}\sup_{x\ge 0}[\hat{F}_2(x) - \hat{F}_1(x)]$. They showed that under H_0,

$$\lim_{n\to\infty} P(T_n > t) = 2(1 - \Phi(t)), \quad t \ge 0, \tag{6.1}$$

where Φ is the standard normal distribution function.

For $k > 2$, we use an extension of the sequential testing procedure in Hogg [7] for testing equality of distribution functions based on independent random samples. For testing $H_{0j} : F_1 = F_2 = \cdots = F_j$ against $H_{aj} - H_{0j}$, where $H_{aj} : F_1 = F_2 = \cdots = F_{j-1} \le F_j, \mathrm{J} = 2, 3, \ldots, k$, we use the test statistic $\sup_{x\ge 0} T_{jn}(x)$ where

$$T_{jn} = \sqrt{n}\sqrt{c_j}[\hat{F}_j - Av[\hat{\mathbf{F}}; 1, j-1]],$$

with $c_j = k(j-1)/j$. We reject H_{0j} for large values T_{jn}, that may be also written as

$$T_{jn} = \sqrt{c_j}[Z_{jn} - Av(\mathbf{Z}_n; 1, j-1)],$$

where $\mathbf{Z}_n = (Z_{1n}, Z_{2n}, \ldots, Z_{kn})^T$. By the weak convergence result in (3.1) and the continuous mapping theorem, $(T_{2n}, T_{3n}, \ldots, T_{kn})^T$ converges weakly to $(T_2, T_3, \ldots, T_k)^T$, where

$$T_j = \sqrt{c_j}[Z_j - Av[\mathbf{Z}; 1, j-1]].$$

A calculation of the covariances shows that the T_j's are independent. Also note that

$$T_j \stackrel{d}{=} B_j(F), \quad 2 \le j \le k,$$

where the B_j's are independent standard Brownian motions and $F = \sum_{i=1}^k F_i = kF_1$ under H_0. We define our test statistic for the overall test of H_0 against $H_a - H_0$ by

$$T_n = \max_{2\le j\le k} \sup_{x\ge 0} T_{jn}(x).$$

By the continuous mapping theorem, T_n converges in distribution to T, where

$$T = \max_{2\le j\le k} \sup_{x\ge 0} T_j(x).$$

Using the distribution of the maximum of a Brownian motion on $[0,1]$ (Billingsley [2]), and using the independence of the B_i's, the distribution of T is given by

$$\begin{aligned} P(T \geq t) &= 1 - P(\sup_x T_j(x) < t, j = 2, \ldots, k) \\ &= 1 - \prod_{j=2}^{k} P(\sup_x B_j(F(x)) < t) \\ &= 1 - [2\Phi(t) - 1]^{k-1}. \end{aligned}$$

This allows us to compute the p-value for an asymptotic test.

7. Censored case

The case when there is censoring in addition to the competing risks is considered next. It is important that the censoring mechanism, that may be a combination of other competing risks, be independent of the k risks of interest; otherwise, the CIFs cannot be estimated nonparametrically. We now denote the causes of failure as $\delta = 0, 1, 2, \ldots, k$, where $\{\delta = 0\}$ is the event that the observation was censored.

Let C_i denote the censoring time, assumed continuous, for the ith subject, and let $L_i = T_i \wedge C_i$. We assume that C_is are identically and independently distributed (IID) with survival function, S_C, and are independent of the life distributions, $\{T_i\}$. For the ith subject we observe (L_i, δ_i), the time and cause of the failure. Here the $\{L_i\}$ are IID by assumption.

7.1. The estimators and consistency

For $j = 1, 2, \ldots, k$, let Λ_j be the cumulative hazard function for risk j, and let $\Lambda = \Lambda_1 + \Lambda_2 + \cdots + \Lambda_k$ be the cumulative hazard function of the life time T. For the censored case, the unrestricted estimators of the CIFs are the sample equivalents of (1.2) using the Kaplan–Meier [9] estimator, $\widehat{S}$, of $S = 1 - F$:

$$\widehat{F}_j(t) = \int_0^t \widehat{S}(u)\, d\widehat{\Lambda}_j(u), \ j = 1, 2, \ldots, k, \tag{7.1}$$

with $\widehat{F} = \widehat{F}_1 + \widehat{F}_2 + \cdots + \hat{F}_k$, where $\widehat{S}$ is chosen to be the left-continuous version for technical reasons, and $\widehat{\Lambda}_j$ is the Nelson–Aalen estimator (see, e.g., Fleming and Harrington, [5]) of Λ_j. Although our estimators use the Kaplan–Meier estimator of S rather than the empirical, we continue to use the same notation for the various estimators and related entities as in the uncensored case for notational simplicity.

As in the uncensored case, we define our restricted estimator of F_i by

$$\begin{aligned} \hat{F}_i^* &= \max_{r \leq i} \min_{s \geq i} Av[\hat{\mathbf{F}}; r, s] \\ &= E((\hat{F}_1, \ldots, \hat{F}_k)^T | \mathcal{I})_i, \quad 1 \leq i \leq k. \end{aligned} \tag{7.2}$$

Let

$$\pi(t) = P[L_i \geq t] = P[T_i \geq t, C_i \geq t] = S(t)S_C(t).$$

Strong uniform consistency of the $\hat{F}_i^*$s on $[0, b]$ for all b with $\pi(b) > 0$ follows from those of the $\hat{F}_i$'s [see, e.g., Shorack and Wellner [18], page 306, and the corrections

posted on the website given in the reference] using the same arguments as in the proof of Theorem 2 in the uncensored case.

7.2. Weak convergence

Let $Z_{jn} = \sqrt{n}[\widehat{F}_j - F_j]$ and $Z^*_{jn} = \sqrt{n}[\hat{F}^*_j - F_j]$, $j = 1, 2, \ldots, k$, be defined as in the uncensored case, except that the unresticted estimators have been obtained via (7.1). Fix b such that $\pi(b) > 0$. Using a counting process-martingale formulation, Lin [11] derived the following representation of Z_{in} on $[0, b]$:

$$Z_{in}(t) = \sqrt{n}\int_0^t \frac{S(u)dM_i(u)}{Y(u)} - \sqrt{n}F_i(t)\int_0^t \frac{\sum_{j=1}^k dM_j(u)}{Y(u)} + \sqrt{n}\int_0^t \frac{F_i(u)\sum_{j=1}^k dM_j(u)}{Y(u)} + o_p(1),$$

where

$$Y(t) = \sum_{j=1}^n I(L_j \geq t) \text{ and } M_i(t) = \sum_{j=1}^n I(L_j \leq t) - \sum_{j=1}^n \int_0^t I(L_j \geq u)d\Lambda_i(u),$$

the M_i's being independent martingales. Using this representation, El Barmi et al. [3] proved the weak convergence of $(Z_{in}, Z_{2n})^T$ to a mean-zero Gaussian process, (Z_i, Z_2), with the covariances given in that paper. A generalization of their results yields the following theorem.

Theorem 7.1. *The process* $(Z_{1n}, Z_{2n}, \ldots, Z_{kn})^T \overset{w}{\Longrightarrow} (Z_1, Z_2, \ldots, Z_k)^T$ *on* $[0, b]^k$, *where* $(Z_1, Z_2, \ldots, Z_k)^T$ *is a mean-zero Gaussian process with the covariance functions, for* $s \leq t$,

$$\begin{aligned} Cov(Z_i(s), Z_i(t)) = &\int_0^s [1-F_i(s) - \sum_{j\neq i} F_j(u))][1-F_i(t) - \sum_{j\neq i} F_j(u))]\,\frac{d\Lambda_i(u)}{\pi(u)} \\ &+ \sum_{j\neq i}\int_0^s [F_i(u) - F_i(s)][F_i(u) - F_i(t)]\,\frac{d\Lambda_j(u)}{\pi(u)}. \end{aligned} \tag{7.3}$$

and, for $i \neq j$,

$$\begin{aligned} Cov(Z_i(s), Z_j(t)) = &\int_0^s [1 - F_i(s) - \sum_{l\neq i} F_l(u)][F_j(u) - F_j(t)]\frac{d\Lambda_i(u)}{\pi(u)} \\ &+ \int_0^s [1 - F_j(t) - \sum_{l\neq j} F_l(u)][F_i(u) - F_i(s)]\frac{d\Lambda_j(u)}{\pi(u)} \\ &+ \sum_{l\neq i,j}\int_0^s [F_j(s) - F_j(u)][F_i(t) - F_i(u)]\frac{d\Lambda_l(u)}{\pi(u)}. \end{aligned} \tag{7.4}$$

The proofs of the weak convergence results for the starred processes in Theorems 3.1 and 3.2 use only the weak convergence of the unrestricted processes and isotonization of the estimators; in particular, they do not depend on the distribution of $(Z_1, \ldots, Z_k)^T$. Thus, the proof of the following theorem is essentially identical to that used in proving Theorems 3.1 and 3.2; the only difference is that the domain has been restricted to $[0, b]^k$.

Theorem 7.2. *The conclusions of Theorems* 3.1 *and* 3.2 *hold for* $(Z_{1n}^*, Z_{2n}^*, \ldots, Z_{kn}^*)^T$ *defined above on* $[0,b]^k$ *under the assumptions of these theorems.*

7.3. Asymptotic properties

In the uncensored case, for a $t > 0$ and an i such that $0 < F_i(t) < 1$, if $S_{it} = \{1, \ldots, l\}$ and $l \geq 2$, then it was shown in Theorem 4 that

$$P(|Z_i^*(t)| \leq u) > P(|Z_i(t)| \leq u) \quad \text{for all } u > 0.$$

The proof only required that $\{Z_j(t)\}$ be a multivariate normal and that the random variables, $\{Z_j(t) : j \in S_{it}\}$, be exchangeable, which imply the independence of $\mathbf{X}^{(i)}(t)$ and $Av(\mathbf{Z}(t); 1, l)$, as defined there. Noting that $F_j(t) = F_i(t)$ for all $j \in S_{it}$, the covariance formulas given in Theorem 7.1 show that the multivariate normality and the exchangeability conditions hold for the censored case also. Thus, the conclusions of Theorem 4.1 continue to hold in the censored case.

All comments and conclusions about asymptotic bias and AMSE in the uncensored case continue to hold in the censored case in view of the results above.

7.4. Hypothesis test

Consider testing $H_0 : F_1 = F_2 = \cdots = F_k$ against $H_a - H_0$, where $H_a : F_1 \leq F_2 \leq \cdots \leq F_k$, using censored observations. As in the uncensored case, it is natural to reject H_0 for large values of $T_n = \max_{2 \leq j \leq k} \sup_{x \geq 0} T_{jn}(x)$, where

$$\begin{aligned} T_{jn}(x) &= \sqrt{n}\sqrt{c_j}[\hat{F}_j(x) - Av(\hat{\mathbf{F}}(x); 1, j-1)] \\ &= \sqrt{c_j}[Z_{jn}(x) - Av(\mathbf{Z}_n(x); 1, j-1)] \end{aligned}$$

with $c_j = k(j-1)/j$, is used to test the sub-hypothesis H_{0j} against $H_{aj} - H_{0j}$, $2 \leq j \leq k$, as in the uncensored case. Using a similar argument as in the uncensored case, under H_0, $(T_{2n}, T_{3n}, \ldots, T_{kn})^T$ converges weakly $(T_2, T_3, \ldots, T_k)^T$ on $[0,b]^k$, where the T_i's are independent mean zero Gaussian processes. For $s \leq t$, $Cov(T_i(s), T_i(t))$ simplifies to exactly the same form as in the 2-sample case in El Barmi et al. [3]:

$$Cov(T_i(s), T_i(t)) = \int_0^s S(u) \frac{d\Lambda(u)}{S_C(u)}.$$

The limiting distribution of

$$T_n = \max_{2 \leq j \leq k} \sup_{x \geq 0} [Z_{jn}(x) - Av(\mathbf{Z}_n(x); 1, j-1)]$$

is intractable. As in the 2-sample case, we utilize the strong uniform convergence of the Kaplan–Meier estimator, $\hat{S}_C$, of S_C, to define

$$T_{jn}^*(t) = \sqrt{n}\sqrt{c_j} \int_0^t \sqrt{\hat{S}_C(u)}\, d[\hat{F}_j(x) - Av(\hat{\mathbf{F}}(x), 1, j-1)], j = 2, 3, \ldots, k,$$

and define $T_n^* = \max_{2 \leq j \leq k} \sup_{x \geq 0} T_{jn}^*(x)$ to be the test statistic for testing the overall hypothesis of H_0 against $H_a - H_0$. By arguments similar to those used in the uncensored case, $(T_{2n}^*, T_{3n}^*, \ldots, T_{kn}^*)^T$ coverges weakly to $(T_2^*, T_3^*, \ldots, T_k^*)^T$, a mean zero Gaussian process with independent components with

$$T_j^* \stackrel{d}{=} B_j(F), \quad 2 \leq j \leq k,$$

where B_j is a standard Brownian motion, and T_n^* converges in distribution to a random variable T^*. Since T_j^* here and T_j in the uncensored case (Section 6) have the same distribution, $2 \leq j \leq k$, T^* has the same distribution as T in Section 6, i.e.,

$$P(T^* \geq t) = 1 - [2\Phi(t) - 1]^{k-1}.$$

Thus the testing problem is identical to that in the uncensored case, with T_n of Section 6 changed to T_n^* as defined above. This is the same test developed by Aly et al. [1], but using a different approach.

8. Example

We analyze a set of mortality data provided by Dr. H. E. Walburg, Jr. of the Oak Ridge National Laboratory and reported by Hoel [6]. The data were obtained from a laboratory experiment on 82 RFM strain male mice who had received a radiation dose of 300 rads at 5–6 weeks of age, and were kept in a conventional laboratory environment. After autopsy, the causes of death were classified as thymic lymphoma, reticulum cell sarcoma, and other causes. Since mice are known to be highly susceptible to sarcoma when irradiated (Kamisaku et al [8]), we illustrate our procedure for the uncensored case considering "other causes" as cause 2, reticulum cell sarcoma as cause 3, and thymic lymphoma as cause 1, making the assumption that $F_1 \leq F_2 \leq F_3$. The unrestricted estimators are displayed in Figure 1, the restricted estimators are displayed in Figure 2. We also considered the large sample test of $H_0 : F_1 = F_2 = F_3$ against $H_a - H_0$, where $H_a : F_1 \leq F_2 \leq F_3$, using the test described in Section 6. The value of the test statistic is 3.592 corresponding to a p-value of 0.00066.

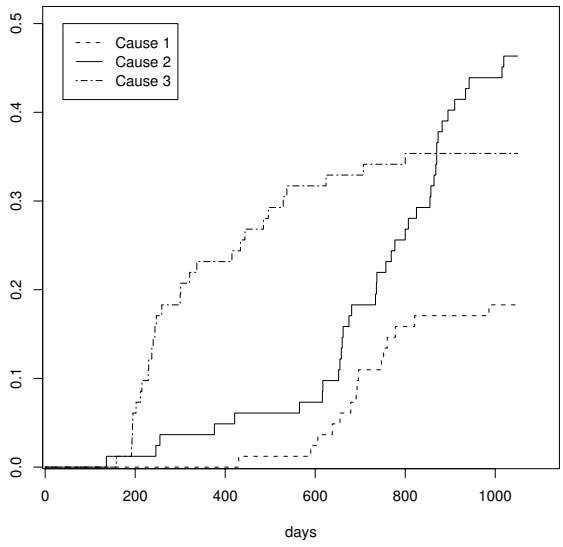


FIG 1. *Unrestricted estimators of the cumulative incidence functions.*

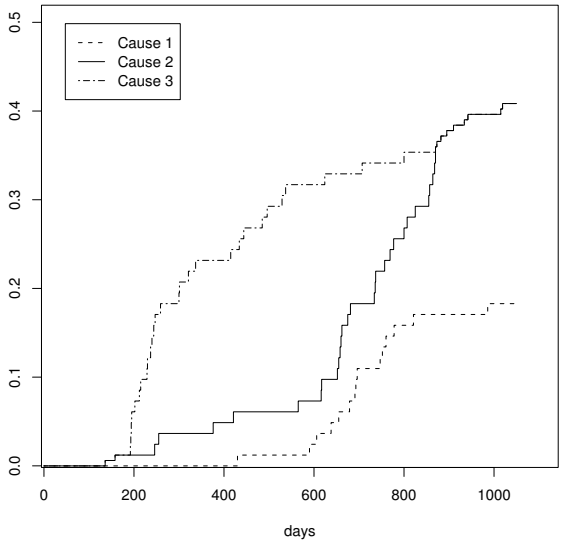


FIG 2. *Restricted estimators of the cumulative incidence functions.*

9. Conclusion

In this paper we have provided estimators of the CIFs of k competing risks under a stochasting ordering constraint, with and without censoring, thus extending the results for $k = 2$ in El Barmi et al. [3]. We have shown that the estimators are uniformly strongly consistent. The weak convergence of the estimators has been derived. We have shown that asymptotic confidence intervals are more conservative when the restricted estimators are used in place of the empiricals. We conjecture that the same is true for asymptotic confidence bands, although we have not been able to prove it. We have provided asymptotic tests for equality of the CIFs against the ordered alternative. The estimators and the test are illustrated using a set of mortality data reported by Hoel [6].

Acknowledgments

The authors are grateful to a referee and the Editor for their careful scrutiny and suggestions. It helped remove some inaccuracies and substantially improve the paper. El Barmi thanks the City University of New York for its support through PSC-CUNY.

References

[1] ALY, E.A.A., KOCHAR, S.C. AND MCKEAGUE, I.W. (1994). Some tests for comparing cumulative incidence functions and cause-specific hazard rates. *J. Amer. Statist. Assoc.* **89**, 994–999.

[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[3] EL BARMI, H., KOCHAR, S., MUKERJEE, H AND SAMANIEGO F. (2004). Estimation of cumulative incidence functions in competing risks studies under an order restriction. *J. Statist. Plann. Inference.* **118**, 145–165.

[4] El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: The k-sample case. *J. Amer. Statist. Assoc.* **100**, 252–261.

[5] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

[6] Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics* **28**, 475–478.

[7] Hogg, R. V. (1965). On models and hypotheses with restricted alternatives. *J. Amer. Statist. Assoc.* **60**, 1153–1162.

[8] Kamisaku, M, Aizawa, S., Kitagawa, M., Ikarashi, Y. and Sado, T. (1997). Limiting dilution analysis of T-cell progenitors in the bone marrow of thymic lymphoma susceptible B10 and resistant C3H mice after fractionated whole-body radiation. *Int. J. Radiat. Biol.* **72**, 191–199.

[9] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimator from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

[10] Kelly, R. (1989). Stochastic reduction of loss in estimating normal means by isotonic regression. *Ann. Statist.* **17**, 937–940.

[11] Lin, D.Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statist. Med.* **16**, 901–910.

[12] Peterson, A.V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.* **72**, 854–858.

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

[14] Praestgaard, J. T. and Huang, J. (1996). Asymptotic theory of nonparametric estimation of survival curves under order restrictions. *Ann. Statist.* **24**, 1679–1716.

[15] Rojo, J. (1995). On the weak convergence of certain estimators of stochastically ordered survival functions. *Nonparametric Statist.* **4**, 349–363.

[16] Rojo, J. (2004). On the estimation of survival functions under a stochastic order constraint. *Lecture Notes–Monograph Series* (J. Rojo and V. Pérez-Abreu, eds.) Vol. 44. Institute of Mathematical Statistics.

[17] Rojo, J. and Ma, Z. (1996). On the estimation of stochastically ordered survival functions. *J. Statist. Comp. Simul.* **55**, 1–21.

[18] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York. Corrections at `www.stat.washington.edu/jaw/RESEARCH/BOOKS/book1.html`

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 253–265
In the public domain
DOI: 10.1214/074921706000000491

Comparison of robust tests for genetic association using case-control studies

Gang Zheng[1], Boris Freidlin[2] and Joseph L. Gastwirth[3,*]

National Heart, Lung and Blood Institute, National Cancer Institute and George Washington University

Abstract: In genetic studies of complex diseases, the underlying mode of inheritance is often not known. Thus, the most powerful test or other optimal procedure for one model, e.g. recessive, may be quite inefficient if another model, e.g. dominant, describes the inheritance process. Rather than choose among the procedures that are optimal for a particular model, it is preferable to see a method that has high efficiency across a family of scientifically realistic models. Statisticians well recognize that this situation is analogous to the selection of an estimator of location when the form of the underlying distribution is not known. We review how the concepts and techniques in the efficiency robustness literature that are used to obtain efficiency robust estimators and rank tests can be adapted for the analysis of genetic data. In particular, several statistics have been used to test for a genetic association between a disease and a candidate allele or marker allele from data collected in case-control studies. Each of them is optimal for a specific inheritance model and we describe and compare several robust methods. The most suitable robust test depends somewhat on the range of plausible genetic models. When little is known about the inheritance process, the maximum of the optimal statistics for the extreme models and an intermediate one is usually the preferred choice. Sometimes one can eliminate a mode of inheritance, e.g. from prior studies of family pedigrees one may know whether the disease skips generations or not. If it does, the disease is much more likely to follow a recessive model than a dominant one. In that case, a simpler linear combination of the optimal tests for the extreme models can be a robust choice.

1. Introduction

For hypothesis testing problems when the model generating the data is known, optimal test statistics can be derived. In practice, however, the precise form of the underlying model is often unknown. Based on prior scientific knowledge a family of possible models is often available. For each model in the family an optimal test statistic is obtained. Hence, we have a collection of optimal test statistics corresponding to each member of the family of scientifically plausible models and need to select one statistic from them or create a robust one, that combines them. Since using any single optimal test in the collection typically results in a substantial loss of efficiency or power when another model is the true one, a robust procedure with reasonable power over the entire family is preferable in practice.

*Supported in part by NSF grant SES-0317956.

[1]Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892-7938, e-mail: zhengg@nhlbi.nih.gov

[2]Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892-7434, e-mail: freidlinb@ctep.nci.nih.gov

[3]Department of Statistics, George Washington University, Washington, DC 20052, e-mail: jlgast@gwu.edu

AMS 2000 subject classifications: primary 62F35, 62G35; secondary 62F03, 62P10.

Keywords and phrases: association, efficiency robustness, genetics, linkage, MAX, MERT, robust test, trend test.

The above situation occurs in many applications. For example, in survival analysis Harrington and Fleming [14] introduced a family of statistics G^ρ. The family includes the log-rank test ($\rho = 0$) that is optimal under the proportional hazards model and the Peto-Peto test ($\rho = 1$, corresponding to the Wilcoxon test without censoring) that is optimal under a logistic shift model. In practice, when the model is unknown, one may apply both tests to survival data. It is difficult to draw scientific conclusions when one of the tests is significant and the other is not. Choosing the significant test *after* one has applied both tests to the data increases the Type I error. A second example is testing for an association between a disease and a risk factor in contingency tables. If the risk factor is a categorical variable and has a natural order, e.g., number of packs of cigarette smoking per day, the Cochran-Armitage trend test is typically used. (Cochran [3] and Armitage [1]) To apply such a trend test, increasing scores as values of a covariate have to be assigned to each category of the risk factor. Thus, the p-value of the trend test may depend on such scores. A collection of trend tests is formed by choosing various increasing scores. (Graubard and Korn [12]) A third example arises in genetic linkage and association studies. In linkage analysis to map quantitative trait loci using affected sib pairs, optimal tests are functions of the number of alleles shared identical-by-descent (IBD) by the two sibs. The IBD probabilities form a family of alternatives which are determined by genetic models. See, e.g., Whittemore and Tu [22] and Gastwirth and Freidlin [10]. In genetic association studies using case-parents trios, the optimal test depends on the mode of inheritance of the disease (recessive, dominant, or co-dominant disease). For complex diseases, the underlying genetic model is often not known. Using a single optimal test does not protect against a substantial loss of power under the worst situation, i.e., when a much different genetic model is the true one. (Zheng, Freidlin and Gastwirth [25])

Robust procedures have been developed and applied when the underlying model is unknown as discussed in Gastwirth [7–9], Birnbaum and Laska [2], Podgor, Gastwirth and Mehta [16], and Freidlin, Podgor and Gastwirth [5]. In this article, we review two useful robust tests. The first one is a linear combination of the two or three extreme optimal tests in a family of optimal statistics and the second one is a suitably chosen maximum statistic, i.e., the maximum of several of the optimum tests for specific models in the family. These two robust procedures are applied to genetic association using case-control studies and compared to other test statistics that are used in practice.

2. Robust procedures: A short review

Suppose we have a collection of alternative models $\{M_i, i \in I\}$ and the corresponding optimal (most powerful) test statistics $\{T_i : i \in I\}$ are obtained, where I can be a finite set or an interval. Under the null hypothesis, assume that each of these test statistics is asymptotically normally distributed, i.e., $Z_i = [T_i - \mathrm{E}(T_i)]/\{\mathrm{Var}(T_i)\}^{1/2}$ converges in law to $N(0,1)$ where $\mathrm{E}(T_i)$ and $\mathrm{Var}(T_i)$ are the mean and the variance of T_i under the null; suppose also that for any $i, j \in I$, Z_i and Z_j are jointly normal with the correlation ρ_{ij}. When M_i is the true model, the optimal test Z_i would be used. When the true model M_i is unknown and the test Z_j is used, assume the Pitman asymptotic relative efficiency (ARE) of Z_j relative to Z_i is $e(Z_j, Z_i) = \rho_{ij}^2$ for $i, j \in I$. These conditions are satisfied in many applications. (van Eeden [21] and Gross [13])

2.1. Maximin efficiency robust tests

When the true model is unknown and each model in the family is scientifically plausible, the minimum ARE compared to the optimum test for each model, Z_i, when Z_j is used is given by $\inf_{i\in I} e(Z_j, Z_i)$ for $j \in I$. One robust test is to choose the optimal test Z_l from the family $\{Z_i : i \in I\}$ which maximizes the minimum ARE, that is,

$$\inf_{i\in I} e(Z_l, Z_i) = \sup_{j\in I} \inf_{i\in I} e(Z_j, Z_i). \tag{2.1}$$

Under the null hypothesis, Z_l converges in distribution to a standard normal random variable and, under the definition (2.1), is the most robust test in $\{Z_i : i \in I\}$. In practice, however, other tests have been studied which may have greater efficiency robustness.

Although a family of models are proposed based on scientific knowledge and the corresponding optimal tests can be obtained, all consistent tests with an asymptotically normal distribution can be used. Denote all these tests for the problem by C. The original family of test statistics can be expanded to C. The purpose is to find a test Z from C, rather than from the original family $\{Z_i : i \in I\}$, such that

$$\inf_{i\in I} e(Z, Z_i) = \sup_{Z\in C} \inf_{i\in I} e(Z, Z_i). \tag{2.2}$$

The test Z satisfying (2.2) is called maximin efficiency robust test (MERT). (Gastwirth [7]) When the family C is restricted to the convex linear combinations of $\{Z_i : i \in I\}$, the resulting robust test is denoted as Z_{MERT}. Since $\{Z_i : i \in I\} \subset C$,

$$\sup_{Z\in C} \inf_{i\in i} e(Z, Z_i) \geq \sup_{j\in I} \inf_{i\in I} e(Z_j, Z_i).$$

Assuming that $\inf_{i,j\in I} \rho_{ij} \geq \epsilon > 0$, Gastwirth [7] proved that Z_{MERT} uniquely exists and can be written as a closed convex combination of optimal tests Z_i in the family $\{Z_i : i \in i\}$. Although a simple algorithm when C is the class of linear combination of $\{Z_i : i \in I\}$ was given in Gastwirth [9] (see also Zucker and Lakatos [27]), the computation of Z_{MERT} is more complicated as it is related to quadratic programming algorithms. (Rosen [18]) For many applications, Z_{MERT} can be easily written as a linear convex combination of two or three optimal tests in $\{Z_i : i \in I\}$ including the extreme pair defined as follows: two optimal tests $Z_s, Z_t \in \{Z_i : i \in I\}$ are called extreme pair if $\rho_{st} = \mathrm{corr}_{H_0}(Z_s, Z_t) = \inf_{i,j\in I} \rho_{ij} > 0$. Define a new test statistic Z_{st} based on the extreme pair as

$$Z_{st} = \frac{Z_s + Z_t}{[2(1+\rho_{st})]^{1/2}}, \tag{2.3}$$

which is the MERT for the extreme pair. A necessary and sufficient condition for Z_{st} to be Z_{MERT} for the whole family $\{Z_i : i \in I\}$ is given, see Gastwirth [8], by

$$\rho_{si} + \rho_{it} \geq 1 + \rho_{st}, \quad \text{for all } i \in I. \tag{2.4}$$

Under the null hypothesis, Z_{MERT} is asymptotically $N(0,1)$. The ARE of the MERT given by (2.4) is $(1+\rho_{st})/2$. To find the MERT, the null correlations ρ_{ij} need to be obtained and the pair is the extreme pair for which ρ_{ij} is smallest.

2.2. Maximum tests

The robust test Z_{MERT} is a linear combination of the optimal test statistics and with modern computers it is useful to extend the family C of possible tests to include non-linear functions of the Z_i. A natural non-linear robust statistic is the maximum over the extreme pair (Z_s, Z_t) or the triple (Z_s, Z_u, Z_t) for the entire family (Freidlin et al. [5]), i.e.,

$$Z_{\mathrm{MAX2}} = \max(Z_s, Z_t) \text{ or } Z_{\mathrm{MAX3}} = \max(Z_s, Z_u, Z_t).$$

There are several choices for Z_u in Z_{MAX3}, e.g., $Z_u = Z_{st}$ (MERT for the extreme pair or entire family). As when obtaining the MERT, the correlation matrix $\{\rho_{ij}\}$ guides the choice of Z_u to be used in MAX3, e.g., it has equal correlation with the extreme tests. A more complicated maximum test statistic is to take the maximum over the entire family $Z_{\mathrm{MAX}} = \max_{i\in I} Z_i$ or $Z_{\mathrm{MAX}} = \max_{i\in C} Z_i$. Z_{MAX} was considered by Davies [4] for some non-standard hypothesis testing whose critical value has to be determined by approximation of its upper bound. In a recent study of several applications in genetic association and linkage analysis, Zheng and Chen [24] showed that Z_{MAX3} and Z_{MAX} have similar power performance in these applications. Moreover, Z_{MAX2} or Z_{MAX3} are much easier to compute than Z_{MAX}. Hence, in the next section, we only consider the two maximum tests Z_{MAX2} and Z_{MAX3}. The critical values for the maximum test statistics can be found by simulation under the null hypothesis as any two or three optimal statistics in $\{Z_i : i \in I\}$ follow multivariate normal distributions with correlation matrix $\{\rho_{ij}\}$. For example, given the data, ρ_{st} can be calculated. Generating a bivariate normal random variable (Z_{sj}, Z_{tj}) with the correlation matrix $\{p_{st}\}$ for $j = 1, \ldots, B$. For each j, Z_{MAX2} is obtained. Then an empirical distribution for Z_{MAX2} can be obtained using these B simulated maximum statistics, from which we can find the critical values. In some applications, if the null hypothesis does not depend on any nuisance parameters, the distribution of Z_{MAX2} or Z_{MAX3} can be simulated exactly without the correlation matrix, e.g., Zheng and Chen [24].

2.3. Comparison of MERT and MAX

Usually, Z_{MERT} is easier to compute and use than Z_{MAX2} (or Z_{MAX3}). Intuitively, however, Z_{MAX3} should have greater efficiency robustness than Z_{MERT} when the range of models is wide. The selection of the robust test depends on the minimum correlation ρ_{st} of the entire family of optimal tests. Results from Freidlin et al. [5] showed that when $\rho_{st} \geq 0.75$, MERT and MAX2 (MAX3) have similar power; thus, the simpler MERT can be used. For example, when $\rho_{st} = 0.75$, the ARE of MERT relative to the optimal test for any model in the family is at least 0.875. When $\rho_{st} < 0.50$, MAX2 (MAX3) is noticeably more powerful than the simple MERT. Hence, MAX2 (MAX3) is recommended. For example, in genetic linkage analysis using affected sib pairs, the minimum correlation is greater than 0.8, and the MERT, MAX2, and MAX3 have similar power. (Whittemore and Tu [22] and Gastwirth and Freidlin [10]) For analysis of case-parents data in genetic association studies where the mode of inheritance can range from pure recessive to pure dominant, the minimum correlation is less than 0.33, and then the MAX3 has desirable power robustness for this problem. (Zheng et al. [25])

3. Genetic association using case-control studies

3.1. Background

It is well known that association studies testing linkage disequilibrium are more powerful than linkage analysis to detect small genetic effects on traits. (Risch and Merikangas [17]) Moreover association studies using cases and controls are easier to conduct as parental genotypes are not required.

Assume that cases are sampled from the study population and that controls are independently sampled from the general population without disease. Cases and controls are not matched. Each individual is genotyped with one of three genotypes MM, MN and NN for a marker with two alleles M and N. The data obtained in case-control studies can be displayed as in Table 1 (genotype-based) or as in Table 2 (allele-based).

Define three penetrances as $f_0 = \Pr(\text{case}|NN)$, $f_1 = \Pr(\text{case}|NM)$, and $f_2 = \Pr(\text{case}|MM)$, which are the disease probabilities given different genotypes. The prevalence of disease is denoted as $D = \Pr(\text{case})$. The probabilities for genotypes (NN, NM, MM) in cases and controls are denoted by (p_0, p_1, p_2) and (q_0, q_1, q_2), respectively. The probabilities for genotypes (NN, NM, MM) in the general population are denoted as (g_0, g_1, g_2). The following relations can be obtained.

$$p_i = \frac{f_i g_i}{D} \text{ and } q_i = \frac{(1-f_i)g_i}{1-D} \text{ for } i = 0, 1, 2. \tag{3.1}$$

Note that, in Table 1, (r_0, r_1, r_2) and (s_0, s_1, s_2) follow multinomial distributions $mul(R; p_0, p_1, p_2)$ and $mul(S; q_0, q_1, q_2)$, respectively. Under the null hypothesis of no association between the disease and the marker, $p_i = q_i = g_i$ for $i = 0, 1, 2$. Hence, from (3.1), the null hypothesis for Table 1 is equivalent to $H_0 : f_0 = f_1 = f_2 = D$. Under the alternative, penetrances are different as one of two alleles is a risk allele, say, M. In genetic analysis, three genetic models (mode of inheritance) are often used. A model is recessive (rec) when $f_0 = f_1$, additive (add) when $f_1 = (f_0 + f_2)/2$, and dominant (dom) when $f_1 = f_2$. For recessive and dominant models, the number of columns in Table 1 can be reduced. Indeed, the columns with NN and NM (NM and MM) can be collapsed for recessive (dominant) model. Testing association using Table 2 is simpler but Sasieni [19] showed that genotype based analysis is preferable unless cases and controls are in Hardy–Weinberg Equilibrium.

TABLE 1
Genotype distribution for case-control studies

	NN	NM	MM	Total
Case	r_0	r_1	r_2	r
Control	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

TABLE 2
Allele distribution for case-control studies

	N	M	Total
Case	$2r_0 + r_1$	$r_1 + 2r_2$	$2r$
Control	$2s_0 + s_1$	$s_1 + 2s_2$	$2s$
Total	$2n_0 + n_1$	$n_1 + 2n_2$	$2n$

3.2. Test statistics

For the 2×3 table (Table 1), a chi-squared test with 2 degrees of freedom (df) can be used. (Gibson and Muse [11]) This test is independent of the underlying genetic model. Note that, under the alternative when M is the risk allele, the penetrances have a natural order: $f_0 \le f_1 \le f_2$ (at least one inequality hold). The Cochran-Armitage (CA) trend test (Cochran [3] and Armitage [1]) taking into account the natural order should be more powerful than the chi-squared test as the trend test has 1 df.

The CA trend test can be obtained as a score test under the logistic regression model with genotype as a covariate, which is coded using scores $\mathbf{x} = (x_0, x_1, x_1)$ for (NN, NM, MM), where $x_0 \le x_1 \le x_2$. The trend test can be written as (Sasieni [19])

$$Z_{\mathbf{x}} = \frac{n^{1/2} \sum_{i=0}^{2} x_i (sr_i - rs_i)}{\{rs[n \sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2]\}^{1/2}}.$$

Since the trend test is invariant to linear transformations of $\mathbf{x}$, without loss of generality, we use the scores $\mathbf{x} = (0, x, 1)$ with $0 \le x \le 1$ and denote $Z_{\mathbf{x}}$ as Z_x. Under the null hypothesis, Z_x has an asymptotic normal distribution $N(0, 1)$. When M is a risk allele, a one-sided test is used. Otherwise, a two-sided test should be used. Results from Sasieni [19] and Zheng, Freidlin, Li and Gastwirth [26] showed that the optimal choices of x for recessive, additive and dominant models are $x = 0$, $x = 1/2$, and $x = 1$, respectively. That is, Z_0, $Z_{1/2}$ or Z_1 is an asymptotically most powerful test when the genetic model is recessive, additive or dominant. The tests using other values of x are optimal for penetrances in the range $0 < f_0 \le f_1 \le f_2 < 1$.

For complex diseases, the genetic model is not known a priori. The optimal test Z_x cannot be used directly as a substantial loss of power may occur when x is misspecified. Applying the robust procedures introduced in Section 2, we have three genetic models and the collection of all consistent tests $C = \{Z_x : x \in [0, 1]\}$. To find a robust test, we need to evaluate the null correlations. Denote these as $\mathrm{corr}_{H_0}(Z_{x_1}, Z_{x_2}) = \rho_{x_1,x_2}$. From appendix C of Freidlin, Zheng, Li and Gastwirth [6],

$$\begin{aligned}
\rho_{0,1/2} &= \frac{p_0(p_1 + 2p_2)}{\{p_0(1-p_0)\}^{1/2}\{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2\}^{1/2}}, \\
\rho_{0,1} &= \frac{p_0 p_2}{\{p_0(1-p_0)\}^{1/2}\{p_2(1-p_2)\}^{1/2}}, \\
\rho_{1/2,1} &= \frac{p_2(p_1 + 2p_0)}{\{p_2(1-p_2)\}^{1/2}\{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2\}^{1/2}}.
\end{aligned}$$

Although the null correlations are functions of the unknown parameters $p_i, i = 0, 1, 2$, it can be shown analytically that $\rho_{0,1} < \rho_{0,1/2}$ and $\rho_{0,1} < \rho_{1/2,1}$. Note that if the above analytical results were not available, the p_i would be estimated by substituting the observed data $\hat{p}_i = n_i/n$ for p_i. Here the minimum correlation among the three optimal tests occurs when Z_0 and Z_1 is the extreme pair for the three genetic models. Freidlin et al. [6] also proved analytically that the condition (2.4) holds. Hence, $Z_{\mathrm{MERT}} = (Z_0 + Z_1)/\{2(1 + \hat{\rho}_{0,1})\}^{1/2}$ is the MERT for the whole family C, where $\hat{\rho}_{0,1}$ is obtained when the p_i are replaced by n_i/n. The two maximum tests can be written as $Z_{\mathrm{MAX2}} = \max(Z_0, Z_1)$ and $Z_{\mathrm{MAX3}} = \max(Z_0, Z_{1/2}, Z_1)$. When the risk allele is unknown, $Z_{\mathrm{MAX2}} = \max(|Z_0|, |Z_1|)$ and $Z_{\mathrm{MAX3}} = \max(|Z_0|, |Z_{1/2}|, |Z_1|)$. Although we considered three genetic models, the

family of genetic models for case-control studies can be extended by defining a genetic model as penetrances restricted to the family $\{(f_0, f_1, f_2) : f_0 \leq f_1 \leq f_2\}$. Three genetic models are contained in this family as the two boundaries and one middle ray of this family. The statistics Z_{MERT} and Z_{MAX3} are also the corresponding robust statistics for this larger family (see, e.g., Freidlin et al. [6] and Zheng et al. [26]).

In analysis of case-control data for genetic association, two other tests are also currently used. However, their robustness and efficiency properties have not been compared to MERT and MAX. The first one is the chi-squared test for the 2×3 contingency table (Table 1), denoted as χ^2_2. (Gibson and Muse [11]) Under the null hypothesis, it has a chi-squared distribution with 2 df. The second test, denoted as Z_{P}, is based on the product of two different tests: (a) the allele association (AA) test and (b) the Hardy-Weinberg disequilibrium (HWD) test. (Hoh, Wile and Ott [15] and Song and Elston [20]) The AA test is a chi-squares test for the 2×2 table given in Table 2, which is written as

$$\chi^2_{\mathrm{AA}} = \frac{2n[(2r_0 + r_1)(s_1 + 2s_2) - (2s_0 + s_1)(r_1 + 2r_2)]^2}{4rs(2n_0 + n_1)(n_1 + 2n_2)}.$$

The HWD test detects the deviation from Hardy–Weinberg equilibrium (HWE) in cases. Assume the allele frequency of M is $p = \Pr(M)$. Using cases, the estimation of p is $\hat{p} = (r_1 + 2r_2)/(2r)$. Let $\hat{q} = 1 - \hat{p}$ be the estimation of allele frequency for N. Under the null hypothesis of HWE, the expected number of genotypes can be written as $\mathrm{E}(NN) = r\hat{q}^2$, $\mathrm{E}(NM) = r2\hat{p}\hat{q}$ and $\mathrm{E}(MM) = r\hat{p}^2$, respectively. Hence, a chi-squared test for HWE is

$$\chi^2_{\mathrm{HWD}} = \frac{(r_0 - \mathrm{E}(NN))^2}{\mathrm{E}(NN)} + \frac{(r_1 - \mathrm{E}(NM))^2}{\mathrm{E}(NM)} + \frac{(r_2 - \mathrm{E}(MM))^2}{\mathrm{E}(MM)}.$$

The product test, proposed by Hoh et al. [15], is $T_{\mathrm{P}} = \chi^2_{\mathrm{AA}} \times \chi^2_{\mathrm{HWD}}$. They noticed that the power performances of these two statistics are complementary. Thus, the product should retain reasonable power as one of the tests has high power when the other does not. Consequently, for a comprehensive comparison, we also consider the maximum of them, $T_{\mathrm{MAX}} = \max(\chi^2_{\mathrm{AA}}, \chi^2_{\mathrm{HWD}})$. Given the data, the critical values of T_{P} and T_{MAX} can be obtained by a permutation procedure as their asymptotic distributions are not available. (Hoh et al. [15]) Note that T_{P} was originally proposed by Hoh et al. [15] as a test statistic for multiple gene selection and was modified by Song and Elston [20] for use as a test statistic for a single gene.

3.3. Power comparison

We conducted a simulation study to compare the power performance of the test statistics. The test statistics were (a) the optimal trend tests for the three genetic models, Z_0, $Z_{1/2}$ and Z_1, (b) MERT Z_{MERT}, (c) maximum tests Z_{MAX2} and Z_{MAX3}, (d) the product test T_{P}, (e) T_{MAX}, and (f) χ^2_2.

In the simulation a two-sided test was used. We assumed that the allele frequency p and the baseline penetrance f_0 are known ($f_0 = .01$). Note that, in practice, the allele frequency and penetrances are unknown. However, they can be estimated empirically (e.g., Song and Elston [20] and Wittke-Thompson, Pluzhnikov and Cox [23]). In our simulation the critical values for all test statistics are simulated under the null hypothesis. Thus, we avoid using asymptotic distributions for the test statistics. The

Type I errors for all tests are expected to be close to the nominal level $\alpha = 0.05$ and the powers of all tests are comparable. When HWE holds, the probabilities (p_0, p_1, p_2) for cases and (q_0, q_1, q_2) for controls can be calculated using (3.1) under the null and alternative hypotheses, where $(g_0, g_1, g_2) = (q^2, 2pq, q^2)$ and (f_1, f_2) are specified by the null or alternative hypotheses and $D = \sum f_i g_i$. After calculating (p_0, p_1, p_2) and (q_0, q_1, q_2) under the null hypothesis, we first simulated the genotype distributions $(r_0, r_1, r_2) \sim mul(R; p_0, p_1, p_2)$ and $(s_0, s_1, s_2) \sim mul(S; q_0, q_1, q_2)$ for cases and controls, respectively (see Table 1). When HWE does not hold, we assumed a mixture of two populations with two different allele frequencies p_1 and p_2. Hence, we simulated two independent samples with different allele frequencies for cases (and controls) and combined these two samples for cases (and for controls). Thus, cases (controls) contain samples from a mixture of two populations with different allele frequencies. When p is small, some counts can be zero. Therefore, we added 1/2 to the count of each genotype in cases and controls in all simulations.

To obtain the critical values, a simulation under the null hypothesis was done with 200,000 replicates. For each replicate, we calculated the test statistics. For each test statistic, we used its empirical distribution function based on 200,000 replicates to calculate the critical value for $\alpha = 0.05$. The alternatives were chosen so that the power of the optimal test $Z_0, Z_{1/2}, Z_1$ was near 80% for the recessive, additive, dominant models, respectively. To determine the empirical power, 10,000 replicates were simulated using multinomial distributions with the above probabilities.

To calculate Z_{MERT}, the correlation $\rho_{0,1}$ was estimated using the simulated data. In Table 3, we present the mean of correlation matrix using 10,000 replicates when $r = s = 250$. The three correlations $\rho_{0,1/2}, \rho_{0,1}, \rho_{1/2,1}$ were estimated by replacing p_i with $n_i/n, i = 0, 1, 2$ using the data simulated under the null and alternatives and various models. The null and alternative hypotheses used in Table 3 were also used to simulate critical values and powers (Table 4). Note that the minimum correlation $\rho_{0,1}$ is less than .50. Hence, the Z_{MAX3} should have greater efficiency robustness than Z_{MERT}. However, when the dominant model can be eliminated based on prior scientific knowledge (e.g. the disease often skips generations), the correlation between Z_0 and $Z_{1/2}$ optimal for the recessive and additive models would be greater than .75. Thus, for these two models, Z_{MERT} should have comparable power to $Z_{\text{MAX2}} = \max(|Z_0|, |Z_{1/2}|)$ and is easier to use.

The correlation matrices used in Table 4 for $r \neq s$ and Table 5 for mixed samples are not presented as they did not differ very much from those given in Table 3. Tables 4 and 5 present simulation results where all three genetic models are plausible. When HWE holds Table 4 shows that the Type I error is indeed close to the $\alpha = 0.05$ level. Since the model is not known, the minimum power across three genetic models is written in bold type. A test with the maximum of the minimum power among all test statistic has the most power robustness. Our comparison fo-

TABLE 3
The mean correlation matrices of three optimal test statistics based on 10,000 replicates when HWE holds

	p								
	.1			.3			.5		
Model	$\rho_{0,1/2}$	$\rho_{0,1}$	$\rho_{1/2,1}$	$\rho_{0,1/2}$	$\rho_{0,1}$	$\rho_{1/2,1}$	$\rho_{0,1/2}$	$\rho_{0,1}$	$\rho_{1/2,1}$
null	.97	.22	.45	.91	.31	.68	.82	.33	.82
rec	.95	.34	.63	.89	.36	.74	.81	.37	.84
add	.96	.23	.48	.89	.32	.71	.79	.33	.84
dom	.97	.21	.44	.90	.29	.69	.79	.30	.83

The same models (rec,add,dom) are used in Table 4 when $r = s = 250$.

TABLE 4
Power comparison when HWE holds in cases and controls under three genetic models with $\alpha = .05$

		Test statistics								
p	Model	Z_0	$Z_{1/2}$	Z_1	Z_{MERT}	Z_{MAX2}	Z_{MAX3}	T_{MAX}	T_{P}	χ^2_2
					$r = 250, s = 250$					
.1	null	.058	.048	.050	.048	.047	.049	.049	.049	.053
	rec	.813	**.364**	**.138**	**.606**	**.725**	**.732**	.941	.862	**.692**
	add	.223	.813	.802	.733	.782	.800	.705	**.424**	.752
	dom	**.108**	.796	.813	.635	.786	.795	**.676**	.556	.763
.3	null	.051	.052	.051	.054	.052	.052	.053	.049	.050
	rec	.793	**.537**	**.178**	.623	**.714**	**.726**	.833	.846	**.691**
	add	.433	.812	.768	.786	.742	.773	.735	**.447**	.733
	dom	**.133**	.717	.809	**.621**	.737	.746	**.722**	.750	.719
.5	null	.049	.047	.051	.047	.047	.047	.050	.051	.050
	rec	.810	.662	**.177**	.644	.738	.738	.772	.813	.709
	add	.575	.802	.684	.807	.729	.760	**.719**	**.450**	.714
	dom	**.131**	**.574**	.787	**.597**	**.714**	**.713**	.747	.802	**.698**
					$r = 50, s = 250$					
.1	null	.035	.052	.049	.052	.051	.053	.045	.051	.052
	rec	.826	**.553**	**.230**	.757	.797	.802	.779	.823	.803
	add	.250	.859	.842	.773	.784	.795	.718	**.423**	.789
	dom	**.114**	.807	.814	**.658**	**.730**	**.734**	**.636**	.447	**.727**
.3	null	.048	.048	.048	.048	.050	.050	.050	.048	.049
	rec	.836	**.616**	**.190**	.715	.787	.787	.733	.752	.771
	add	.507	.844	.794	.821	.786	.813	.789	**.500**	.778
	dom	**.171**	.728	.812	**.633**	**.746**	**.749**	**.696**	.684	**.729**
.5	null	.049	.046	.046	.046	.047	.047	.046	.049	.046
	rec	.838	.692	**.150**	.682	.771	.765	.705	.676	.748
	add	.615	.818	.697	.820	.744	.780	.743	**.493**	.728
	dom	**.151**	**.556**	.799	**.565**	**.710**	**.708**	**.662**	.746	**.684**

cuses on the test statistics: Z_{MAX3}, T_{MAX}, T_{P} and χ^2_2. Our results show that T_{MAX} has greater efficiency robustness than T_{P} while Z_{MAX3}, T_{MAX} and χ^2_2 have similar minimum powers. Notice that Z_{MAX3} is preferable to χ^2_2 although the difference in minimum powers depends on the allele frequency. When HWE does not hold, Z_{MAX3} still possesses its efficiency robustness, but T_{MAX} and T_{P} do not perform as well. Thus, population stratification affects their performance. Z_{MAX3} also remains more robust than χ^2_2 even when HWE does not hold. From both Table 4 and Table 5, χ^2_2 is more powerful than Z_{MERT} except for the additive model. However, when the genetic model is known, the corresponding optimal CA trend test is more powerful than χ^2_2 with 2 df.

From Tables 4 and 5, one sees that the robust test Z_{MAX3} tends to be more powerful than χ^2_2 under the various scenarios we simulated. Further comparisons of these two test statistics using p-values and the same data given in Table 4 (when $r = s = 250$) are reported in Table 6. Following Zheng et al. [25], which reported a matched study, where both tests are applied to the same data, the p-values for each test are grouped as $< .01$, $(.01, .05)$, $(.05, .10)$, and $> .10$. Cross classification of the p-values are given in Table 6 for allele frequencies $p = .1$ and .3 under all three genetic models. Table 6 is consistent with the results of Tables 4 and 5, i.e., Z_{MAX3} is more powerful than the chi-squared test with 2 degrees of freedom when the genetic model is unknown. This is seen by comparing the counts at the upper right corner with the counts at lower left corner. When the counts at the upper right corner are greater than the corresponding counts at the lower left corner, Z_{MAX3} usually has smaller p-values than χ^2_2. In particular, we compare two tests with p-values $< .01$ versus p-values in $(.01, .05)$. Notice that in most situations

TABLE 5

Power comparison when HWE does not hold in cases and controls under three genetic models (Mixed samples with different allele frequencies (p_1, p_2) and sample sizes (R_1, S_1) and (R_2, S_2) with $r = R_1 + R_2$, $s = S_1 + S_2$, and $\alpha = .05$).

		Test statistics								
(p_1, p_2)	Model	Z_0	$Z_{1/2}$	Z_1	Z_{MERT}	Z_{MAX2}	Z_{MAX3}	T_{MAX}	T_{P}	χ^2_2
		$R_1 = 250, S_1 = 250$ and $R_2 = 100, S_2 = 100$								
(.1,.4)	null	.047	.049	.046	.050	.047	.046	.048	.046	.048
	rec	.805	**.519**	**.135**	.641	.747	.744	.936	.868	.723
	add	.361	.817	.771	.776	.737	.757	.122	.490	.688
	dom	**.098**	.715	.805	**.586**	**.723**	**.724**	**.049**	**.097**	**.681**
(.1,.5)	null	.047	.050	.046	.048	.046	.046	.052	.050	.052
	rec	.797	**.537**	**.121**	.620	.746	.750	.920	.839	.724
	add	.383	.794	.732	.764	**.681**	.709	.033	.620	**.631**
	dom	**.095**	.695	.812	**.581**	.697	**.703**	**.001**	**.133**	.670
(.2,.5)	null	.048	.052	.052	.054	.052	.052	.050	.047	.050
	rec	.816	**.576**	**.157**	.647	.760	.749	.889	.881	.725
	add	.417	.802	.754	.782	**.726**	.743	.265	.365	**.684**
	dom	**.112**	.679	.812	**.600**	.729	**.715**	**.137**	**.122**	.696
		$R_1 = 30, S_1 = 150$ and $R_2 = 20, S_2 = 100$								
(.1,.4)	null	.046	.048	.048	.046	.048	.046	.055	.045	.048
	rec	.847	**.603**	**.163**	.720	.807	.799	.768	.843	.779
	add	.387	.798	.762	.749	.725	.746	.449	.309	**.691**
	dom	**.139**	.733	.810	**.603**	**.721**	**.728**	**.345**	**.223**	.699
(.1,.5)	null	.053	.055	.050	.053	.053	.053	.059	.047	.054
	rec	.816	**.603**	**.139**	.688	.776	.780	.697	.827	.741
	add	.415	.839	.797	.798	.763	.781	.276	.358	**.703**
	dom	**.120**	.726	.845	**.612**	**.750**	**.752**	**.149**	**.133**	.716
(.2,.5)	null	.047	.048	.050	.051	.048	.048	.049	.050	.044
	rec	.858	**.647**	**.167**	.722	.808	.804	.768	.852	.783
	add	.472	.839	.790	.811	.776	.799	.625	**.325**	.740
	dom	**.139**	.708	.815	**.614**	**.741**	**.743**	**.442**	.390	**.708**

the number of times χ^2_2 has a p-value $< .01$ and Z_{MAX3} has a p-value in $(.01, .05)$ is much less than the corresponding number of times when Z_{MAX3} has a p-value $< .01$ and χ^2_2 has a p-value in $(.01, .05)$. For example, when $p = .3$ and the additive model holds, there are 289 simulated datasets where χ^2_2 has a p-value in (.01,.05) while Z_{MAX3} has a p-value $< .01$ versus only 14 such datasets when Z_{MAX3} has a p-value in (.01,.05) while χ^2_2 has a p-value $< .01$. The only exception occurs at the recessive model under which they have similar counts (165 vs. 140). Combining results from Tables 4 and 6, Z_{MAX3} is more powerful than χ^2_2, but the difference of power between Z_{MAX3} and χ^2_2 is usually less than 5% in the simulation. Hence χ^2_2 is also an efficiency robust test, which is very useful for genome-wide association studies, where hundreds of thousands of tests are performed.

From prior studies of family pedigrees one may know whether the disease skips generations or not. If it does, the disease is less likely to follow a pure-dominant model. Thus, when genetic evidence strongly suggests that the underlying genetic model is between the recessive and additive inclusive, we compared the performance of tests $Z_{\text{MERT}} = (Z_0 + Z_{1/2})/\{2(1 + \hat{\rho}_{0,1/2})\}^{1/2}$, $Z_{\text{MAX2}} = \max(Z_0, Z_{1/2})$, and χ^2_2. The results are presented in Table 7. The alternatives used in Table 4 for rec and add with $r = s = 250$ were also used to obtain Table 7. For a family with the recessive and additive models, the minimum correlation is increased compared to the family with three genetic models (rec, add and dom). For example, from Table 3, the minimum correlation with the family of three models that ranges from .21 to .37 is increased to the range of .79 to .97 with only two models. From Table 7, under the recessive and additive models, while Z_{MAX2} remains more powerful than

TABLE 6
Matched p-value comparison of Z_{MAX3} and χ^2_2 when HWE holds in cases and controls under three genetic models (Sample sizes $r = s = 250$ and 5,000 replications)

			χ^2_2			
p	Model	Z_{MAX3}	$< .01$	$.01 - .05$	$.05 - .10$	$> .10$
.10	rec	$< .01$	2069	165	0	0
		$.01 - .05$	140	1008	251	0
		$.05 - .10$	7	52	227	203
		$> .10$	1	25	47	805
	add	$< .01$	2658	295	0	0
		$.01 - .05$	44	776	212	0
		$.05 - .10$	3	23	159	198
		$> .10$	0	5	16	611
	dom	$< .01$	2785	214	0	0
		$.01 - .05$	80	712	214	0
		$.05 - .10$	10	42	130	169
		$> .10$	1	14	27	602
.30	rec	$< .01$	2159	220	0	0
		$.01 - .05$	85	880	211	0
		$.05 - .10$	6	44	260	157
		$> .10$	2	33	40	903
	add	$< .01$	2485	289	0	0
		$.01 - .05$	14	849	229	0
		$.05 - .10$	1	8	212	215
		$> .10$	0	1	6	691
	dom	$< .01$	2291	226	0	0
		$.01 - .05$	90	894	204	0
		$.05 - .10$	7	52	235	160
		$> .10$	0	26	49	766

TABLE 7
Power comparison when HWE holds in cases and controls assuming two genetic models (rec and add) based on 10,000 replicates ($r = s = 250$ and $f_0 = .01$)

	p								
	.1			.3			.5		
Model	Z_{MERT}	Z_{MAX2}	χ^2_2	Z_{MERT}	Z_{MAX2}	χ^2_2	Z_{MERT}	Z_{MAX2}	χ^2_2
null	.046	.042	.048	.052	.052	.052	.053	.053	.053
rec	.738	.729	.703	.743	.732	.687	.768	.766	.702
add	.681	.778	.778	.714	.755	.726	.752	.764	.728
other[1]	.617	.830	.836	.637	.844	.901	.378	.547	.734
other[2]	.513	.675	.677	.632	.774	.803	.489	.581	.652
other[3]	.385	.465	.455	.616	.672	.655	.617	.641	.606

other[1] is dominant and the other two are semi-dominant, all with $f_2 = .019$.
other [1]: $f_1 = .019$. other[2]: $f_1 = .017$. other[3]: $f_1 = .015$.

χ^2_2 and Z_{MERT}, the difference in minimum power is much less than in the previous simulation study (Table 4). Indeed, when studying complex common diseases where the allele frequency is thought to be fairly high, Z_{MAX2} and Z_{MERT} have similar power. Thus, when a genetic model is between the recessive and additive models inclusive, MAX2 and MERT should be used. In Table 7, some other models were also included in simulations when we do not have sound genetic knowledge to eliminate the dominant model. In this case, MAX2 and MERT lose some efficiency compared to χ^2_2. However, MAX3 still has greater efficiency robustness than other tests. In particular, MAX3 is more powerful than χ^2_2 (not reported) as in Table 4. Thus, MAX3 should be used when prior genetic studies do not justify excluding one of the basic three models.

4. Discussion

In this article, we review robust procedures for testing hypothesis when the underlying model is unknown. The implementation of these robust procedures is illustrated by applying them to testing genetic association in case-control studies. Simulation studies demonstrated the usefulness of these robust procedures when the underlying genetic model is unknown.

When the genetic model is known (e.g., recessive, dominant or additive model), the optimal Cochran-Armitage trend test with the appropriate choice of x is more powerful than the chi-squared test with 2 df for testing an association. The genetic model is usually not known for complex diseases. In this situation, the maximum of three optimal tests (including the two extreme tests), Z_{MAX3}, is shown to be efficient robust compared to other available tests. In particular, Z_{MAX3} is slightly more powerful than the chi-squared test with 2 df. Based on prior scientific knowledge, if the dominant model can be eliminated, then MERT, the maximum test, and the chi-squared test have roughly comparable power for a genetic model that ranges from recessive model to additive model and the allele frequency is not small. In this situation, the MERT and the chi-squared test are easier to apply than the maximum test and can be used by researchers. Otherwise, with current computational tools, Z_{MAX3} is recommended.

Acknowledgements

It is a pleasure to thank Prof. Javier Rojo for inviting us to participate in this important conference, in honor of Prof. Lehmann, and the members of the Department of Statistics at Rice University for their kind hospitality during it. We would also like to thank two referees for their useful comments and suggestions which improved our presentation.

References

[1] Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

[2] Birnbaum, A. and Laska, E. (1967). Optimal robustness: a general method, with applications to linear estimators of location. *J. Am. Statist. Assoc.* **62**, 1230–1240.

[3] Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics* **10**, 417–451.

[4] Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

[5] Freidlin, B., Podgor, M. J. and Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics* **55**, 883–886.

[6] Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53**, 146–152.

[7] Gastwirth, J. L. (1966). On robust procedures. *J. Am. Statist. Assoc.* **61**, 929–948.

[8] Gastwirth, J. L. (1970). On robust rank tests. In *Nonparametric Techniques in Statistical Inference*. Ed. M. L. Puri. Cambridge University Press, London.

[9] Gastwirth, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Am. Statist. Assoc.* **80**, 380–384.

[10] Gastwirth, J. L. and Freidlin, B. (2000). On power and efficiency robust linkage tests for affected sibs. *Ann. Hum. Genet.* **64**, 443–453.

[11] Gibson, G. and Muse, S. (2001). *A Primer of Genome Science*. Sinnauer, Sunderland, MA.

[12] Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**, 471–476.

[13] Gross, S. T. (1981). On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *J. Am. Statist. Assoc.* **76**, 935–941.

[14] Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.

[15] Hoh, J., Wile, A. and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* **11**, 269–293.

[16] Podgor, M. J., Gastwirth, J. L. and Mehta, C. R. (1996). Efficiency robust tests of independence in contingency tables with ordered classifications. *Statist. Med.* **15**, 2095–2105.

[17] Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

[18] Rosen, J. B. (1960). The gradient projection method for non-linear programming. Part I: Linear constraints. *SIAM J.* **8**, 181–217.

[19] Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

[20] Song, K. and Elston. R. C. (2006). A powerful method of combining measures of association and Hardy–Weinberg disequilibrium for fine-mapping in case-control studies. *Statist. Med.* **25**, 105–126.

[21] van Eeden, C. (1964). The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *Ann. Math. Statist.* **34**, 1442–1451.

[22] Whittemore, A. S. and Tu, I.-P. (1998). Simple, robust linkage tests for affected sibs. *Am. J. Hum. Genet.* **62**, 1228–1242.

[23] Wittke-Thompson, J. K., Pluzhnikov A. and Cox, N. J. (2005). Rational inference about departures from Hardy–Weinberg Equilibrium. *Am. J. Hum. Genet.* **76**, 967–986.

[24] Zheng, G. and Chen, Z. (2005). Comparison of maximum statistics for hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrics* **61**, 254–258.

[25] Zheng, G., Freidlin, B. and Gastwirth, J. L. (2002). Robust TDT-type candidate-gene association tests. *Ann. Hum. Genet.* **66**, 145–155.

[26] Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical J.* **45**, 335–348.

[27] Zucker, D. M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77**, 853–864.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 266–290
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000509

Optimal sampling strategies for multiscale stochastic processes

Vinay J. Ribeiro[1], Rudolf H. Riedi[2] and Richard G. Baraniuk[1,*]

Rice University

Abstract: In this paper, we determine which non-random sampling of fixed size gives the best linear predictor of the sum of a finite spatial population. We employ different multiscale superpopulation models and use the minimum mean-squared error as our optimality criterion. In multiscale superpopulation tree models, the leaves represent the units of the population, interior nodes represent partial sums of the population, and the root node represents the total sum of the population. We prove that the optimal sampling pattern varies dramatically with the correlation structure of the tree nodes. While *uniform sampling* is optimal for trees with "positive correlation progression", it provides the worst possible sampling with "negative correlation progression." As an analysis tool, we introduce and study a class of *independent innovations trees* that are of interest in their own right. We derive a fast water-filling algorithm to determine the optimal sampling of the leaves to estimate the root of an independent innovations tree.

1. Introduction

In this paper we design optimal sampling strategies for spatial populations under different multiscale superpopulation models. Spatial sampling plays an important role in a number of disciplines, including geology, ecology, and environmental science. See, e.g., Cressie [5].

1.1. Optimal spatial sampling

Consider a finite population consisting of a rectangular grid of $R \times C$ units as depicted in Fig. 1(a). Associated with the unit in the i^{th} row and j^{th} column is an unknown value $\ell_{i,j}$. We treat the $\ell_{i,j}$'s as one realization of a superpopulation model.

Our goal is to determine which sample, among all samples of size n, gives the best linear estimator of the population sum, $S := \sum_{i,j} \ell_{i,j}$. We abbreviate *variance*, *covariance*, and *expectation* by "var", "cov", and "$\mathbb{E}$" respectively. Without loss of generality we assume that $\mathbb{E}(\ell_{i,j}) = 0$ for all locations (i, j).

[1]Department of Statistics, 6100 Main Street, MS-138, Rice University, Houston, TX 77005, e-mail: vinay@rice.edu; riedi@rice.edu

[2]Department of Electrical and Computer Engineering, 6100 Main Street, MS-380, Rice University, Houston, TX 77005, e-mail: richb@rice.edu, url: dsp.rice.edu, spin.rice.edu

*Supported by NSF Grants ANI-9979465, ANI-0099148, and ANI-0338856, DoE SciDAC Grant DE-FC02-01ER25462, DARPA/AFRL Grant F30602-00-2-0557, Texas ATP Grant 003604-0036-2003, and the Texas Instruments Leadership University program.

AMS 2000 subject classifications: primary 94A20, 62M30, 60G18; secondary 62H11, 62H12, 78M50.

Keywords and phrases: multiscale stochastic processes, finite population, spatial data, networks, sampling, convex, concave, optimization, trees, sensor networks.

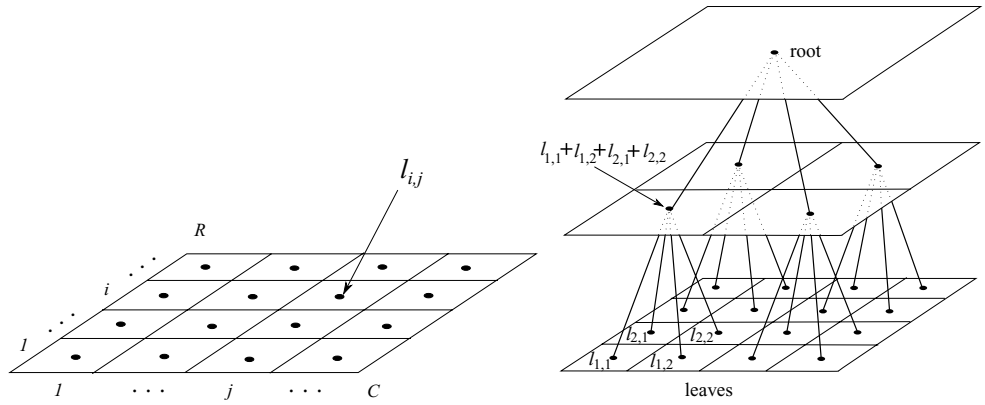


FIG 1. (a) *Finite population on a spatial rectangular grid of size $R \times C$ units. Associated with the unit at position (i, j) is an unknown value $\ell_{i,j}$.* (b) *Multiscale superpopulation model for a finite population. Nodes at the bottom are called leaves and the topmost node the root. Each leaf node corresponds to one value $\ell_{i,j}$. All nodes, except for the leaves, correspond to the sum of their children at the next lower level.*

Denote an arbitrary sample of size n by L. We consider linear estimators of S that take the form

$$\widehat{S}(L, \boldsymbol{\alpha}) := \boldsymbol{\alpha}^T L, \tag{1.1}$$

where $\boldsymbol{\alpha}$ is an arbitrary set of coefficients. We measure the accuracy of $\widehat{S}(L, \boldsymbol{\alpha})$ in terms of the *mean-squared error* (MSE)

$$\mathcal{E}(S|L, \boldsymbol{\alpha}) := \mathbb{E}\left(S - \widehat{S}(L, \boldsymbol{\alpha})\right)^2 \tag{1.2}$$

and define the *linear minimum mean-squared error* (LMMSE) of estimating S from L as

$$\mathcal{E}(S|L) := \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \mathcal{E}(S|L, \boldsymbol{\alpha}). \tag{1.3}$$

Restated, our goal is to determine

$$L^* := \arg\min_{L} \mathcal{E}(S|L). \tag{1.4}$$

Our results are particularly applicable to Gaussian processes for which linear estimation is optimal in terms of mean-squared error. We note that for certain multimodal and discrete processes linear estimation may be sub-optimal.

1.2. Multiscale superpopulation models

We assume that the population is one realization of a multiscale stochastic process (see Fig. 1(b)) (see Willsky [20]). Such processes consist of random variables organized on a tree. Nodes at the bottom, called *leaves*, correspond to the population $\ell_{i,j}$. All nodes, except for the leaves, represent the sum total of their children at the next lower level. The topmost node, the *root*, hence represents the sum of the entire population. The problem we address in this paper is thus equivalent to the following: *Among all possible sets of leaves of size n, which set provides the best linear estimator for the root in terms of MSE?*

Multiscale stochastic processes efficiently capture the correlation structure of a wide range of phenomena, from uncorrelated data to complex *fractal* data. They

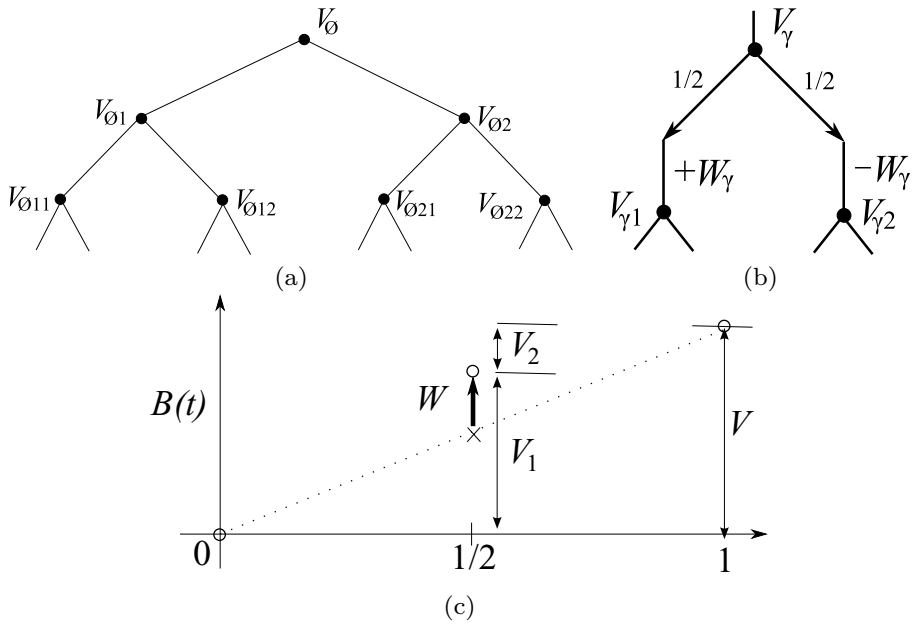


FIG 2. (a) *Binary tree for interpolation of Brownian motion, $B(t)$.* (b) *Form child nodes $V_{\gamma 1}$ and $V_{\gamma 2}$ by adding and subtracting an independent Gaussian random variable W_γ from $V_\gamma/2$. (c) Mid-point displacement. Set $B(1) = V_\emptyset$ and form $B(1/2) = (B(1) - B(0))/2 + W_\emptyset = V_{\emptyset 1}$. Then $B(1) - B(1/2) = V_\emptyset/2 - W_\emptyset = V_{\emptyset 2}$. In general a node at scale j and position k from the left of the tree corresponds to $B((k+1)2^{-j}) - B(k2^{-j})$.*

do so through a simple probabilistic relationship between each parent node and its children. They also provide fast algorithms for analysis and synthesis of data and are often physically motivated. As a result multiscale processes have been used in a number of fields, including oceanography, hydrology, imaging, physics, computer networks, and sensor networks (see Willsky [20] and references therein, Riedi et al. [15], and Willett et al. [19]).

We illustrate the essentials of multiscale modeling through a tree-based interpolation of one-dimensional *standard Brownian motion*. Brownian motion, $B(t)$, is a zero-mean Gaussian process with $B(0) := 0$ and $\mathrm{var}(B(t)) = t$. Our goal is to begin with $B(t)$ specified only at $t = 1$ and then interpolate it at all time instants $t = k2^{-j}$, $k = 1, 2, \ldots, 2^j$ for any given value j.

Consider a binary tree as shown in Fig. 2(a). We denote the root by $V_\emptyset$. Each node V_γ is the *parent* of two nodes connected to it at the next lower level, $V_{\gamma 1}$ and $V_{\gamma 2}$, which are called its *child* nodes. The address γ of any node V_γ is thus a concatenation of the form $\emptyset k_1 k_2 \ldots k_j$, where j is the node's *scale* or depth in the tree.

We begin by generating a zero-mean Gaussian random variable with unit variance and assign this value to the root, $V_\emptyset$. The root is now a realization of $B(1)$. We next interpolate $B(0)$ and $B(1)$ to obtain $B(1/2)$ using a "mid-point displacement" technique. We generate independent *innovation* $W_\emptyset$ of variance $\mathrm{var}(W_\emptyset) = 1/4$ and set $B(1/2) = V_\emptyset/2 + W_\emptyset$ (see Fig. 2(c)).

Random variables of the form $B((k+1)2^{-j}) - B(k2^{-j})$ are called *increments* of Brownian motion at time-scale j. We assign the increments of the Brownian motion at time-scale 1 to the children of $V_\emptyset$. That is, we set

$$\begin{aligned} V_{\emptyset 1} &= B(1/2) - B(0) = V_\emptyset/2 + W_\emptyset, \text{ and} \\ V_{\emptyset 2} &= B(1) - B(1/2) = V_\emptyset/2 - W_\emptyset \end{aligned} \tag{1.5}$$

as depicted in Fig. 2(c). We continue the interpolation by repeating the procedure described above, replacing $V_{\emptyset}$ by each of its children and reducing the variance of the innovations by half, to obtain $V_{\emptyset 11}$, $V_{\emptyset 12}$, $V_{\emptyset 21}$, and $V_{\emptyset 22}$.

Proceeding in this fashion we go down the tree assigning values to the different tree nodes (see Fig. 2(b)). It is easily shown that the nodes at scale j are now realizations of $B((k+1)2^{-j}) - B(k2^{-j})$. That is, increments at time-scale j. For a given value of j we thus obtain the interpolated values of Brownian motion, $B(k2^{-j})$ for $k = 0, 1, \ldots, 2^j - 1$, by cumulatively summing up the nodes at scale j.

By appropriately setting the variances of the innovations W_γ, we can use the procedure outlined above for Brownian motion interpolation to interpolate several other Gaussian processes (Abry et al. [1], Ma and Ji [12]). One of these is *fractional Brownian motion* (fBm), $B_H(t)$ $(0 < H < 1))$, that has variance $\mathrm{var}(B_H(t)) = t^{2H}$. The parameter H is called the *Hurst* parameter. Unlike the interpolation for Brownian motion which is exact, however, the interpolation for fBm is only approximate. By setting the variance of innovations at different scales appropriately we ensure that nodes at scale j have the same variance as the increments of fBm at time-scale j. However, except for the special case when $H = 1/2$, the covariance between any two arbitrary nodes at scale j is not always identical to the covariance of the corresponding increments of fBm at time-scale j. Thus the tree-based interpolation captures the variance of the increments of fBm at all time-scales j but does not perfectly capture the entire covariance (second-order) structure.

This approximate interpolation of fBm, nevertheless, suffices for several applications including network traffic synthesis and queuing experiments (Ma and Ji [12]). They provide fast $O(N)$ algorithms for both synthesis and analysis of data sets of size N. By assigning multivariate random variables to the tree nodes V_γ as well as innovations W_γ, the accuracy of the interpolations for fBm can be further improved (Willsky [20]).

In this paper we restrict our attention to two types of multiscale stochastic processes: *covariance trees* (Ma and Ji [12], Riedi et al. [15]) and *independent innovations trees* (Chou et al. [3], Willsky [20]). In covariance trees the covariance between pairs of leaves is purely a function of their distance. In independent innovations trees, each node is related to its parent nodes through a unique independent additive innovation. One example of a covariance tree is the multiscale process described above for the interpolation of Brownian motion (see Fig. 2).

1.3. Summary of results and paper organization

We analyze covariance trees belonging to two broad classes: those with *positive correlation progression* and those with *negative correlation progression*. In trees with positive correlation progression, leaves closer together are more correlated than leaves father apart. The opposite is true for trees with negative correlation progression. While most spatial data sets are better modeled by trees with positive correlation progression, there exist several phenomena in finance, computer networks, and nature that exhibit anti-persistent behavior, which is better modeled by a tree with negative correlation progression (Li and Mills [11], Kuchment and Gelfan [9], Jamdee and Los [8]).

For covariance trees with positive correlation progression we prove that uniformly spaced leaves are optimal and that clustered leaf nodes provides the worst possible MSE among all samples of fixed size. The optimal solution can, however, change with the correlation structure of the tree. In fact for covariance trees with negative

correlation progression we prove that uniformly spaced leaf nodes give the *worst* possible MSE!

In order to prove optimality results for covariance trees we investigate the closely related independent innovations trees. In these trees, a parent node cannot equal the sum of its children. As a result they cannot be used as superpopulation models in the scenario described in Section 1.1. Independent innovations trees are however of interest in their own right. For independent innovations trees we describe an efficient algorithm to determine an optimal leaf set of size n called *water-filling*. Note that the general problem of determining which n random variables from a given set provide the best linear estimate of another random variable that is not in the same set is an NP-hard problem. In contrast, the water-filling algorithm solves one problem of this type in polynomial-time.

The paper is organized as follows. Section 2 describes various multiscale stochastic processes used in the paper. In Section 3 we describe the water-filling technique to obtain optimal solutions for independent innovations trees. We then prove optimal and worst case solutions for covariance trees in Section 4. Through numerical experiments in Section 5 we demonstrate that optimal solutions for multiscale processes can vary depending on their topology and correlation structure. We describe related work on optimal sampling in Section 6. We summarize the paper and discuss future work in Section 7. The proofs can be found in the Appendix. The pseudo-code and analysis of the computational complexity of the water-filling algorithm are available online (Ribeiro et al. [14]).

2. Multiscale stochastic processes

Trees occur naturally in many applications as an efficient data structure with a simple dependence structure. Of particular interest are trees which arise from representing and analyzing stochastic processes and time series on different time scales. In this section we describe various trees and related background material relevant to this paper.

2.1. Terminology and notation

A tree is a special graph, i.e., a set of nodes together with a list of pairs of nodes which can be pictured as directed edges pointing from one node to another with the following special properties (see Fig. 3): (1) There is a unique node called the *root* to which no edge points to. (2) There is exactly one edge pointing to any node, with the exception of the root. The starting node of the edge is called the *parent* of the ending node. The ending node is called a *child* of its parent. (3) The tree is *connected*, meaning that it is possible to reach any node from the root by following edges.

These simple rules imply that there are no cycles in the tree, in particular, there is exactly one way to reach a node from the root. Consequently, unique addresses can be assigned to the nodes which also reflect the level of a node in the tree. The topmost node is the root whose address we denote by $\emptyset$. Given an arbitrary node γ, its child nodes are said to be one level lower in the tree and are addressed by γk ($k = 1, 2, \ldots, P_\gamma$), where $P_\gamma \geq 0$. The address of each node is thus a concatenation of the form $\emptyset k_1 k_2 \ldots k_j$, or $k_1 k_2 \ldots k_j$ for short, where j is the node's *scale* or depth in the tree. The largest scale of any node in the tree is called the *depth* of the tree.

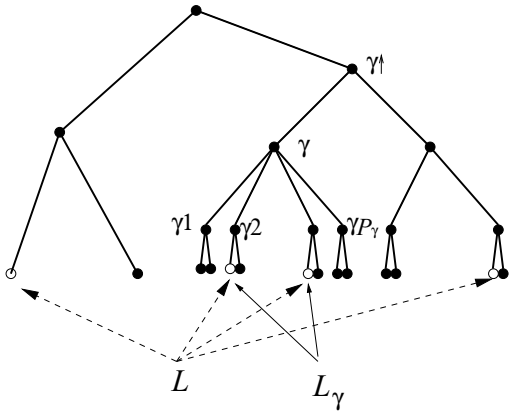


FIG 3. *Notation for multiscale stochastic processes.*

Nodes with no child nodes are termed *leaves* or *leaf nodes*. As usual, we denote the number of elements of a set of leaf nodes L by $|L|$. We define the operator $\uparrow$ such that $\gamma k \uparrow = \gamma$. Thus, the operator $\uparrow$ takes us one level higher in the tree to the parent of the current node. Nodes that can be reached from γ by repeated $\uparrow$ operations are called *ancestors* of γ. We term γ a *descendant* of all of its ancestors.

The set of nodes and edges formed by γ and all its descendants is termed the *tree of* γ. Clearly, it satisfies all rules of a tree. Let L_γ denote the subset of L that belong to the tree of γ. Let $\mathcal{N}_\gamma$ be the total number of leaves of the tree of γ.

To every node γ we associate a single (univariate) random variable V_γ. For the sake of brevity we often refer to V_γ as simply "the node V_γ" rather than "the random variable associated with node γ."

2.2. Covariance trees

Covariance trees are multiscale stochastic processes defined on the basis of the covariance between the leaf nodes which is purely a function of their *proximity*. Examples of covariance trees are the Wavelet-domain Independent Gaussian model (WIG) and the Multifractal Wavelet Model (MWM) proposed for network traffic (Ma and Ji [12], Riedi et al. [15]). Precise definitions follow.

Definition 2.1. The proximity of two leaf nodes is the scale of their lowest common ancestor.

Note that the larger the proximity of a pair of leaf nodes, the closer the nodes are to each other in the tree.

Definition 2.2. A covariance tree is a multiscale stochastic process with two properties. (1) The covariance of any two leaf nodes depends only on their proximity. In other words, if the leaves γ' and γ have proximity k then $\mathrm{cov}(V_\gamma, V_{\gamma'}) =: c_k$. (2) All leaf nodes are at the same scale D and the root is equally correlated with all leaves.

In this paper we consider covariance trees of two classes: trees with positive correlation progression and trees with negative correlation progression.

Definition 2.3. A covariance tree has a positive correlation progression if $c_k > c_{k-1} > 0$ for $k = 1, \ldots, D-1$. A covariance tree has a *negative correlation progression if* $c_k < c_{k-1}$ *for* $k = 1, \ldots, D-1$.

Intuitively in trees with positive correlation progression leaf nodes "closer" to each other in the tree are more strongly correlated than leaf nodes "farther apart." Our results take on a special form for covariance trees that are also symmetric trees.

Definition 2.4. A symmetric tree is a multiscale stochastic process in which P_γ, the number of child nodes of V_γ, is purely a function of the scale of γ.

2.3. Independent innovations trees

Independent innovations trees are particular multiscale stochastic processes defined as follows.

Definition 2.5. An independent innovations tree is a multiscale stochastic process in which each node V_γ, excluding the root, is defined through

$$V_\gamma := \varrho_\gamma V_{\gamma\uparrow} + W_\gamma. \tag{2.1}$$

Here, ϱ_γ is a scalar and W_γ is a random variable independent of $V_{\gamma\uparrow}$ as well as of $W_{\gamma'}$ for all $\gamma' \neq \gamma$. The root, $V_\emptyset$, is independent of W_γ for all γ. In addition $\varrho_\gamma \neq 0$, $\mathrm{var}(W_\gamma) > 0\ \forall\gamma$ and $\mathrm{var}(V_\emptyset) > 0$.

Note that the above definition guarantees that $\mathrm{var}(V_\gamma) > 0\ \forall\gamma$ as well as the *linear independence*[1] of any set of tree nodes.

The fact that each node is the sum of a scaled version of its parent and an independent random variable makes these trees amenable to analysis (Chou et al. [3], Willsky [20]). We prove optimality results for independent innovations trees in Section 3. Our results take on a special form for scale-invariant trees defined below.

Definition 2.6. A scale-invariant tree is an independent innovations tree which is symmetric and where ϱ_γ and the distribution of W_γ are purely functions of the scale of γ.

While independent innovations trees are not covariance trees in general, it is easy to see that scale-invariant trees are indeed covariance trees with positive correlation progression.

3. Optimal leaf sets for independent innovations trees

In this section we determine the optimal leaf sets of independent innovations trees to estimate the root. We first describe the concept of water-filling which we later use to prove optimality results. We also outline an efficient numerical method to obtain the optimal solutions.

3.1. Water-filling

While obtaining optimal sets of leaves to estimate the root we maximize a sum of concave functions under certain constraints. We now develop the tools to solve this problem.

[1] A set of random variables is linearly independent if none of them can be written as a linear combination of finitely many other random variables in the set.

Definition 3.1. A real function ψ defined on the set of integers $\{0, 1, \ldots, M\}$ is discrete-concave if

$$\psi(x+1) - \psi(x) \geq \psi(x+2) - \psi(x+1), \text{ for } x = 0, 1, \ldots, M-2. \tag{3.1}$$

The optimization problem we are faced with can be cast as follows. Given integers $P \geq 2$, $M_k > 0$ $(k = 1, \ldots, P)$ and $n \leq \sum_{k=1}^{P} M_k$ consider the discrete space

$$\Delta_n(M_1, \ldots, M_P) := \left\{ X = [x_k]_{k=1}^{P} : \sum_{k=1}^{P} x_k = n; x_k \in \{0, 1, \ldots, M_k\}, \forall k \right\}. \tag{3.2}$$

Given non-decreasing, discrete-concave functions ψ_k $(k = 1, \ldots, P)$ with domains $\{0, \ldots, M_k\}$ we are interested in

$$h(n) := \max \left\{ \sum_{k=1}^{P} \psi_k(x_k) \, : \, X \in \Delta_n(M_1, \ldots, M_P) \right\}. \tag{3.3}$$

In the context of optimal estimation on a tree, P will play the role of the number of children that a parent node V_γ has, M_k the total number of leaf node descendants of the k-th child $V_{\gamma k}$, and ψ_k the reciprocal of the optimal LMMSE of estimating V_γ given x_k leaf nodes in the tree of $V_{\gamma k}$. The quantity $h(n)$ corresponds to the reciprocal of the optimal LMMSE of estimating node V_γ given n leaf nodes in its tree.

The following iterative procedure solves the optimization problem (3.3). Form vectors $G^{(n)} = [g_k^{(n)}]_{k=1}^{P}$, $n = 0, \ldots, \sum_k M_k$ as follows:
Step (i): Set $g_k^{(0)} = 0$, $\forall k$.
Step (ii): Set

$$g_k^{(n+1)} = \begin{cases} g_k^{(n)} + 1, & k = m \\ g_k^{(n)}, & k \neq m \end{cases} \tag{3.4}$$

where

$$m \in \arg\max_k \left\{ \psi_k \left(g_k^{(n)} + 1 \right) - \psi_k \left(g_k^{(n)} \right) : g_k^{(n)} < M_k \right\}. \tag{3.5}$$

The procedure described in Steps (i) and (ii) is termed *water-filling* because it resembles the solution to the problem of filling buckets with water to maximize the sum of the heights of the water levels. These buckets are narrow at the bottom and monotonically widen towards the top. Initially all buckets are empty (compare Step (i)). At each step we are allowed to pour one unit of water into any one bucket with the goal of maximizing the sum of water levels. Intuitively at any step we must pour the water into that bucket which will give the maximum increase in water level among all the buckets not yet full (compare Step (ii)). Variants of this water-filling procedure appear as solutions to different information theoretic and communication problems (Cover and Thomas [4]).

Lemma 3.1. *The function $h(n)$ is non-decreasing and discrete-concave. In addition,*

$$h(n) = \sum_k \psi_k \left(g_k^{(n)} \right), \tag{3.6}$$

where $g_k^{(n)}$ is defined through water-filling.

When all functions ψ_k in Lemma 3.1 are identical, the maximum of $\sum_{k=1}^{P} \psi_k(x_k)$ is achieved by choosing the x_k's to be "near-equal". The following Corollary states this rigorously.

Corollary 3.1. *If $\psi_k = \psi$ for all $k = 1, 2, \ldots, P$ with ψ non-decreasing and discrete-concave, then*

$$h(n) = \left(P - n + P\left\lfloor \frac{n}{P} \right\rfloor\right) \psi\left(\left\lfloor \frac{n}{P} \right\rfloor\right) + \left(n - P\left\lfloor \frac{n}{P} \right\rfloor\right) \psi\left(\left\lfloor \frac{n}{P} \right\rfloor + 1\right). \tag{3.7}$$

The maximizing values of the x_k are apparent from (3.7). In particular, if n is a multiple of P then this reduces to

$$h(n) = P\psi\left(\frac{n}{P}\right). \tag{3.8}$$

Corollary 3.1 is key to proving our results for scale-invariant trees.

3.2. Optimal leaf sets through recursive water-filling

Our goal is to determine a choice of n leaf nodes that gives the smallest possible LMMSE of the root. Recall that the LMMSE of V_γ given L_γ is defined as

$$\mathcal{E}(V_\gamma | L_\gamma) := \min_{\boldsymbol{\alpha}} \mathbb{E}(V_\gamma - \boldsymbol{\alpha}^T L_\gamma)^2, \tag{3.9}$$

where, in an abuse of notation, $\boldsymbol{\alpha}^T L_\gamma$ denotes a linear combination of the elements of L_γ with coefficients $\boldsymbol{\alpha}$. Crucial to our proofs is the fact that (Chou et al. [3] and Willsky [20]),

$$\frac{1}{\mathcal{E}(V_\gamma | L_\gamma)} + \frac{P_\gamma - 1}{\mathrm{var}(V_\gamma)} = \sum_{k=1}^{P_\gamma} \frac{1}{\mathcal{E}(V_\gamma | L_{\gamma k})}. \tag{3.10}$$

Denote the set consisting of all subsets of leaves of the tree of γ of size n by $\Lambda_\gamma(n)$. Motivated by (3.10) we introduce

$$\mu_\gamma(n) := \max_{L \in \Lambda_\gamma(n)} \mathcal{E}(V_\gamma | L)^{-1} \tag{3.11}$$

and define

$$\mathcal{L}_\gamma(n) := \{L \in \Lambda_\gamma(n) : \mathcal{E}(V_\gamma | L)^{-1} = \mu_\gamma(n)\}. \tag{3.12}$$

Restated, our goal is to determine one element of $\mathcal{L}_\emptyset(n)$. To allow a recursive approach through scale we generalize (3.11) and (3.12) by defining

$$\mu_{\gamma,\gamma'}(n) := \max_{L \in \Lambda_{\gamma'}(n)} \mathcal{E}(V_\gamma | L)^{-1} \quad \text{and} \tag{3.13}$$

$$\mathcal{L}_{\gamma,\gamma'}(n) := \{L \in \Lambda_{\gamma'}(n) : \mathcal{E}(V_\gamma | L)^{-1} = \mu_{\gamma,\gamma'}(n)\}. \tag{3.14}$$

Of course, $\mathcal{L}_\gamma(n) = \mathcal{L}_{\gamma,\gamma}(n)$. For the recursion, we are mostly interested in $\mathcal{L}_{\gamma,\gamma k}(n)$, i.e., the optimal estimation of a parent node from a sample of leaf nodes of one of its children. The following will be useful notation

$$X^* = [x_k^*]_{k=1}^{P_\gamma} := \arg \max_{X \in \Delta_n(\mathcal{N}_{\gamma 1}, \ldots, \mathcal{N}_{\gamma P_\gamma})} \sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k). \tag{3.15}$$

Using (3.10) we can decompose the problem of determining $L \in \mathcal{L}_\gamma(n)$ into smaller problems of determining elements of $\mathcal{L}_{\gamma,\gamma k}(x_k^*)$ for all k as stated in the next theorem.

Theorem 3.1. *For an independent innovations tree, let there be given one leaf set $L^{(k)}$ belonging to $\mathcal{L}_{\gamma,\gamma k}(x_k^*)$ for all k. Then $\bigcup_{k=1}^{P_\gamma} L^{(k)} \in \mathcal{L}_\gamma(n)$. Moreover, $\mathcal{L}_{\gamma k}(n) = \mathcal{L}_{\gamma k,\gamma k}(n) = \mathcal{L}_{\gamma,\gamma k}(n)$. Also $\mu_{\gamma,\gamma k}(n)$ is a positive, non-decreasing, and discrete-concave function of n, $\forall k, \gamma$.*

Theorem 3.1 gives us a two step procedure to obtain the best set of n leaves in the tree of γ to estimate V_γ. We first obtain the best set of x_k^* leaves in the tree of γk to estimate $V_{\gamma k}$ for all children γk of γ. We then take the union of these sets of leaves to obtain the required optimal set.

By sub-dividing the problem of obtaining optimal leaf nodes into smaller sub-problems we arrive at the following recursive technique to construct $L \in \mathcal{L}_\gamma(n)$. Starting at γ we move downward determining how many of the n leaf nodes of $L \in \mathcal{L}_\gamma(n)$ lie in the trees of the different descendants of γ until we reach the bottom. Assume for the moment that the functions $\mu_{\gamma,\gamma k}(n)$, for all γ, are given.

Scale-Recursive Water-filling scheme $\gamma \to \gamma k$

Step (a): Split n leaf nodes between the trees of γk, $k = 1, 2, \ldots, P_\gamma$.
First determine how to split the n leaf nodes between the trees of γk by maximizing $\sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k)$ over $X \in \Delta_n(\mathcal{N}_{\gamma 1}, \ldots, \mathcal{N}_{\gamma P_\gamma})$ (see (3.15)). The split is given by X^* which is easily obtained using the water-filling procedure for discrete-concave functions (defined in (3.4)) since $\mu_{\gamma,\gamma k}(n)$ is discrete-concave for all k. Determine $L^{(k)} \in \mathcal{L}_{\gamma,\gamma k}(x_k^*)$ since $L = \bigcup_{k=1}^{P_\gamma} L^{(k)} \in \mathcal{L}_\gamma(n)$.

Step (b): Split x_k^* nodes between the trees of child nodes of γk.
It turns out that $L^{(k)} \in \mathcal{L}_{\gamma,\gamma k}(x_k^*)$ if and only if $L^{(k)} \in \mathcal{L}_{\gamma k}(x_k^*)$. Thus repeat Step (a) with $\gamma = \gamma k$ and $n = x_k^*$ to construct $L^{(k)}$. Stop when we have reached the bottom of the tree.

We outline an efficient implementation of the scale-recursive water-filling algorithm. This implementation first computes $L \in \mathcal{L}_\gamma(n)$ for $n = 1$ and then inductively obtains the same for larger values of n. Given $L \in \mathcal{L}_\gamma(n)$ we obtain $L \in \mathcal{L}_\gamma(n+1)$ as follows. Note from Step (a) above that we determine how to split the n leaves at γ. We are now required to split $n+1$ leaves at γ. We easily obtain this from the earlier split of n leaves using (3.4). The water-filling technique maintains the split of n leaf nodes at γ while adding just one leaf node to the tree of one of the child nodes (say $\gamma k'$) of γ. We thus have to perform Step (b) only for $k = k'$. In this way the new leaf node "percolates" down the tree until we find its location at the bottom of the tree. The pseudo-code for determining $L \in \mathcal{L}_\gamma(n)$ given $\mathrm{var}(W_\gamma)$ for all γ as well as the proof that the recursive water-filling algorithm can be computed in polynomial-time are available online (Ribeiro et al. [14]).

3.3. Uniform leaf nodes are optimal for scale-invariant trees

The symmetry in scale-invariant trees forces the optimal solution to take a particular form irrespective of the variances of the innovations W_γ. We use the following notion of *uniform split* to prove that in a scale-invariant tree a more or less equal spread of sample leaf nodes across the tree gives the best linear estimate of the root.

Definition 3.2. Given a scale-invariant tree, a vector of leaf nodes L has uniform split of size n at node γ if $|L_\gamma| = n$ and $|L_{\gamma k}|$ is either $\lfloor \frac{n}{P_\gamma} \rfloor$ or $\lfloor \frac{n}{P_\gamma} \rfloor + 1$ for all values of k. It follows that $\#\{k : |L_{\gamma k}| = \lfloor \frac{n}{P_\gamma} \rfloor + 1\} = n - P_\gamma \lfloor \frac{n}{P_\gamma} \rfloor$.

Definition 3.3. Given a scale-invariant tree, a vector of leaf nodes is called a uniform leaf sample if it has a uniform split at all tree nodes.

The next theorem gives the optimal leaf node set for scale-invariant trees.

Theorem 3.2. *Given a scale-invariant tree, the uniform leaf sample of size n gives the best LMMSE estimate of the tree-root among all possible choices of n leaf nodes.*

Proof. For a scale-invariant tree, $\mu_{\gamma,\gamma k}(n)$ is identical for all k given any location γ. Corollary 3.1 and Theorem 3.1 then prove the theorem. □

4. Covariance trees

In this section we prove optimal and worst case solutions for covariance trees. For the optimal solutions we leverage our results for independent innovations trees and for the worst case solutions we employ eigenanalysis. We begin by formulating the problem.

4.1. Problem formulation

Let us compute the LMMSE of estimating the root $V_\emptyset$ given a set of leaf nodes L of size n. Recall that for a covariance tree the correlation between any leaf node and the root node is identical. We denote this correlation by ρ. Denote an $i \times j$ matrix with all elements equal to 1 by $\mathbf{1}_{i\times j}$. It is well known (Stark and Woods [17]) that the optimal linear estimate of $V_\emptyset$ given L (assuming zero-mean random variables) is given by $\rho \mathbf{1}_{1\times n} Q_L^{-1} L$, where Q_L is the covariance matrix of L and that the resulting LMMSE is

$$\begin{aligned} \mathcal{E}(V_\emptyset|L) &= \operatorname{var}(V_\emptyset) - \operatorname{cov}(L, V_\emptyset)^T Q_L^{-1} \operatorname{cov}(L, V_\emptyset) \\ &= \operatorname{var}(V_\emptyset) - \rho^2 \mathbf{1}_{1\times n} Q_L^{-1} \mathbf{1}_{n\times 1}. \end{aligned} \tag{4.1}$$

Clearly obtaining the best and worst-case choices for L is equivalent to maximizing and minimizing the sum of the elements of Q_L^{-1}. The exact value of ρ does not affect the solution. We assume that no element of L can be expressed as a linear combination of the other elements of L which implies that Q_L is invertible.

4.2. Optimal solutions

We use our results of Section 3 for independent innovations trees to determine the optimal solutions for covariance trees. Note from (4.2) that the estimation error for a covariance tree is a function only of the covariance *between leaf nodes*. Exploiting this fact, we first construct an independent innovations tree whose leaf nodes have the same correlation structure as that of the covariance tree and then prove that both trees must have the same optimal solution. Previous results then provide the optimal solution for the independent innovations tree which is also optimal for the covariance tree.

Definition 4.1. A matched innovations tree of a given covariance tree with positive correlation progression is an independent innovations tree with the following properties. It has (1) the same topology (2) and the same correlation structure between leaf nodes as the covariance tree, and (3) the root is equally correlated with all leaf nodes (though the exact value of the correlation between the root and a leaf node may differ from that of the covariance tree).

All covariance trees with positive correlation progression have corresponding matched innovations trees. We construct a matched innovations tree for a given covariance tree as follows. Consider an independent innovations tree with the same topology as the covariance tree. Set $\varrho_\gamma = 1$ for all γ,

$$\operatorname{var}(V_\emptyset) = c_0 \tag{4.2}$$

and

$$\operatorname{var}(W^{(j)}) = c_j - c_{j-1},\ j = 1, 2, \ldots, D, \tag{4.3}$$

where c_j is the covariance of leaf nodes of the covariance tree with proximity j and $\operatorname{var}(W^{(j)})$ is the common variance of all innovations of the independent innovations tree at scale j. Call c'_j the covariance of leaf nodes with proximity j in the independent innovations tree. From (2.1) we have

$$c'_j = \operatorname{var}(V_\emptyset) + \sum_{k=1}^{j} \operatorname{var}\left(W^{(k)}\right),\quad j = 1, \ldots, D. \tag{4.4}$$

Thus, $c'_j = c_j$ for all j and hence this independent innovations tree is the required matched innovations tree.

The next lemma relates the optimal solutions of a covariance tree and its matched innovations tree.

Lemma 4.1. *A covariance tree with positive correlation progression and its matched innovations tree have the same optimal leaf sets.*

Proof. Note that (4.2) applies to any tree whose root is equally correlated with all its leaves. This includes both the covariance tree and its matched innovations tree. From (4.2) we see that the choice of L that maximizes the sum of elements of Q_L^{-1} is optimal. Since Q_L^{-1} is identical for both the covariance tree and its matched innovations tree for any choice of L, they must have the same optimal solution. □

For a symmetric covariance tree that has positive correlation progression, the optimal solution takes on a specific form irrespective of the actual covariance between leaf nodes.

Theorem 4.1. *Given a symmetric covariance tree that has positive correlation progression, the uniform leaf sample of size n gives the best LMMSE of the tree-root among all possible choices of n leaf nodes.*

Proof. Form a matched innovations tree using the procedure outlined previously. This tree is by construction a scale-invariant tree. The result then follows from Theorem 3.2 and Lemma 4.1. □

While the uniform leaf sample is the optimal solution for a symmetric covariance tree with positive correlation progression, it is surprisingly the worst case solution for certain trees with a different correlation structure, which we prove next.

4.3. Worst case solutions

The *worst case solution* is any choice of $L \in \Lambda_\emptyset(n)$ that maximizes $\mathcal{E}(V_\emptyset|L)$. We now highlight the fact that the best and worst case solutions can change dramatically depending on the correlation structure of the tree. Of particular relevance to our discussion is the set of *clustered leaf nodes* defined as follows.

Definition 4.2. The set consisting of all leaf nodes of the tree of V_γ is called the set of clustered leaves of γ.

We provide the worst case solutions for covariance trees in which every node (with the exception of the leaves) has the same number of child nodes. The following theorem summarizes our results.

Theorem 4.2. *Consider a covariance tree of depth D in which every node (excluding the leaves) has the same number of child nodes σ. Then for leaf sets of size σ^p, $p = 0, 1, \ldots, D$, the worst case solution when the tree has positive correlation progression is given by the sets of clustered leaves of γ, where γ is any node at scale $D - p$. The worst case solution is given by the sets of uniform leaf nodes when the tree has negative correlation progression.*

Theorem 4.2 gives us the intuition that "more correlated" leaf nodes give worse estimates of the root. In the case of covariance trees with positive correlation progression, clustered leaf nodes are strongly correlated when compared to uniform leaf nodes. The opposite is true in the negative correlation progression case. Essentially if leaf nodes are highly correlated then they contain more redundant information which leads to poor estimation of the root.

While we have proved the optimal solution for covariance trees with positive correlation progression. we have not yet proved the same for those with negative correlation progression. Based on the intuition just gained we make the following conjecture.

Conjecture 4.1. Consider a covariance tree of depth D in which every node (excluding the leaves) has the same number of child nodes σ. Then for leaf sets of size σ^p, $p = 0, 1, \ldots, D$, the optimal solution when the tree has negative correlation progression is given by the sets of clustered leaves of γ, where γ is any node at scale $D - p$.

Using numerical techniques we support this conjecture in the next section.

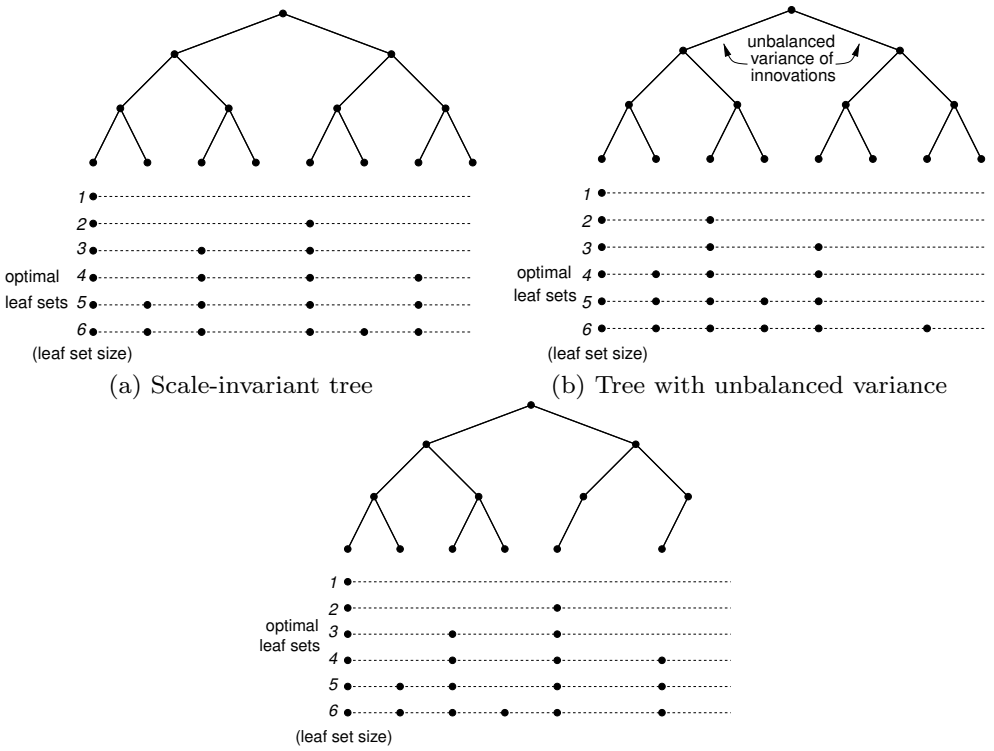
5. Numerical results

In this section, using the scale-recursive water-filling algorithm we evaluate the optimal leaf sets for independent innovations trees that are not scale-invariant. In addition we provide numerical support for Conjecture 4.1.

5.1. Independent innovations trees: scale-recursive water-filling

We consider trees with depth $D = 3$ and in which all nodes have at most two child nodes. The results demonstrate that the optimal leaf sets are a function of the correlation structure and topology of the multiscale trees.

In Fig. 4(a) we plot the optimal leaf node sets of different sizes for a scale-invariant tree. As expected the uniform leaf nodes sets are optimal.



(a) Scale-invariant tree

(b) Tree with unbalanced variance

(c) Tree with missing leaves

FIG 4. *Optimal leaf node sets for three different independent innovations trees:* (a) *scale-invariant tree,* (b) *symmetric tree with unbalanced variance of innovations at scale* 1*, and* (c) *tree with missing leaves at the finest scale. Observe that the uniform leaf node sets are optimal in* (a) *as expected. In* (b)*, however, the nodes on the left half of the tree are more preferable to those on the right. In* (c) *the solution is similar to* (a) *for optimal sets of size* $n = 5$ *or lower but changes for* $n = 6$ *due to the missing nodes.*

We consider a symmetric tree in Fig. 4(b), that is a tree in which all nodes have the same number of children (excepting leaf nodes). All parameters are constant within each scale except for the variance of the innovations W_γ at scale 1. The variance of the innovation on the right side is five times larger than the variance of the innovation on the left. Observe that leaves on the left of the tree are now preferable to those on the right and hence dominate the optimal sets. Comparing this result to Fig. 4(a) we see that the optimal sets are dependent on the correlation structure of the tree.

In Fig. 4(c) we consider the same tree as in Fig. 4(a) with two leaf nodes missing. These two leaves do not belong to the optimal leaf sets of size $n = 1$ to $n = 5$ in Fig. 4(a) but are elements of the optimal set for $n = 6$. As a result the optimal sets of size 1 to 5 in Fig. 4(c) are identical to those in Fig. 4(a) whereas that for $n = 6$ differs. This result suggests that the optimal sets depend on the tree topology.

Our results have important implications for applications because situations arise where we must model physical processes using trees with different correlation structures and topologies. For example, if the process to be measured is non-stationary over space then the multiscale tree may be unbalanced as in Fig. 4(b). In some applications it may not be possible to sample at certain locations due to physical constraints. We would thus have to exclude certain leaf nodes in our analysis as in Fig. 4(c).

The above experiments with tree-depth $D = 3$ are "toy-examples" to illustrate key concepts. In practice, the water-filling algorithm can solve much larger real-

world problems with ease. For example, on a Pentium IV machine running Matlab, the water-filling algorithm takes 22 seconds to obtain the optimal leaf set of size 100 to estimate the root of a binary tree with depth 11, that is a tree with 2048 leaves.

5.2. Covariance trees: best and worst cases

This section provides numerical support for Conjecture 4.1 that states that the clustered leaf node sets are optimal for covariance trees with negative correlation progression. We employ the WIG tree, a covariance tree in which each node has $\sigma = 2$ child nodes (Ma and Ji [12]). We provide numerical support for our claim using a WIG model of depth $D = 6$ possessing a fractional Gaussian noise-like[2] correlation structure corresponding to $H = 0.8$ and $H = 0.3$. To be precise, we choose the WIG model parameters such that the variance of nodes at scale j is proportional to 2^{-2jH} (see Ma and Ji [12] for further details). Note that $H > 0.5$ corresponds to positive correlation progression while $H \leq 0.5$ corresponds to negative correlation progression.

Fig. 5 compares the LMMSE of the estimated root node (normalized by the variance of the root) of the uniform and clustered sampling patterns. Since an exhaustive search of all possible patterns is very computationally expensive (for example there are over 10^{18} ways of choosing 32 leaf nodes from among 64) we instead compute the LMMSE for 10^4 randomly selected patterns. Observe that the clustered pattern gives the smallest LMMSE for the tree with negative correlation progression in Fig. 5(a) supporting our Conjecture 4.1 while the uniform pattern gives the smallest LMMSE for the positively correlation progression one in Fig. 5(b) as stated in Theorem 4.1. As proved in Theorem 4.2, the clustered and uniform patterns give the worst LMMSE for the positive and negative correlation progression cases respectively.

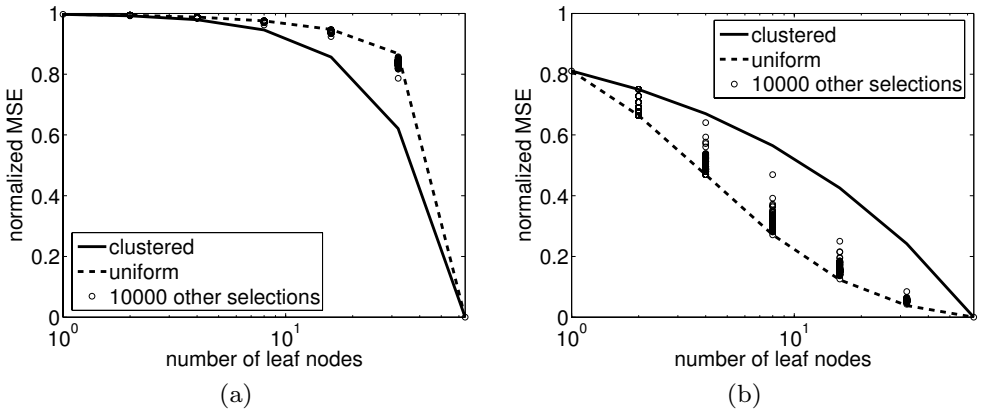


FIG 5. *Comparison of sampling schemes for a WIG model with* (a) *negative correlation progression and* (b) *positive correlation progression. Observe that the clustered nodes are optimal in* (a) *while the uniform is optimal in* (b). *The uniform and the clustered leaf sets give the worst performance in* (a) *and* (b) *respectively, as expected from our theoretical results.*

[2]Fractional Gaussian noise is the increments process of fBm (Mandelbrot and Ness [13]).

6. Related work

Earlier work has studied the problem of designing optimal samples of size n to linearly estimate the sum total of a process. For a one dimensional process which is wide-sense stationary with positive and convex correlation, within a class of unbiased estimators of the sum of the population, it was shown that systematic sampling of the process (uniform patterns with random starting points) is optimal (Hájek [6]).

For a two dimensional process on an $n_1 \times n_2$ grid with positive and convex correlation it was shown that an optimal sampling scheme does not lie in the class of schemes that ensure equal inclusion probability of $n/(n_1 n_2)$ for every point on the grid (Bellhouse [2]). In Bellhouse [2], an "optimal scheme" refers to a sampling scheme that achieves a particular lower bound on the error variance. The requirement of equal inclusion probability guarantees an unbiased estimator. The optimal schemes *within* certain sub-classes of this larger "equal inclusion probability" class were obtained using systematic sampling. More recent analysis refines these results to show that optimal designs do exist in the equal inclusion probability class for certain values of n, n_1, and n_2 and are obtained by Latin square sampling (Lawry and Bellhouse [10], Salehi [16]).

Our results differ from the above works in that we provide optimal solutions for the entire class of linear estimators and study a different set of random processes.

Other work on sampling fractional Brownian motion to estimate its Hurst parameter demonstrated that geometric sampling is superior to uniform sampling (Vidàcs and Virtamo [18]).

Recent work compared different probing schemes for traffic estimation through numerical simulations (He and Hou [7]). It was shown that a scheme which used uniformly spaced probes outperformed other schemes that used clustered probes. These results are similar to our findings for independent innovation trees and covariance trees with positive correlation progression.

7. Conclusions

This paper has addressed the problem of obtaining optimal leaf sets to estimate the root node of two types of multiscale stochastic processes: independent innovations trees and covariance trees. Our findings are particularly useful for applications which require the estimation of the sum total of a correlated population from a finite sample.

We have proved for an independent innovations tree that the optimal solution can be obtained using an efficient water-filling algorithm. Our results show that the optimal solutions can vary drastically depending on the correlation structure of the tree. For covariance trees with positive correlation progression as well as scale-invariant trees we obtained that uniformly spaced leaf nodes are optimal. However, uniform leaf nodes give the worst estimates for covariance trees with negative correlation progression. Numerical experiments support our conjecture that clustered nodes provide the optimal solution for covariance trees with negative correlation progression.

This paper raises several interesting questions for future research. The general problem of determining which n random variables from a given set provide the best linear estimate of another random variable that is not in the same set is an NP-hard problem. We, however, devised a fast polynomial-time algorithm to solve one

problem of this type, namely determining the optimal leaf set for an independent innovations tree. Clearly, the structure of independent innovations trees was an important factor that enabled a fast algorithm. The question arises as to whether there are similar problems that have polynomial-time solutions.

We have proved optimal results for covariance trees by reducing the problem to one for independent innovations trees. Such techniques of reducing one optimization problem to another problem that has an efficient solution can be very powerful. If a problem can be reduced to one of determining optimal leaf sets for independent innovations trees in polynomial-time, then its solution is also polynomial-time. Which other problems are malleable to this reduction is an open question.

Appendix

Proof of Lemma 3.1. We first prove the following statement.
Claim (1): If there exists $X^* = [x_k^*] \in \Delta_n(M_1, \ldots, M_P)$ that has the following property:

$$\psi_i(x_i^*) - \psi_i(x_i^* - 1) \geq \psi_j(x_j^* + 1) - \psi_j(x_j^*), \tag{7.1}$$

$\forall i \neq j$ such that $x_i^* > 0$ and $x_j^* < M_j$, then

$$h(n) = \sum_{k=1}^{P} \psi_k(x_k^*). \tag{7.2}$$

We then prove that such an X^* always exists and can be constructed using the water-filling technique.

Consider any $\widehat{X} \in \Delta_n(M_1, \ldots, M_P)$. Using the following steps, we transform the vector $\widehat{X}$ two elements at a time to obtain X^*.
Step 1: (Initialization) Set $X = \widehat{X}$.
Step 2: If $X \neq X^*$, then since the elements of both X and X^* sum up to n, there must exist a pair i, j such that $x_i \neq x_i^*$ and $x_j \neq x_j^*$. Without loss of generality assume that $x_i < x_i^*$ and $x_j > x_j^*$. This assumption implies that $x_i^* > 0$ and $x_j^* < M_j$. Now form vector Y such that $y_i = x_i + 1$, $y_j = x_j - 1$, and $y_k = x_k$ for $k \neq i, j$. From (7.1) and the concavity of ψ_i and ψ_j we have

$$\begin{aligned} \psi_i(y_i) - \psi_i(x_i) &= \psi_i(x_i + 1) - \psi_i(x_i) \geq \psi_i(x_i^*) - \psi_i(x_i^* - 1) \\ &\geq \psi_j(x_j^* + 1) - \psi_j(x_j^*) \geq \psi_j(x_j) - \psi_j(x_j - 1) \\ &\geq \psi_j(x_j) - \psi_j(y_j). \end{aligned} \tag{7.3}$$

As a consequence

$$\sum_k (\psi_k(y_k) - \psi_k(x_k)) = \psi_i(y_i) - \psi_i(x_i) + \psi_j(y_j) - \psi_j(x_j) \geq 0. \tag{7.4}$$

Step 3: If $Y \neq X^*$ then set $X = Y$ and repeat Step 2, otherwise stop.

After performing the above steps at most $\sum_k M_k$ times, $Y = X^*$ and (7.4) gives

$$\sum_k \psi_k(x_k^*) = \sum_k \psi_k(y_k) \geq \sum_k \psi_k(\widehat{x}_k). \tag{7.5}$$

This proves Claim (1).

Indeed for any $\widetilde{X} \neq X^*$ satisfying (7.1) we must have $\sum_k \psi_k(\widetilde{x}_k) = \sum_k \psi_k(x_k^*)$. We now prove the following claim by induction.

Claim (2): $G^{(n)} \in \Delta_n(M_1, \ldots, M_P)$ and $G^{(n)}$ satisfies (7.1).
(Initial Condition) The claim is trivial for $n = 0$.
(Induction Step) Clearly from (3.4) and (3.5)

$$\sum_k g_k^{(n+1)} = 1 + \sum_k g_k^{(n)} = n + 1, \tag{7.6}$$

and $0 \le g_k^{(n+1)} \le M_k$. Thus $G^{(n+1)} \in \Delta_{n+1}(M_1, \ldots, M_P)$. We now prove that $G^{(n+1)}$ satisfies property (7.1). We need to consider pairs i, j as in (7.1) for which either $i = m$ or $j = m$ because all other cases directly follow from the fact that $G^{(n)}$ satisfies (7.1).

Case (i) $j = m$, where m is defined as in (3.5). Assuming that $g_m^{(n+1)} < M_m$, for all $i \ne m$ such that $g_i^{(n+1)} > 0$ we have

$$\begin{aligned} \psi_i\left(g_i^{(n+1)}\right) - \psi_i\left(g_i^{(n+1)} - 1\right) &= \psi_i\left(g_i^{(n)}\right) - \psi_i\left(g_i^{(n)} - 1\right) \\ &\ge \psi_m\left(g_m^{(n)} + 1\right) - \psi_m\left(g_m^{(n)}\right) \\ &\ge \psi_m\left(g_m^{(n)} + 2\right) - \psi_m\left(g_m^{(n)} + 1\right) \\ &= \psi_m\left(g_m^{(n+1)} + 1\right) - \psi_m\left(g_m^{(n+1)}\right). \end{aligned} \tag{7.7}$$

Case (ii) $i = m$. Consider $j \ne m$ such that $g_j^{(n+1)} < M_j$. We have from (3.5) that

$$\begin{aligned} \psi_m\left(g_m^{(n+1)}\right) - \psi_m\left(g_m^{(n+1)} - 1\right) &= \psi_m\left(g_m^{(n)} + 1\right) - \psi_m\left(g_m^{(n)}\right) \\ &\ge \psi_j\left(g_j^{(n)} + 1\right) - \psi_j\left(g_j^{(n)}\right) \\ &= \psi_j\left(g_j^{(n+1)} + 1\right) - \psi_j\left(g_j^{(n+1)}\right). \end{aligned} \tag{7.8}$$

Thus Claim (2) is proved.

It only remains to prove the next claim.

Claim (3): $h(n)$, or equivalently $\sum_k \psi_k(g_k^{(n)})$, is non-decreasing and discrete-concave.

Since ψ_k is non-decreasing for all k, from (3.4) we have that $\sum_k \psi_k(g_k^{(n)})$ is a non-decreasing function of n. We have from (3.5)

$$\begin{aligned} h(n+1) - h(n) &= \sum_k \left(\psi_k(g_k^{(n+1)}) - \psi_k(g_k^{(n)})\right) \\ &= \max_{k: g_k^{(n)} < M_k} \left\{\psi_k(g_k^{(n)} + 1) - \psi_k(g_k^{(n)})\right\}. \end{aligned} \tag{7.9}$$

From the concavity of ψ_k and the fact that $g_k^{(n+1)} \ge g_k^{(n)}$ we have that

$$\psi_k(g_k^{(n)} + 1) - \psi_k(g_k^{(n)}) \ge \psi_k(g_k^{(n+1)} + 1) - \psi_k(g_k^{(n+1)}), \tag{7.10}$$

for all k. Thus from (7.10) and (7.10), $h(n)$ is discrete-concave. □

Proof of Corollary 3.1. Set $x_k^* = \left\lfloor \frac{n}{P} \right\rfloor$ for $1 \le k \le P - n + P\left\lfloor \frac{n}{P} \right\rfloor$ and $x_k^* = 1 + \left\lfloor \frac{n}{P} \right\rfloor$ for all other k. Then $X^* = [x_k^*] \in \Delta_n(M_1, \ldots, M_P)$ and X^* satisfies (7.1) from which the result follows. □

The following two lemmas are required to prove Theorem 3.1.

Lemma 7.1. *Given independent random variables A, W, F, define Z and E through $Z := \zeta A + W$ and $E := \eta Z + F$ where ζ, η are constants. We then have the result*

$$\frac{\mathrm{var}(A)}{\mathrm{cov}(A,E)^2} \cdot \frac{\mathrm{cov}(Z,E)^2}{\mathrm{var}(Z)} = \frac{\zeta^2 + \mathrm{var}(W)/\mathrm{var}(A)}{\zeta^2} \geq 1. \tag{7.11}$$

Proof. Without loss of generality assume all random variables have zero mean. We have

$$\mathrm{cov}(E,Z) = \mathbb{E}(\eta Z^2 + FZ) = \eta \mathrm{var}(Z), \tag{7.12}$$

$$\mathrm{cov}(A,E) = \mathbb{E}((\eta(\zeta A + W) + F)A)\zeta\eta\mathrm{var}(A), \tag{7.13}$$

and

$$\mathrm{var}(Z) = \mathbb{E}(\zeta^2 A^2 + W^2 + 2\zeta AW) = \zeta^2 \mathrm{var}(A) + \mathrm{var}(W). \tag{7.14}$$

Thus from (7.12), (7.13) and (7.14)

$$\frac{\mathrm{cov}(Z,E)^2}{\mathrm{var}(Z)} \cdot \frac{\mathrm{var}(A)}{\mathrm{cov}(A,E)^2} = \frac{\eta^2 \mathrm{var}(Z)}{\zeta^2 \eta^2 \mathrm{var}(A)} = \frac{\zeta^2 + \mathrm{var}(W)/\mathrm{var}(A)}{\zeta^2} \geq 1. \tag{7.15}$$

□

Lemma 7.2. *Given a positive function $z_i, i \in \mathbb{Z}$ and constant $\alpha > 0$ such that*

$$r_i := \frac{1}{1 - \alpha z_i} \tag{7.16}$$

is positive, discrete-concave, and non-decreasing, we have that

$$\delta_i := \frac{1}{1 - \beta z_i} \tag{7.17}$$

is also positive, discrete-concave, and non-decreasing for all β with $0 < \beta \leq \alpha$.

Proof. Define $\kappa_i := z_i - z_{i-1}$. Since z_i is positive and r_i is positive and non-decreasing, $\alpha z_i < 1$ and z_i must increase with i, that is $\kappa_i \geq 0$. This combined with the fact that $\beta z_i \leq \alpha z_i < 1$ guarantees that δ_i must be positive and non-decreasing.

It only remains to prove the concavity of δ_i. From (7.16)

$$r_{i+1} - r_i = \frac{\alpha(z_{i+1} - z_i)}{(1 - \alpha z_{i+1})(1 - \alpha z_i)} = \alpha \kappa_{i+1} r_{i+1} r_i. \tag{7.18}$$

We are given that r_i is discrete-concave, that is

$$\begin{aligned} 0 &\geq (r_{i+2} - r_{i+1}) - (r_{i+1} - r_i) \\ &= \alpha r_i r_{i+1} \left[\kappa_{i+2} \left(\frac{1 - \alpha z_i}{1 - \alpha z_{i+2}} \right) - \kappa_{i+1} \right]. \end{aligned} \tag{7.19}$$

Since $r_i > 0 \ \forall i$, we must have

$$\left[\kappa_{i+2} \left(\frac{1 - \alpha z_i}{1 - \alpha z_{i+2}} \right) - \kappa_{i+1} \right] \leq 0. \tag{7.20}$$

Similar to (7.20) we have that

$$(\delta_{i+2} - \delta_{i+1}) - (\delta_{i+1} - \delta_i) = \beta \delta_i \delta_{i+1} \left[\kappa_{i+2} \left(\frac{1 - \beta z_i}{1 - \beta z_{i+2}} \right) - \kappa_{i+1} \right]. \tag{7.21}$$

Since $\delta_i > 0\ \forall i$, for the concavity of δ_i it suffices to show that

$$\left[\kappa_{i+2}\frac{1-\beta z_i}{1-\beta z_{i+2}} - \kappa_{i+1}\right] \leq 0. \tag{7.22}$$

Now

$$\frac{1-\alpha z_i}{1-\alpha z_{i+2}} - \frac{1-\beta z_i}{1-\beta z_{i+2}} = \frac{(\alpha-\beta)(z_{i+2}-z_i)}{(1-\alpha z_{i+2})(1-\beta z_{i+2})} \geq 0. \tag{7.23}$$

Then (7.20) and (7.23) combined with the fact that $\kappa_i \geq 0,\ \forall i$ proves (7.22). □

Proof of Theorem 3.1. We split the theorem into three claims.

Claim (1): $L^* := \cup_k L^{(k)}(x_k^*) \in \mathcal{L}_\gamma(n)$.

From (3.10), (3.11), and (3.13) we obtain

$$\begin{aligned} \mu_\gamma(n) + \frac{P_\gamma - 1}{\mathrm{var}(V_\gamma)} &= \max_{L\in\Lambda_\gamma(n)} \sum_{k=1}^{P_\gamma} \mathcal{E}(V_\gamma|L_{\gamma k})^{-1} \\ &\leq \max_{X\in\Delta_n(\mathcal{N}_{\gamma 1},\ldots,\mathcal{N}_{\gamma P_\gamma})} \sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k). \end{aligned} \tag{7.24}$$

Clearly $L^* \in \Lambda_\gamma(n)$. We then have from (3.10) and (3.11) that

$$\begin{aligned} \mu_\gamma(n) + \frac{P_\gamma - 1}{\mathrm{var}(V_\gamma)} &\geq \mathcal{E}(V_\gamma|L^*)^{-1} + \frac{P_\gamma - 1}{\mathrm{var}(V_\gamma)} = \sum_{k=1}^{P_\gamma} \mathcal{E}(V_\gamma|L_{\gamma k}^*)^{-1} \\ &= \sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k^*) = \max_{X\in\Delta_n(\mathcal{N}_{\gamma 1},\ldots,\mathcal{N}_{\gamma P_\gamma})} \sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k). \end{aligned} \tag{7.25}$$

Thus from (7.25) and (7.26) we have

$$\mu_\gamma(n) = \mathcal{E}(V_\gamma|L^*)^{-1} = \max_{X\in\Delta_n(\mathcal{N}_{\gamma 1},\ldots,\mathcal{N}_{\gamma P_\gamma})} \sum_{k=1}^{P_\gamma} \mu_{\gamma,\gamma k}(x_k) - \frac{P_\gamma - 1}{\mathrm{var}(V_\gamma)}, \tag{7.26}$$

which proves Claim (1).

Claim (2): If $L \in \mathcal{L}_{\gamma k}(n)$ then $L \in \mathcal{L}_{\gamma,\gamma k}(n)$ and vice versa.
Denote an arbitrary leaf node of the tree of γk as E. Then V_γ, $V_{\gamma k}$, and E are related through

$$V_{\gamma k} = \varrho_{\gamma k} V_\gamma + W_{\gamma k}, \tag{7.27}$$

and

$$E = \eta V_{\gamma k} + F \tag{7.28}$$

where η and $\varrho_{\gamma k}$ are scalars and $W_{\gamma k},\ F$ and V_γ are independent random variables. We note that by definition $\mathrm{var}(V_\gamma) > 0\ \forall\gamma$ (see Definition 2.5). From Lemma 7.1

we have

$$\begin{aligned}\frac{\mathrm{cov}(V_{\gamma k},E)}{\mathrm{cov}(V_\gamma,E)} &= \left(\frac{\mathrm{var}(V_{\gamma k})}{\mathrm{var}(V_\gamma)}\right)^{1/2}\left(\frac{\varrho_{\gamma k}^2+\frac{\mathrm{var}(W_{\gamma k})}{\mathrm{var}(V_\gamma)}}{\varrho_{\gamma k}^2}\right)^{1/2}\\ &=: \xi_{\gamma,k}\geq\left(\frac{\mathrm{var}(V_{\gamma k})}{\mathrm{var}(V_\gamma)}\right)^{1/2}.\end{aligned}\tag{7.29}$$

From (7.30) we see that $\xi_{\gamma,k}$ is not a function of E.

Denote the covariance between V_γ and leaf node vector $L=[\ell_i]\in\Lambda_{\gamma k}(n)$ as $\Theta_{\gamma,L}=[\mathrm{cov}(V_\gamma,\ell_i)]^T$. Then (7.30) gives

$$\Theta_{\gamma k,L}=\xi_{\gamma,k}\Theta_{\gamma,L}.\tag{7.30}$$

From (4.2) we have

$$\mathcal{E}(V_\gamma|L)=\mathrm{var}(V_\gamma)-\varphi(\gamma,L)\tag{7.31}$$

where $\varphi(\gamma,L)=\Theta_{\gamma,L}^T Q_L^{-1}\Theta_{\gamma,L}$. Note that $\varphi(\gamma,L)\geq 0$ since Q_L^{-1} is positive semi-definite. Using (7.30) we similarly get

$$\mathcal{E}(V_{\gamma k}|L)=\mathrm{var}(V_{\gamma k})-\frac{\varphi(\gamma,L)}{\xi_{\gamma,k}^2}.\tag{7.32}$$

From (7.31) and (7.32) we see that $\mathcal{E}(V_\gamma|L)$ and $\mathcal{E}(V_{\gamma k}|L)$ are both minimized over $L\in\Lambda_{\gamma k}(n)$ by the same leaf vector that maximizes $\varphi(\gamma,L)$. This proves Claim (2).

Claim (3): $\mu_{\gamma,\gamma k}(n)$ is a positive, non-decreasing, and discrete-concave function of n, $\forall k,\gamma$.

We start at a node γ at one scale from the bottom of the tree and then move up the tree.

Initial Condition: Note that $V_{\gamma k}$ is a leaf node. From (2.1) and (**??**) we obtain

$$\mathcal{E}(V_\gamma|V_{\gamma k})=\mathrm{var}(V_\gamma)-\frac{(\varrho_{\gamma k}\mathrm{var}(V_\gamma))^2}{\mathrm{var}(V_{\gamma k})}\leq\mathrm{var}(V_\gamma).\tag{7.33}$$

For our choice of γ, $\mu_{\gamma,\gamma k}(1)$ corresponds to $\mathcal{E}(V_\gamma|V_{\gamma k})^{-1}$ and $\mu_{\gamma,\gamma k}(0)$ corresponds to $1/\mathrm{var}(V_\gamma)$. Thus from (7.33), $\mu_{\gamma,\gamma k}(n)$ is positive, non-decreasing, and discrete-concave (trivially since n takes only two values here).

Induction Step: Given that $\mu_{\gamma,\gamma k}(n)$ is a positive, non-decreasing, and discrete-concave function of n for $k=1,\ldots,P_\gamma$, we prove the same when γ is replaced by $\gamma\uparrow$. Without loss of generality choose k such that $(\gamma\uparrow)k=\gamma$. From (3.11), (3.13), (7.31), (7.32) and Claim (2), we have for $L\in\mathcal{L}_\gamma(n)$

$$\begin{aligned}\mu_\gamma(n)&=\frac{1}{\mathrm{var}(V_\gamma)}\cdot\frac{1}{1-\frac{\varphi(\gamma,L)}{\mathrm{var}(V_\gamma)}},\quad\text{and}\\ \mu_{\gamma\uparrow,k}(n)&=\frac{1}{\mathrm{var}(V_{\gamma\uparrow})}\cdot\frac{1}{1-\frac{\varphi(\gamma,L)}{\xi_{\gamma\uparrow,k}^2\mathrm{var}(V_{\gamma\uparrow})}}.\end{aligned}\tag{7.34}$$

From (7.26), the assumption that $\mu_{\gamma,\gamma k}(n)$ $\forall k$ is a positive, non-decreasing, and discrete-concave function of n, and Lemma 3.1 we have that $\mu_\gamma(n)$ is a non-decreasing and discrete-concave function of n. Note that by definition (see (3.11))

$\mu_\gamma(n)$ is positive. This combined with (2.1), (7.35), (7.30) and Lemma 7.2, then prove that $\mu_{\gamma\uparrow,k}(n)$ is also positive, non-decreasing, and discrete-concave. □

We now prove a lemma to be used to prove Theorem 4.2. As a first step we compute the leaf arrangements L which maximize and minimize the sum of all elements of $Q_L = [q_{i,j}(L)]$. We restrict our analysis to a covariance tree with depth D and in which each node (excluding leaf nodes) has σ child nodes. We introduce some notation. Define

$$\Gamma^{(u)}(p) := \{L : L \in \Lambda_\emptyset(\sigma^p) \text{ and } L \text{ is a uniform leaf node set}\} \quad \text{and} \tag{7.35}$$

$$\Gamma^{(c)}(p) := \{L : L \text{ is a clustered leaf set of a node at scale } D-p\} \tag{7.36}$$

for $p = 0, 1, \ldots, D$. We number nodes at scale m in an arbitrary order from $q = 0, 1, \ldots, \sigma^m - 1$ and refer to a node by the pair (m, q).

Lemma 7.3. *Assume a positive correlation progression. Then, $\sum_{i,j} q_{i,j}(L)$ is minimized over $L \in \Lambda_\emptyset(\sigma^p)$ by every $L \in \Gamma^{(u)}(p)$ and maximized by every $L \in \Gamma^{(c)}(p)$. For a negative correlation progression, $\sum_{i,j} q_{i,j}(L)$ is maximized by every $L \in \Gamma^{(u)}(p)$ and minimized by every $L \in \Gamma^{(c)}(p)$.*

Proof. Set p to be an arbitrary element in $\{1, \ldots, D-1\}$. The cases of $p = 0$ and $p = D$ are trivial. Let $\vartheta_m = \#\{q_{i,j}(L) \in Q_L : q_{i,j}(L) = c_m\}$ be the number of elements of Q_L equal to c_m. Define $a_m := \sum_{k=0}^m \vartheta_k, m \geq 0$ and set $a_{-1} = 0$. Then

$$\begin{aligned}
\sum_{i,j} q_{i,j} &= \sum_{m=0}^{D} c_m \vartheta_m = \sum_{m=0}^{D-1} c_m (a_m - a_{m-1}) + c_D \vartheta_D \\
&= \sum_{m=0}^{D-1} c_m a_m - \sum_{m=-1}^{D-2} c_{m+1} a_m + c_D \vartheta_D \\
&= \sum_{m=0}^{D-2} (c_m - c_{m+1}) a_m + c_{D-1} a_{D-1} - c_0 a_{-1} + c_D \vartheta_D \\
&= \sum_{m=0}^{D-2} (c_m - c_{m+1}) a_m + \text{constant},
\end{aligned} \tag{7.37}$$

where we used the fact that $a_{D-1} = a_D - \vartheta_D$ is a constant independent of the choice of L, since $\vartheta_D = \sigma^p$ and $a_D = \sigma^{2p}$.

We now show that $L \in \Gamma^{(u)}(p)$ maximizes $a_m, \forall m$ while $L \in \Gamma^{(c)}(p)$ minimizes $a_m, \forall m$. First we prove the results for $L \in \Gamma^{(u)}(p)$. Note that L has one element in the tree of every node at scale p.

Case (i) $m \geq p$. Since every element of L has proximity at most $p-1$ with all other elements, $a_m = \sigma^p$ which is the maximum value it can take.

Case (ii) $m < p$ (assuming $p > 0$). Consider an arbitrary ordering of nodes at scale $m+1$. We refer to the q^{th} node in this ordering as "the q^{th} node at scale $m+1$".

Let the number of elements of L belonging to the sub-tree of the q^{th} node at scale $m+1$ be $g_q, q = 0, \ldots, \sigma^{m+1} - 1$. We have

$$a_m = \sum_{q=0}^{\sigma^{m+1}-1} g_q(\sigma^p - g_q) = \frac{\sigma^{2p+1+m}}{4} - \sum_{q=0}^{\sigma^{m+1}-1} (g_q - \sigma^p/2)^2 \tag{7.38}$$

since every element of L in the tree of the q^{th} node at scale $m+1$ must have proximity *at most* m with all nodes *not* in the same tree but must have proximity *at least* $m+1$ with all nodes *within* the same tree.

The choice of g_q's is constrained to lie on the hyperplane $\sum_q g_q = \sigma^p$. Obviously the quadratic form of (7.38) is maximized by the point on this hyperplane closest to the point $(\sigma^p/2, \ldots, \sigma^p/2)$ which is $(\sigma^{p-m-1}, \ldots, \sigma^{p-m-1})$. This is clearly achieved by $L \in \Gamma^{(u)}(p)$.

Now we prove the results for $L \in \Gamma^{(c)}(p)$.
Case (i) $m < D - p$. We have $a_m = 0$, the smallest value it can take.
Case (ii) $D - p \leq m < D$. Consider leaf node $\ell_i \in L$ which without any loss of generality belongs to the tree of first node at scale $m+1$. Let $a_m(\ell_i)$ be the number of elements of L to which ℓ_i has proximity less than or equal to m. Now since ℓ_i has proximity less than or equal to m only with those elements of L not in the same tree, we must have $a_m(\ell_i) \geq \sigma^p - \sigma^{D-m-1}$. Since $L \in \Gamma^{(c)}(p)$ achieves this lower bound for $a_m(\ell_i), \forall i$ and $a_m = \sum_i a_m(\ell_i)$, $L \in \Gamma^{(c)}$ minimizes a_m in turn. □

We now study to what extent the above results transfer to the actual matrix of interest Q_L^{-1}. We start with a useful lemma.

Lemma 7.4. *Denote the eigenvalues of Q_L by $\lambda_j, j = 1, \ldots, \sigma^p$. Assume that no leaf node of the tree can be expressed as a linear combination of other leaf nodes, implying that $\lambda_j > 0, \forall j$. Set $\mathcal{D}_L = [d_{i,j}]_{\sigma^p \times \sigma^p} := Q_L^{-1}$. Then there exist positive numbers f_i with $f_1 + \ldots + f_p = 1$ such that*

$$\sum_{i,j=1}^{\sigma^p} q_{i,j} = \sigma^p \sum_{j=1}^{\sigma^p} f_j \lambda_j, \text{ and} \tag{7.39}$$

$$\sum_{i,j=1}^{\sigma^p} d_{i,j} = \sigma^p \sum_{j=1}^{\sigma^p} f_j / \lambda_j. \tag{7.40}$$

Furthermore, for both special cases, $L \in \Gamma^{(u)}(p)$ and $L \in \Gamma^{(c)}(p)$, we may choose the weights f_j such that only one is non-zero.

Proof. Since the matrix Q_L is real and symmetric there exists an orthonormal eigenvector matrix $U = [u_{i,j}]$ that diagonalizes Q_L, that is $Q_L = U \Xi U^T$ where Ξ is diagonal with eigenvalues $\lambda_j, j = 1, \ldots, \sigma^p$. Define $w_j := \sum_i u_{i,j}$. Then

$$\begin{aligned} \sum_{i,j} q_{i,j} &= \mathbf{1}_{1\times\sigma^p} Q_L \mathbf{1}_{\sigma^p\times 1} = (\mathbf{1}_{1\times\sigma^p} U)\Xi(\mathbf{1}_{1\times\sigma^p} U)^T \\ &= [w_1 \ldots w_{\sigma^p}]\Xi[w_1 \ldots w_{\sigma^p}]^T = \sum_j \lambda_j w_j^2. \end{aligned} \tag{7.41}$$

Further, since $U^T = U^{-1}$ we have

$$\sum_j w_j^2 = (\mathbf{1}_{1\times\sigma^p} U)(U^T \mathbf{1}_{\sigma^p\times 1}) = \mathbf{1}_{1\times\sigma^p} I \mathbf{1}_{\sigma^p\times 1} = \sigma^p. \tag{7.42}$$

Setting $f_i = w_i^2/\sigma^p$ establishes (7.39). Using the decomposition

$$Q_L^{-1} = (U^T)^{-1}\Xi^{-1}U^{-1} = U\Xi^{-1}U^T \tag{7.43}$$

similarly gives (7.40).

Consider the case $L \in \Gamma^{(u)}(p)$. Since $L = [\ell_i]$ consists of a symmetrical set of leaf nodes (the set of proximities between any element ℓ_i and the rest does not depend

on i) the sum of the covariances of a leaf node ℓ_i with its fellow leaf nodes does not depend on i, and we can set

$$\lambda^{(u)} := \sum_{j=1}^{\sigma^p} q_{i,j}(L) = c_D + \sum_{m=1}^{p} \sigma^{p-m} c_m. \tag{7.44}$$

With the sum of the elements of any row of Q_L being identical, the vector $\mathbf{1}_{\sigma^p \times 1}$ is an eigenvector of Q_L with eigenvalue $\lambda^{(u)}$ equal to (7.44).

Recall that we can always choose a basis of orthogonal eigenvectors that includes $\mathbf{1}_{\sigma^p \times 1}$ as the first basis vector. It is well known that the rows of the corresponding basis transformation matrix U will then be exactly these normalized eigenvectors. Since they are orthogonal to $\mathbf{1}_{\sigma^p \times 1}$, the sum of their coordinates w_j ($j = 2, \ldots, \sigma^p$) must be zero. Thus, all f_i but f_1 vanish. (The last claim follows also from the observation that the sum of coordinates of the normalized $\mathbf{1}_{\sigma^p \times 1}$ equals $w_1 = \sigma^p \sigma^{-p/2} = \sigma^{p/2}$; due to (7.42) $w_j = 0$ for all other j.)

Consider the case $L \in \Gamma^{(u)}(p)$. The reasoning is similar to the above, and we can define

$$\lambda^{(c)} := \sum_{j=1}^{\sigma^p} q_{i,j}(L) = c_D + \sum_{m=1}^{p} \sigma^{m} c_{D-m}. \tag{7.45}$$

□

Proof of Theorem 4.2. Due to the special form of the covariance vector $\mathrm{cov}(L, V_\emptyset) = \rho \mathbf{1}_{1 \times \sigma^k}$ we observe from (4.2) that minimizing the LMMSE $\mathcal{E}(V_\emptyset | L)$ over $L \in \Lambda_\emptyset(n)$ is equivalent to maximizing $\sum_{i,j} d_{i,j}(L)$ the sum of the elements of Q_L^{-1}.

Note that the weights f_i and the eigenvalues λ_i of Lemma 7.4 depend on the arrangement of the leaf nodes L. To avoid confusion, denote by λ_i the eigenvalues of Q_L for an arbitrary fixed set of leaf nodes L, and by $\lambda^{(u)}$ and $\lambda^{(c)}$ the only relevant eigenvalues of $L \in \Gamma^{(u)}(p)$ and $L \in \Gamma^{(c)}(p)$ according to (7.44) and (7.45).

Assume a positive correlation progression, and let L be an arbitrary set of σ^p leaf nodes. Lemma 7.3 and Lemma 7.4 then imply that

$$\lambda^{(u)} \le \sum_j \lambda_j f_j \le \lambda^{(c)}. \tag{7.46}$$

Since Q_L is positive definite, we must have $\lambda_j > 0$. We may then interpret the middle expression as an expectation of the positive "random variable" λ with discrete law given by f_i. By Jensen's inequality,

$$\sum_j (1/\lambda_j) f_j \ge \frac{1}{\sum_j \lambda_j f_j} \ge \frac{1}{\lambda^{(c)}}. \tag{7.47}$$

Thus, $\sum_{i,j} d_{i,j}$ is minimized by $L \in \Gamma^{(c)}(p)$; that is, clustering of nodes gives the worst LMMSE. A similar argument holds for the negative correlation progression case which proves the Theorem. □

References

[1] Abry, P., Flandrin, P., Taqqu, M. and Veitch, D. (2000). Wavelets for the analysis, estimation and synthesis of scaling data. In *Self-similar Network Traffic and Performance Evaluation*. Wiley.

[2] Bellhouse, D. R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika* **64**, 3 (Dec.), 605–611.

[3] Chou, K. C., Willsky, A. S. and Benveniste, A. (1994). Multiscale recursive estimation, data fusion, and regularization. *IEEE Trans. on Automatic Control* **39**, 3, 464–478.

[4] Cover, T. M. and Thomas, J. A. (1991). *Information Theory*. Wiley Interscience.

[5] Cressie, N. (1993). *Statistics for Spatial Data*. Revised edition. Wiley, New York.

[6] Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pěst. Mat.* **84**, 387–423. Also available in *Collected Works of Jaroslav Hájek – With Commentary* by M. Hušková, R. Beran and V. Dupač, Wiley, 1998.

[7] He, G. and Hou, J. C. (2003). On exploiting long-range dependency of network traffic in measuring cross-traffic on an end-to-end basis. *IEEE INFOCOM*.

[8] Jamdee, S. and Los, C. A. (2004). Dynamic risk profile of the US term structure by wavelet MRA. Tech. Rep. 0409045, Economics Working Paper Archive at WUSTL.

[9] Kuchment, L. S. and Gelfan, A. N. (2001). Statistical self-similarity of spatial variations of snow cover: verification of the hypothesis and application in the snowmelt runoff generation models. *Hydrological Processes* **15**, 3343–3355.

[10] Lawry, K. A. and Bellhouse, D. R. (1992). Relative efficiency of certian randomization procedures in an $n \times n$ array when spatial correlation is present. *Jour. Statist. Plann. Inference* **32**, 385–399.

[11] Li, Q. and Mills, D. L. (1999). Investigating the scaling behavior, crossover and anti-persistence of Internet packet delay dynamics. *Proc. IEEE GLOBECOM Symposium*, 1843–1852.

[12] Ma, S. and Ji, C. (1998). Modeling video traffic in the wavelet domain. *IEEE INFOCOM*, 201–208.

[13] Mandelbrot, B. B. and Ness, J. W. V. (1968). Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review* **10**, 4 (Oct.), 422–437.

[14] Ribeiro, V. J., Riedi, R. H., and Baraniuk, R. G. Pseudo-code and computational complexity of water-filling algorithm for independent innovations trees. Available at `http://www.stat.rice.edu/~vinay/waterfilling/pseudo.pdf`.

[15] Riedi, R., Crouse, M. S., Ribeiro, V., and Baraniuk, R. G. (1999). A multifractal wavelet model with application to TCP network traffic. *IEEE Trans. on Information Theory* **45**, 3, 992–1018.

[16] Salehi, M. M. (2004). Optimal sampling design under a spatial correlation model. *J. of Statistical Planning and Inference* **118**, 9–18.

[17] Stark, H. and Woods, J. W. (1986). *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice-Hall.

[18] Vidàcs, A. and Virtamo, J. T. (1999). ML estimation of the parameters of fBm traffic with geometrical sampling. *COST257* **99**, 14.

[19] Willett, R., Martin, A., and Nowak, R. (2004). Backcasting: Adaptive sampling for sensor networks. *Information Processing in Sensor Networks (IPSN)*.

[20] Willsky, A. (2002). Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE* **90**, 8, 1396–1458.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 291–311
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000518

The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations

Hannes Leeb[1,*]

Yale University

Abstract: We analyze the (unconditional) distribution of a linear predictor that is constructed after a data-driven model selection step in a linear regression model. First, we derive the exact finite-sample cumulative distribution function (cdf) of the linear predictor, and a simple approximation to this (complicated) cdf. We then analyze the large-sample limit behavior of these cdfs, in the fixed-parameter case and under local alternatives.

1. Introduction

The analysis of the unconditional distribution of linear predictors after model selection given in this paper complements and completes the results of Leeb [1], where the corresponding conditional distribution is considered, conditional on the outcome of the model selection step. The present paper builds on Leeb [1] as far as finite-sample results are concerned. For a large-sample analysis, however, we can not rely on that paper; the limit behavior of the unconditional cdf differs from that of the conditional cdfs so that a separate analysis is necessary. For a review of the relevant related literature and for an outline of applications of our results, we refer the reader to Leeb [1].

We consider a linear regression model $Y = X\theta + u$ with normal errors. (The normal linear model facilitates a detailed finite-sample analysis. Also note that asymptotic properties of the Gaussian location model can be generalized to a much larger model class including nonlinear models and models for dependent data, as long as appropriate standard regularity conditions guaranteeing asymptotic normality of the maximum likelihood estimator are satisfied.) We consider model selection by a sequence of 'general-to-specific' hypothesis tests; that is, starting from the overall model, a sequence of tests is used to simplify the model. The cdf of a linear function of the post-model-selection estimator (properly scaled and centered) is denoted by $G_{n,\theta,\sigma}(t)$. The notation suggests that this cdf depends on the sample size n, the regression parameter θ, and on the error variance σ^2. An explicit formula for $G_{n,\theta,\sigma}(t)$ is given in (3.10) below. From this formula, we see that the distribution of, say, a linear predictor after model selection is significantly different from

[1]Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511.

*Research supported by the Max Kade Foundation and by the Austrian Science Foundation (FWF), project no. P13868-MAT. A preliminary version of this manuscript was written in February 2002.

AMS 2000 subject classifications: primary 62E15; secondary 62F10, 62F12, 62J05.

Keywords and phrases: model uncertainty, model selection, inference after model selection, distribution of post-model-selection estimators, linear predictor constructed after model selection, pre-test estimator.

(and more complex than) the Gaussian distribution that one would get without model selection. Because the cdf $G_{n,\theta,\sigma}(t)$ is quite difficult to analyze directly, we also provide a uniform asymptotic approximation to this cdf. This approximation, which we shall denote by $G^*_{n,\theta,\sigma}(t)$, is obtained by considering an 'idealized' scenario where the error variance σ^2 is treated as known and is used by the model selection procedure. The approximating cdf $G^*_{n,\theta,\sigma}(t)$ is much simpler and allows us to observe the main effects of model selection. Moreover, this approximation allows us to study the large-sample limit behavior of $G_{n,\theta,\sigma}(t)$ not only in the fixed-parameter case but also along sequences of parameters. The consideration of asymptotics along sequences of parameters is necessitated by a complication that seems to be inherent to post-model-selection estimators: Convergence of the finite-sample distributions to the large-sample limit distribution is non-uniform in the underlying parameters. (See Corollary 5.5 in Leeb and Pötscher [3], Appendix B in Leeb and Pötscher [4].) For applications like the computation of large-sample limit minimal coverage probabilities, it therefore appears to be necessary to study the limit behavior of $G_{n,\theta,\sigma}(t)$ along *sequences of parameters* $\theta^{(n)}$ and $\sigma^{(n)}$. We characterize all accumulation points of $G_{n,\theta^{(n)},\sigma^{(n)}}(t)$ for such sequences (with respect to weak convergence). Ex post, it turns out that, as far as possible accumulation points are concerned, it suffices to consider only a particular class of parameter sequences, namely local alternatives. Of course, the large-sample limit behavior of $G_{n,\theta,\sigma}(t)$ in the fixed-parameter case is contained in this analysis. Besides, we also consider the model selection probabilities, i.e., the probabilities of selecting each candidate model under consideration, in the finite-sample and in the large-sample limit case.

The remainder of the paper is organized as follows: In Section 2, we describe the basic framework of our analysis and the quantities of interest: The post-model-selection estimator $\tilde{\theta}$ and the cdf $G_{n,\theta,\sigma}(t)$. Besides, we also introduce the 'idealized post-model-selection estimator' $\tilde{\theta}^*$ and the cdf $G^*_{n,\theta,\sigma}(t)$, which correspond to the case where the error variance is known. In Section 3, we derive finite-sample expansions of the aforementioned cdfs, and we discuss and illustrate the effects of the model selection step in finite samples. Section 4 contains an approximation result which shows that $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$ are asymptotically uniformly close to each other. With this, we can analyze the large-sample limit behavior of the two cdfs in Section 5. All proofs are relegated to the appendices.

2. The model and estimators

Consider the linear regression model

$$Y = X\theta + u, \tag{2.1}$$

where X is a non-stochastic $n \times P$ matrix with $\operatorname{rank}(X) = P$ and $u \sim N(0, \sigma^2 I_n)$, $\sigma^2 > 0$. Here n denotes the sample size and we assume $n > P \geq 1$. In addition, we assume that $Q = \lim_{n\to\infty} X'X/n$ exists and is non-singular (this assumption is not needed in its full strength for all of the asymptotic results; cf. Remark 2.1). Similarly as in Pötscher [6], we consider model selection from a collection of nested models $M_{\mathcal{O}} \subseteq M_{\mathcal{O}+1} \subseteq \cdots \subseteq M_P$ which are given by $M_p = \{(\theta_1, \ldots, \theta_P)' \in \mathbb{R}^P : \theta_{p+1} = \cdots = \theta_P = 0\}$ $(0 \leq p \leq P)$. Hence, the model M_p corresponds to the situation where only the first p regressors in (2.1) are included. For the most parsimonious model under consideration, i.e., for $M_{\mathcal{O}}$, we assume that

$\mathcal{O}$ satisfies $0 \le \mathcal{O} < P$; if $\mathcal{O} > 0$, this model contains those components of the parameter that will not be subject to model selection. Note that $M_0 = \{(0, \ldots, 0)'\}$ and $M_P = \mathbb{R}^P$. We call M_p the regression model of order p.

The following notation will prove useful. For matrices B and C of the same row-dimension, the column-wise concatenation of B and C is denoted by $(B : C)$. If D is an $m \times P$ matrix, let $D[p]$ denote the matrix of the first p columns of D. Similarly, let $D[\neg p]$ denote the matrix of the last $P - p$ columns of D. If x is a $P \times 1$ (column-) vector, we write in abuse of notation $x[p]$ and $x[\neg p]$ for $(x'[p])'$ and $(x'[\neg p])'$, respectively. (We shall use these definitions also in the 'boundary' cases $p = 0$ and $p = P$. It will always be clear from the context how expressions like $D[0]$, $D[\neg P]$, $x[0]$, or $x[\neg P]$ are to be interpreted.) As usual the i-th component of a vector x will be denoted by x_i; in a similar fashion, denote the entry in the i-th row and j-th column of a matrix B by $B_{i,j}$.

The restricted least-squares estimator for θ under the restriction $\theta[\neg p] = 0$ will be denoted by $\tilde{\theta}(p)$, $0 \le p \le P$ (in case $p = P$, the restriction is void, of course). Note that $\tilde{\theta}(p)$ is given by the $P \times 1$ vector whose first p components are given by $(X[p]'X[p])^{-1}X[p]'Y$, and whose last $P - p$ components are equal to zero; the expressions $\tilde{\theta}(0)$ and $\tilde{\theta}(P)$, respectively, are to be interpreted as the zero-vector in $\mathbb{R}^P$ and as the unrestricted least-squares estimator for θ. Given a parameter vector θ in $\mathbb{R}^P$, the order of θ, relative to the set of models $M_0, \ldots, M_P$, is defined as $p_0(\theta) = \min\{p : \ 0 \le p \le P, \ \theta \in M_p\}$. Hence, if θ is the true parameter vector, only models M_p of order $p \ge p_0(\theta)$ are correct models, and $M_{p_0(\theta)}$ is the most parsimonious correct model for θ among $M_0, \ldots, M_P$. We stress that $p_0(\theta)$ is a property of a *single parameter*, and hence needs to be distinguished from the notion of the order of the model M_p introduced earlier, which is a property of the *set of parameters* M_p.

A model selection procedure in general is now nothing else than a data-driven (measurable) rule $\hat{p}$ that selects a value from $\{\mathcal{O}, \ldots, P\}$ and thus selects a model from the list of candidate models $M_{\mathcal{O}}, \ldots, M_P$. In this paper, we shall consider a model selection procedure based on a sequence of 'general-to-specific' hypothesis tests, which is given as follows: The sequence of hypotheses $H_0^p : \ p_0(\theta) < p$ is tested against the alternatives $H_1^p : \ p_0(\theta) = p$ in decreasing order starting at $p = P$. If, for some $p > \mathcal{O}$, H_0^p is the first hypothesis in the process that is rejected, we set $\hat{p} = p$. If no rejection occurs until even $H_0^{\mathcal{O}+1}$ is accepted, we set $\hat{p} = \mathcal{O}$. Each hypothesis in this sequence is tested by a kind of t-test where the error variance is always estimated from the overall model. More formally, we have

$$\hat{p} \ = \ \max\{p : \ |T_p| \ge c_p, \, 0 \le p \le P\},$$

where the test-statistics are given by $T_0 = 0$ and by $T_p = \sqrt{n}\tilde{\theta}_p(p)/(\hat{\sigma}\xi_{n,p})$ with

$$\xi_{n,p} \ = \ \left(\left[\left(\frac{X[p]'X[p]}{n} \right)^{-1} \right]_{p,p} \right)^{\frac{1}{2}} \qquad (0 < p \le P) \tag{2.2}$$

being the (non-negative) square root of the p-th diagonal element of the matrix indicated, and with $\hat{\sigma}^2 = (n - P)^{-1}(Y - X\tilde{\theta}(P))'(Y - X\tilde{\theta}(P))$ (cf. also Remark 6.2 in Leeb [1] concerning other variance estimators). The critical values c_p are independent of sample size (cf., however, Remark 2.1) and satisfy $0 < c_p < \infty$ for $\mathcal{O} < p \le P$. We also set $c_{\mathcal{O}} = 0$ in order to restrict $\hat{p}$ to the range of candidate models under consideration, i.e., to $\{\mathcal{O}, \mathcal{O} + 1, \ldots, P\}$. Note that under the hypothesis

H_0^p the statistic T_p is t-distributed with $n-P$ degrees of freedom for $0 < p \leq P$. The so defined model selection procedure $\hat{p}$ is conservative (or over-consistent): The probability of selecting an incorrect model, i.e., the probability of the event $\{\hat{p} < p_0(\theta)\}$, converges to zero as the sample size increases; the probability of selecting a correct (but possibly over-parameterized) model, i.e., the probability of the event $\{\hat{p} = p\}$ for p satisfying $\max\{p_0(\theta), \mathcal{O}\} \leq p \leq P$, converges to a positive limit; cf. (5.7) below.

The post-model-selection estimator $\tilde{\theta}$ is now defined as follows: On the event $\hat{p} = p$, $\tilde{\theta}$ is given by the restricted least-squares estimator $\tilde{\theta}(p)$, i.e.,

$$\tilde{\theta} \;=\; \sum_{p=\mathcal{O}}^{P} \tilde{\theta}(p)\, \mathbf{1}\{\hat{p} = p\}. \tag{2.3}$$

To study the distribution of a linear function of $\tilde{\theta}$, let A be a non-stochastic $k \times P$ matrix of rank k ($1 \leq k \leq P$). Examples for A include the case where A equals a $1 \times P$ (row-) vector x_f if the object of interest is the linear predictor $x_f\tilde{\theta}$, or the case where $A = (I_s : 0)$, say, if the object of interest is an $s \times 1$ subvector of θ. We shall consider the cdf

$$G_{n,\theta,\sigma}(t) \;=\; P_{n,\theta,\sigma}\left(\sqrt{n}A(\tilde{\theta} - \theta) \;\leq\; t\right) \qquad (t \in \mathbb{R}^k). \tag{2.4}$$

Here and in the following, $P_{n,\theta,\sigma}(\cdot)$ denotes the probability measure corresponding to a sample of size n from (2.1) under the true parameters θ and σ. For convenience we shall refer to (2.4) as the cdf of $A\tilde{\theta}$, although (2.4) is in fact the cdf of an affine transformation of $A\tilde{\theta}$.

For theoretical reasons we shall also be interested in the idealized model selection procedure which assumes knowledge of σ^2 and hence uses T_p^* instead of T_p, where $T_p^* = \sqrt{n}\tilde{\theta}_p(p)/(\sigma\xi_{n,p})$, $0 < p \leq P$, and $T_0^* = 0$. The corresponding model selector is denoted by $\hat{p}^*$ and the resulting idealized 'post-model-selection estimator' by $\tilde{\theta}^*$. Note that under the hypothesis H_0^p the variable T_p^* is standard normally distributed for $0 < p \leq P$. The corresponding cdf will be denoted by $G_{n,\theta,\sigma}^*(t)$, i.e.,

$$G_{n,\theta,\sigma}^*(t) \;=\; P_{n,\theta,\sigma}\left(\sqrt{n}A(\tilde{\theta}^* - \theta) \;\leq\; t\right) \qquad (t \in \mathbb{R}^k). \tag{2.5}$$

For convenience we shall also refer to (2.5) as the cdf of $A\tilde{\theta}^*$.

Remark 2.1. Some of the assumptions introduced above are made only to simplify the exposition and can hence easily be relaxed. This includes, in particular, the assumption that the critical values c_p used by the model selection procedure do not depend on sample size, and the assumption that the regressor matrix X is such that $X'X/n$ converges to a positive definite limit Q as $n \to \infty$. For the finite-sample results in Section 3 below, these assumptions are clearly inconsequential. Moreover, for the large-sample limit results in Sections 4 and 5 below, these assumptions can be relaxed considerably. For the details, see Remark 6.1(i)–(iii) in Leeb [1], which also applies, mutatis mutandis, to the results in the present paper.

3. Finite-sample results

Some further preliminaries are required before we can proceed. The expected value of the restricted least-squares estimator $\tilde{\theta}(p)$ will be denoted by $\eta_n(p)$ and is given

by the $P \times 1$ vector

$$\eta_n(p) \;=\; \begin{pmatrix} \theta[p] \,+\, (X[p]'X[p])^{-1}X[p]'X[\neg p]\theta[\neg p] \\ (0,\ldots,0)' \end{pmatrix} \tag{3.1}$$

with the conventions that $\eta_n(0) = (0,\ldots,0)' \in \mathbb{R}^P$ and $\eta_n(P) = \theta$. Furthermore, let $\Phi_{n,p}(t)$, $t \in \mathbb{R}^k$, denote the cdf of $\sqrt{n}A(\tilde{\theta}(p) - \eta_n(p))$, i.e., $\Phi_{n,p}(t)$ is the cdf of a centered Gaussian random vector with covariance matrix $\sigma^2 A[p](X[p]'X[p]/n)^{-1}A[p]'$ in case $p > 0$, and the cdf of point-mass at zero in $\mathbb{R}^k$ in case $p = 0$. If $p > 0$ and if the matrix $A[p]$ has rank k, then $\Phi_{n,p}(t)$ has a density with respect to Lebesgue measure, and we shall denote this density by $\phi_{n,p}(t)$. We note that $\eta_n(p)$ depends on θ and that $\Phi_{n,p}(t)$ depends on σ (in case $p > 0$), although these dependencies are not shown explicitly in the notation.

For $p > 0$, the conditional distribution of $\sqrt{n}\tilde{\theta}_p(p)$ given $\sqrt{n}A(\tilde{\theta}(p) - \eta_n(p)) = z$ is a Gaussian distribution with mean $\sqrt{n}\eta_{n,p}(p) + b_{n,p}z$ and variance $\sigma^2\zeta_{n,p}^2$, where

$$b_{n,p} \;=\; C_n^{(p)'}(A[p](X[p]'X[p]/n)^{-1}A[p]')^{-}, \text{ and} \tag{3.2}$$

$$\zeta_{n,p}^2 \;=\; \xi_{n,p}^2 - b_{n,p}C_n^{(p)}. \tag{3.3}$$

In the displays above, $C_n^{(p)}$ stands for $A[p](X[p]'X[p]/n)^{-1}e_p$, with e_p denoting the p-th standard basis vector in $\mathbb{R}^p$, and $(A[p](X[p]'X[p]/n)^{-1}A[p]')^{-}$ denotes a generalized inverse of the matrix indicated (cf. Note 3(v) in Section 8a.2 of Rao [7]). Note that, in general, the quantity $b_{n,p}z$ depends on the choice of generalized inverse in (3.2); however, for z in the column-space of $A[p]$, $b_{n,p}z$ is invariant under the choice of inverse; cf. Lemma A.2 in Leeb [1]. Since $\sqrt{n}A(\tilde{\theta}(p) - \eta_n(p))$ lies in the column-space of $A[p]$, the conditional distribution of $\sqrt{n}\tilde{\theta}_p(p)$ given $\sqrt{n}A(\tilde{\theta}(p) - \eta_n(p)) = z$ is thus well-defined by the above. Observe that the vector of covariances between $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ is given by $\sigma^2 n^{-1}C_n^{(p)}$. In particular, note that $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ are uncorrelated if and only if $\zeta_{n,p}^2 = \xi_{n,p}^2$ (or, equivalently, if and only if $b_{n,p}z = 0$ for all z in the column-space of $A[p]$); again, see Lemma A.2 in Leeb [1].

Finally, for M denoting a univariate Gaussian random variable with zero mean and variance $s^2 \geq 0$, we abbreviate the probability $P(|M - a| < b)$ by $\Delta_s(a, b)$, $a \in \mathbb{R} \cup \{-\infty, \infty\}$, $b \in \mathbb{R}$. Note that $\Delta_s(\cdot,\cdot)$ is symmetric around zero in its first argument, and that $\Delta_s(-\infty, b) = \Delta_s(\infty, b) = 0$ holds. In case $s = 0$, M is to be interpreted as being equal to zero, such that $\Delta_0(a, b)$ equals one if $|a| < b$ and zero otherwise; i.e., $\Delta_0(a, b)$ reduces to an indicator function.

3.1. The known-variance case

The cdf $G_{n,\theta,\sigma}^*(t)$ can be expanded as a weighted sum of conditional cdfs, conditional on the outcome of the model selection step, where the weights are given by the corresponding model selection probabilities. To this end, let $G_{n,\theta,\sigma}^*(t|p)$ denote the conditional cdf of $\sqrt{n}A(\tilde{\theta}^* - \theta)$ given that $\hat{p}^*$ equals p for $\mathcal{O} \leq p \leq P$; that is, $G_{n,\theta,\sigma}^*(t|p) = P_{n,\theta,\sigma}(\sqrt{n}A(\tilde{\theta}^* - \theta) \leq t \,|\, \hat{p}^* = p)$, with $t \in \mathbb{R}^k$. Moreover, let $\pi_{n,\theta,\sigma}^*(p) = P_{n,\theta,\sigma}(\hat{p}^* = p)$ denote the corresponding model selection probability. Then the unconditional cdf $G_{n,\theta,\sigma}^*(t)$ can be written as

$$G_{n,\theta,\sigma}^*(t) \;=\; \sum_{p=\mathcal{O}}^{P} G_{n,\theta,\sigma}^*(t|p)\pi_{n,\theta,\sigma}^*(p). \tag{3.4}$$

Explicit finite-sample formulas for $G^*_{n,\theta,\sigma}(t|p)$, $\mathcal{O} \le p \le P$, are given in Leeb [1], equations (10) and (13). Let $\gamma(\xi_{n,q}, s) = \Delta_{\sigma\xi_{n,q}}(\sqrt{n}\eta_{n,q}(q), sc_q\sigma\xi_{n,q})$, and $\gamma^*(\zeta_{n,q}, z, s) = \Delta_{\sigma\zeta_{n,q}}(\sqrt{n}\eta_{n,q}(q) + b_{n,q}z, sc_q\sigma\xi_{n,q})$ It is elementary to verify that $\pi^*_{n,\theta,\sigma}(\mathcal{O})$ is given by

$$\pi^*_{n,\theta,\sigma}(\mathcal{O}) = \prod_{q=\mathcal{O}+1}^{P} \gamma(\xi_{n,q}, 1) \tag{3.5}$$

while, for $p > \mathcal{O}$, we have

$$\pi^*_{n,\theta,\sigma}(p) = (1 - \gamma(\xi_{n,p}, 1)) \times \prod_{q=p+1}^{P} \gamma(\xi_{n,q}, 1). \tag{3.6}$$

(This follows by arguing as in the discussion leading up to (12) of Leeb [1], and by using Proposition 3.1 of that paper.) Observe that the model selection probability $\pi^*_{n,\theta,\sigma}(p)$ is always positive for each p, $\mathcal{O} \le p \le P$.

Plugging the formulas for the conditional cdfs obtained in Leeb [1] and the above formulas for the model selection probabilities into (3.4), we obtain that $G^*_{n,\theta,\sigma}(t)$ is given by

$$\begin{aligned} G^*_{n,\theta,\sigma}(t) \;=\; & \Phi_{n,\mathcal{O}}(t - \sqrt{n}A(\eta_n(\mathcal{O}) - \theta)) \prod_{q=\mathcal{O}+1}^{P} \gamma(\xi_{n,q}, 1) \\ & + \sum_{p=\mathcal{O}+1}^{P} \int_{z \le t - \sqrt{n}A(\eta_n(p)-\theta)} (1 - \gamma^*(\zeta_{n,p}, z, 1))\, \Phi_{n,p}(dz) \\ & \times \prod_{q=p+1}^{P} \gamma(\xi_{n,q}, 1). \end{aligned} \tag{3.7}$$

In the above display, $\Phi_{n,p}(dz)$ denotes integration with respect to the measure induced by the cdf $\Phi_{n,p}(t)$ on $\mathbb{R}^k$.

3.2. The unknown-variance case

Similar to the known-variance case, define $G_{n,\theta,\sigma}(t|p) = P_{n,\theta,\sigma}(\sqrt{n}A(\tilde{\theta} - \theta) \le t|\hat{p} = p)$ and $\pi_{n,\theta,\sigma}(p) = P_{n,\theta,\sigma}(\hat{p} = p)$, $\mathcal{O} \le p \le P$. Then $G_{n,\theta,\sigma}(t)$ can be expanded as the sum of the terms $G_{n,\theta,\sigma}(t|p)\pi_{n,\theta,\sigma}(p)$ for $p = \mathcal{O}, \ldots, P$, similar to (3.4).

For the model selection probabilities, we argue as in Section 3.2 of Leeb and Pötscher [3] to obtain that $\pi_{n,\theta,\sigma}(\mathcal{O})$ equals

$$\pi_{n,\theta,\sigma}(\mathcal{O}) = \int_0^\infty \prod_{q=\mathcal{O}+1}^{P} \gamma(\xi_{n,q}, s) h(s) ds, \tag{3.8}$$

where h denotes the density of $\hat{\sigma}/\sigma$, i.e., h is the density of $(n - P)^{-1/2}$ times the square-root of a chi-square distributed random variable with $n - P$ degrees of freedom. In a similar fashion, for $p > \mathcal{O}$, we get

$$\pi_{n,\theta,\sigma}(p) = \int_0^\infty (1 - \gamma(\xi_{n,p}, s)) \prod_{q=p+1}^{P} \gamma(\xi_{n,q}, s) h(s) ds; \tag{3.9}$$

cf. the argument leading up to (18) in Leeb [1]. As in the known-variance case, the model selection probabilities are all positive.

Using the formulas for the conditional cdfs $G_{n,\theta,\sigma}(t|p)$, $\mathcal{O} \le p \le P$, given in Leeb [1], equations (14) and (16)–(18), the unconditional cdf $G_{n,\theta,\sigma}(t)$ is thus seen to be given by

$$\begin{aligned} G_{n,\theta,\sigma}(t) &= \Phi_{n,\mathcal{O}}(t - \sqrt{n}A(\eta_n(\mathcal{O}) - \theta)) \int_0^\infty \prod_{q=\mathcal{O}+1}^{P} \gamma(\xi_{n,q}, s)h(s)ds \\ &\quad + \sum_{p=\mathcal{O}+1}^{P} \int_{z \le t - \sqrt{n}A(\eta_n(p)-\theta)} \Big[\int_0^\infty (1 - \gamma^*(\zeta_{n,p}, z, s)) \\ &\qquad \times \prod_{q=p+1}^{P} \gamma(\xi_{n,q}, s)h(s)ds \Big] \Phi_{n,p}(dz). \end{aligned} \tag{3.10}$$

Observe that $G_{n,\theta,\sigma}(t)$ is in fact a smoothed version of $G^*_{n,\theta,\sigma}(t)$: Indeed, the right-hand side of the formula (3.10) for $G_{n,\theta,\sigma}(t)$ is obtained by taking the right-hand side of formula (3.7) for $G^*_{n,\theta,\sigma}(t)$, changing the last argument of $\gamma(\xi_{n,q}, 1)$ and $\gamma^*(\zeta_{n,q}, z, 1)$ from 1 to s for $q = \mathcal{O}+1, \ldots, P$, integrating with respect to $h(s)ds$, and interchanging the order of integration. Similar considerations apply, mutatis mutandis, to the model selection probabilities $\pi_{n,\theta,\sigma}(p)$ and $\pi^*_{n,\theta,\sigma}(p)$ for $\mathcal{O} \le p \le P$.

3.3. Discussion

3.3.1. General Observations

The cdfs $G^*_{n,\theta,\sigma}(t)$ and $G_{n,\theta,\sigma}(t)$ need not have densities with respect to Lebesgue measure on $\mathbb{R}^k$. However, densities do exist if $\mathcal{O} > 0$ and the matrix $A[\mathcal{O}]$ has rank k. In that case, the density of $G_{n,\theta,\sigma}(t)$ is given by

$$\begin{aligned} &\phi_{n,\mathcal{O}}(t - \sqrt{n}A(\eta_n(\mathcal{O}) - \theta)) \int_0^\infty \prod_{q=\mathcal{O}+1}^{P} \gamma(\xi_{n,q}, s)h(s)ds \\ &\quad + \sum_{p=\mathcal{O}+1}^{P} \Big[\int_0^\infty (1 - \gamma^*(\zeta_{n,p}, t - \sqrt{n}A(\eta_n(p) - \theta), s)) \\ &\qquad \times \prod_{q=p+1}^{P} \gamma(\xi_{n,q}, s)h(s)ds \Big] \phi_{n,p}(t - \sqrt{n}A(\eta_n(p) - \theta)). \end{aligned} \tag{3.11}$$

(Given that $\mathcal{O} > 0$ and that $A[\mathcal{O}]$ has rank k, we see that $A[p]$ has rank k and that the Lebesgue density $\phi_{n,p}(t)$ of $\Phi_{n,p}(t)$ exists for each $p = \mathcal{O}, \ldots, P$. We hence may write the integrals with respect to $\Phi_{n,p}(dz)$ in (3.10) as integrals with respect to $\phi_{n,p}(z)dz$. Differentiating the resulting formula for $G_{n,\theta,\sigma}(t)$ with respect to t, we get (3.11).) Similarly, the Lebesgue density of $G^*_{n,\theta,\sigma}(t)$ can be obtained by differentiating the right-hand side of (3.7), provided that $\mathcal{O} > 0$ and $A[\mathcal{O}]$ has rank k. Conversely, if that condition is violated, then some of the conditional cdfs are degenerate and Lebesgue densities do not exist. (Note that on the event $\hat{p} = p$, $A\tilde{\theta}$ equals $A\tilde{\theta}(p)$, and recall that the last $P-p$ coordinates of $\tilde{\theta}(p)$ are constant equal to zero. Therefore $A\tilde{\theta}(0)$ is the zero-vector in $\mathbb{R}^k$ and, for $p > 0$, $A\tilde{\theta}(p)$ is concentrated

in the column space of $A[p]$. On the event $\hat{p}^* = p$, a similar argument applies to $A\tilde{\theta}^*$.)

Both cdfs $G^*_{n,\theta,\sigma}(t)$ and $G_{n,\theta,\sigma}(t)$ are given by a weighted sum of conditional cdfs, cf. (3.7) and (3.10), where the weights are given by the model-selection probabilities (which are always positive in finite samples). For a detailed discussion of the conditional cdfs, the reader is referred to Section 3.3 of Leeb [1].

The cdf $G_{n,\theta,\sigma}(t)$ is typically highly non-Gaussian. A notable exception where $G_{n,\theta,\sigma}(t)$ reduces to the Gaussian cdf $\Phi_{n,P}(t)$ for each $\theta \in \mathbb{R}^P$ occurs in the special case where $\tilde{\theta}_p(p)$ is uncorrelated with $A\tilde{\theta}(p)$ for each $p = \mathcal{O}+1, \ldots, P$. In this case, we have $A\tilde{\theta}(p) = A\tilde{\theta}(P)$ for each $p = \mathcal{O}, \ldots, P$ (cf. the discussion following (20) in Leeb [1]). From this and in view of (2.3), it immediately follows that $G_{n,\theta,\sigma}(t) = \Phi_{n,P}(t)$, independent of θ and σ. (The same considerations apply, mutatis mutandis, to $G^*_{n,\theta,\sigma}(t)$.) Clearly, this case is rather special, because it entails that fitting the overall model with P regressors gives the same estimator for $A\theta$ as fitting the restricted model with $\mathcal{O}$ regressors only.

To compare the distribution of a linear function of the post-model-selection estimator with the distribution of the post-model-selection estimator itself, note that the cdf of $\tilde{\theta}$ can be studied in our framework by setting A equal to I_P (and k equal to P). Obviously, the distribution of $\tilde{\theta}$ does not have a density with respect to Lebesgue measure. Moreover, $\tilde{\theta}_p(p)$ is always perfectly correlated with $\tilde{\theta}(p)$ for each $p = 1, \ldots, P$, such that the special case discussed above can not occur (for A equal to I_P).

3.3.2. An illustrative example

We now exemplify the possible shapes of the finite-sample distributions in a simple setting. To this end, we set $P = 2$, $\mathcal{O} = 1$, $A = (1, 0)$, and $k = 1$ for the rest of this section. The choice of $P = 2$ gives a special case of the model (2.1), namely

$$Y_i = \theta_1 X_{i,1} + \theta_2 X_{i,2} + u_i \qquad (1 \leq i \leq n). \tag{3.12}$$

With $\mathcal{O} = 1$, the first regressor is always included in the model, and a pre-test will be employed to decide whether or not to include the second one. The two model selectors $\hat{p}$ and $\hat{p}^*$ thus decide between two candidate models, $M_1 = \{(\theta_1, \theta_2)' \in \mathbb{R}^2 : \theta_2 = 0\}$ and $M_2 = \{(\theta_1, \theta_2)' \in \mathbb{R}^2\}$. The critical value for the test between M_1 and M_2, i.e., c_2, will be chosen later (recall that we have set $c_{\mathcal{O}} = c_1 = 0$). With our choice of $A = (1, 0)$, we see that $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$ are the cdfs of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ and $\sqrt{n}(\tilde{\theta}^*_1 - \theta_1)$, respectively.

Since the matrix $A[\mathcal{O}]$ has rank one and $k = 1$, the cdfs of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ and $\sqrt{n}(\tilde{\theta}^*_1 - \theta_1)$ both have Lebesgue densities. To obtain a convenient expression for these densities, we write $\sigma^2(X'X/n)^{-1}$, i.e., the covariance matrix of the least-squares estimator based on the overall model (3.12), as

$$\sigma^2 \left(\frac{X'X}{n} \right)^{-1} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}.$$

The elements of this matrix depend on sample size n, but we shall suppress this dependence in the notation. It will prove useful to define $\rho = \sigma_{1,2}/(\sigma_1\sigma_2)$, i.e., ρ is the correlation coefficient between the least-squares estimators for θ_1 and θ_2 in model (3.12). Note that here we have $\phi_{n,2}(t) = \sigma_1^{-1}\phi(t/\sigma_1)$ and $\phi_{n,1}(t) =$

$\sigma_1^{-1}(1-\rho^2)^{-1/2}\phi(t(1-\rho^2)^{-1/2}/\sigma_1)$ with $\phi(t)$ denoting the univariate standard Gaussian density. The density of $\sqrt{n}(\tilde{\theta}_1-\theta_1)$ is given by

$$\begin{aligned}(3.13)\qquad &\phi_{n,1}(t+\sqrt{n}\theta_2\rho\sigma_1/\sigma_2)\int_0^\infty \Delta_1(\sqrt{n}\theta_2/\sigma_2, sc_2)h(s)ds \\ &\quad+\phi_{n,2}(t)\int_0^\infty (1-\Delta_1(\frac{\sqrt{n}\theta_2/\sigma_2+\rho t/\sigma_1}{\sqrt{1-\rho^2}}, \frac{sc_2}{\sqrt{1-\rho^2}}))h(s)ds;\end{aligned}$$

recall that $\Delta_1(a,b)$ is equal to $\Phi(a+b)-\Phi(a-b)$, where $\Phi(t)$ denotes the standard univariate Gaussian cdf, and note that here $h(s)$ denotes the density of $(n-2)^{-1/2}$ times the square-root of a chi-square distributed random variable with $n-2$ degrees of freedom. Similarly, the density of $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$ is given by

$$\begin{aligned}(3.14)\qquad &\phi_{n,1}(t+\sqrt{n}\theta_2\rho\sigma_1/\sigma_2)\Delta_1(\sqrt{n}\theta_2/\sigma_2, c_2) \\ &\quad+\phi_{n,2}(t)(1-\Delta_1(\frac{\sqrt{n}\theta_2/\sigma_2+\rho t/\sigma_1}{\sqrt{1-\rho^2}}, \frac{c_2}{\sqrt{1-\rho^2}})).\end{aligned}$$

Note that both densities depend on the regression parameter $(\theta_1,\theta_2)'$ only through θ_2, and that these densities depend on the error variance σ^2 and on the regressor matrix X only through σ_1, σ_2, and ρ. Also note that the expressions in (3.13) and (3.14) are unchanged if ρ is replaced by $-\rho$ and, at the same time, the argument t is replaced by $-t$. Similarly, replacing θ_2 and t by $-\theta_2$ and $-t$, respectively, leaves (3.13) and (3.14) unchanged. The same applies also to the conditional densities considered below; cf. (3.15) and (3.16). We therefore consider only non-negative values of ρ and θ_2 in the numerical examples below.

From (3.14) we can also read-off the conditional densities of $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$, conditional on selecting the model M_p for $p=1$ and $p=2$, which will be useful later: The unconditional cdf of $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$ is the weighted sum of two conditional cdfs, conditional on selecting the model M_1 and M_2, respectively, weighted by the corresponding model selection probabilities; cf. (3.4) and the attending discussion. Hence, the unconditional density is the sum of the conditional densities multiplied by the corresponding model selection probabilities. In the simple setting considered here, the probability of $\hat{p}^*$ selecting M_1, i.e., $\pi^*_{n,\theta,\sigma}(1)$, equals $\Delta_1(\sqrt{n}\theta_2/\sigma_2, c_2)$ in view of (3.5) and because $\mathcal{O}=1$, and $\pi^*_{n,\theta,\sigma}(2)=1-\pi^*_{n,\theta,\sigma}(1)$. Thus, conditional on selecting the model M_1, the density of $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$ is given by

$$(3.15)\qquad \phi_{n,1}(t+\sqrt{n}\theta_2\rho\sigma_1/\sigma_2).$$

Conditional on selecting M_2, the density of $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$ equals

$$(3.16)\qquad \phi_{n,2}(t)\frac{1-\Delta_1((\sqrt{n}\theta_2/\sigma_2+\rho t/\sigma_1)/\sqrt{1-\rho^2}, c_2/\sqrt{1-\rho^2})}{1-\Delta_1(\sqrt{n}\theta_2/\sigma_2, c_2)}.$$

This can be viewed as a 'deformed' version of $\phi_{n,2}(t)$, i.e., the density of $\sqrt{n}(\tilde{\theta}_1(2)-\theta_1)$, where the deformation is governed by the fraction in (3.16). The conditional densities of $\sqrt{n}(\tilde{\theta}_1-\theta_1)$ can be obtained and interpreted in a similar fashion from (3.13), upon observing that $\pi_{n,\theta,\sigma}(1)$ here equals $\int_0^\infty \Delta_1(\sqrt{n}\theta_2/\sigma_2, sc_2)h(s)ds$ in view of (3.8)

Figure 1 illustrates some typical shapes of the densities of $\sqrt{n}(\tilde{\theta}_1-\theta_1)$ and $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$ given in (3.13) and (3.14), respectively, for $\rho=0.75$, $n=7$, and for various values of θ_2. Note that the densities of $\sqrt{n}(\tilde{\theta}_1-\theta_1)$ and $\sqrt{n}(\tilde{\theta}_1^*-\theta_1)$,

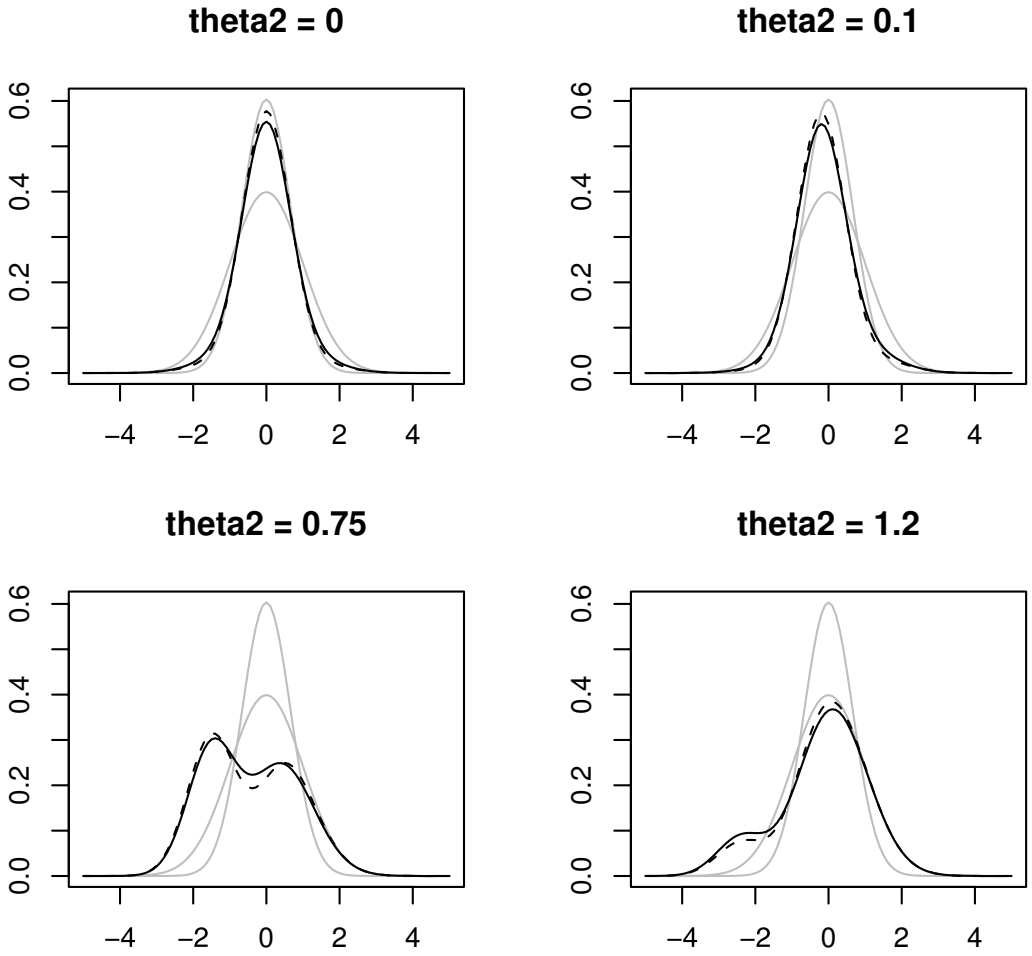


FIG 1. *The densities of* $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ *(black solid line) and of* $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$ *(black dashed line) for the indicated values of* θ_2*,* $n = 7$*,* $\rho = 0.75$*, and* $\sigma_1 = \sigma_2 = 1$*. The critical value of the test between* M_1 *and* M_2 *was set to* $c_2 = 2.015$*, corresponding to a t-test with significance level* 0.9*. For reference, the gray curves are Gaussian densities* $\phi_{n,1}(t)$ *(larger peak) and* $\phi_{n,2}(t)$ *(smaller peak).*

corresponding to the unknown-variance case and the (idealized) known-variance case, are very close to each other. In fact, the small sample size, i.e., $n = 7$, was chosen because for larger n these two densities are visually indistinguishable in plots as in Figure 1 (this phenomenon is analyzed in detail in the next section). For $\theta_2 = 0$ in Figure 1, the density of $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$, although seemingly close to being Gaussian, is in fact a mixture of a Gaussian density and a bimodal density; this is explained in detail below. For the remaining values of θ_2 considered in Figure 1, the density of $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$ is clearly non-Gaussian, namely skewed in case $\theta_2 = 0.1$, bimodal in case $\theta_2 = 0.75$, and highly non-symmetric in case $\theta_2 = 1.2$. Overall, we see that the finite-sample density of $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$ can exhibit a variety of different shapes. Exactly the same applies to the density of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$. As a point of interest, we note that these different shapes occur for values of θ_2 in a quite narrow range: For example, in the setting of Figure 1, the uniformly most powerful test of the hypothesis $\theta_2 = 0$ against $\theta_2 > 0$ with level 0.95, i.e., a one-sided t-test, has a power of only 0.27 at the alternative $\theta_2 = 1.2$. This suggests that estimating the distribution of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ is difficult here. (See also Leeb and Pöstcher [4] as well as Leeb and Pöstcher [2] for a thorough analysis of this difficulty.)

We stress that the phenomena shown in Figure 1 are not caused by the small

sample size, i.e., $n = 7$. This becomes clear upon inspection of (3.13) and (3.14), which depend on θ_2 through $\sqrt{n}\theta_2$ (for fixed σ_1, σ_2 and ρ). Hence, for other values of n, one obtains plots essentially similar to Figure 1, provided that the range of values of θ_2 is adapted accordingly.

We now show how the shape of the unconditional densities can be explained by the shapes of the conditional densities together with the model selection probabilities. Since the unknown-variance case and the known-variance case are very similar as seen above, we focus on the latter. In Figure 2 below, we give the conditional densities of $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$, conditional on selecting the model M_p, $p = 1, 2$, cf. (3.15) and (3.16), and the corresponding model selection probabilities in the same setting as in Figure 1.

The unconditional densities of $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$ in each panel of Figure 1 are the sum of the two conditional densities in the corresponding panel in Figure 2, weighted by the model selection probabilities, i.e, $\pi^*_{n,\theta,\sigma}(1)$ and $\pi^*_{n,\theta,\sigma}(2)$. In other words, in each panel of Figure 2, the solid black curve gets the weight given in parentheses, and the dashed black curve gets one minus that weight. In case $\theta_2 = 0$, the probability of selecting model M_1 is very large, and the corresponding conditional density (solid curve) is the dominant factor in the unconditional density in Figure 1. For $\theta_2 = 0.1$, the situation is similar if slightly less pronounced. In case $\theta_2 = 0.75$, the solid and

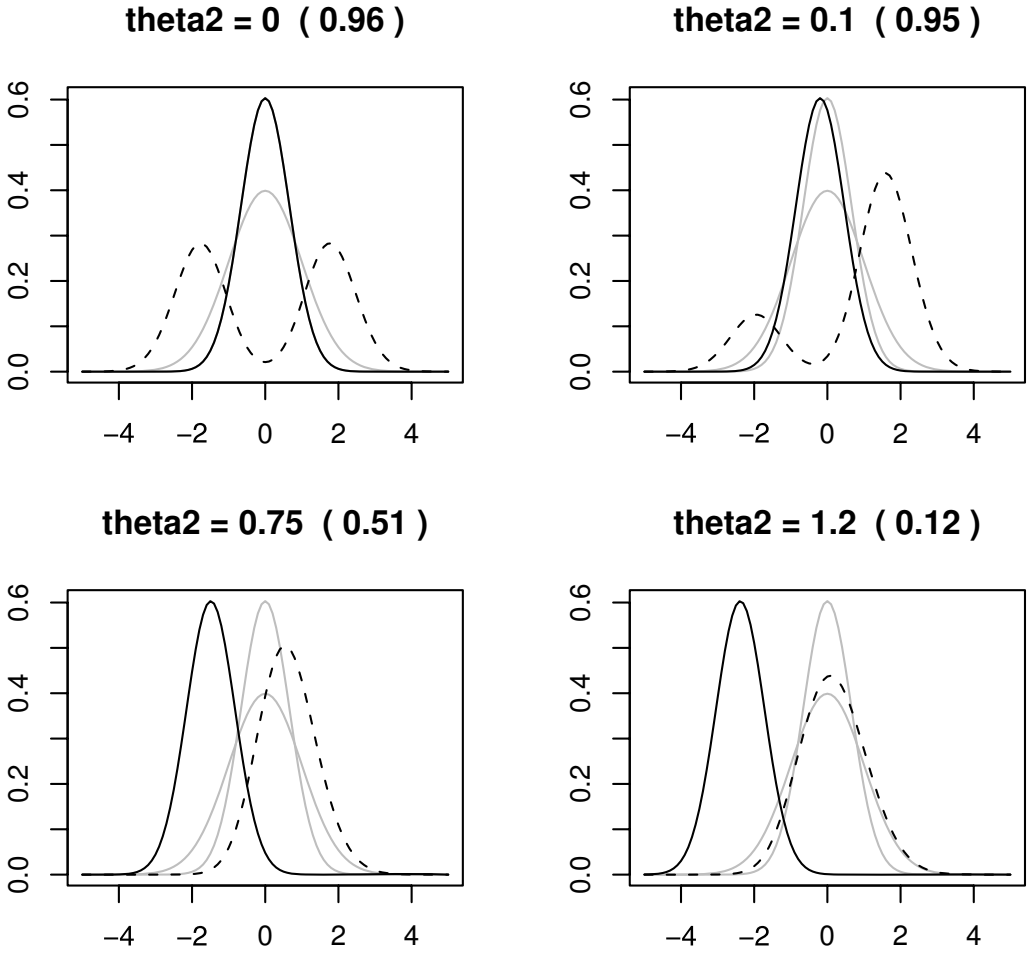


FIG 2. *The conditional density of* $\sqrt{n}(\tilde{\theta}_1^* - \theta_1)$*, conditional on selecting model* M_1 *(black solid line), and conditional on selecting model* M_2 *(black dashed line), for the same parameters as used for Figure* 1*. The number in parentheses in each panel header is the probability of selecting* M_1*, i.e.,* $\pi^*_{n,\theta,\sigma}(1)$*. The gray curves are as in Figure* 1*.*

the dashed curve in Figure 2 get approximately equal weight, i.e., 0.51 and 0.49, respectively, resulting in a bimodal unconditional density in Figure 1. Finally, in case $\theta_2 = 1.2$, the weight of the solid curve is 0.12 while that of the dashed curve is 0.88; the resulting unconditional density in Figure 1 is unimodal but has a 'hump' in the left tail. For a detailed discussion of the conditional distributions and densities themselves, we refer to Section 3.3 of Leeb [1].

Results similar to Figure 1 and Figure 2 can be obtained for any other sample size (by appropriate choice of θ_2 as noted above), and also for other choices of the critical value c_2 that is used by the model selectors. Larger values of c_2 result in model selectors that more strongly favor the smaller model M_1, and for which the phenomena observed above are more pronounced (see also Section 2.1 of Leeb and Pötscher [5] for results on the case where the critical value increases with sample size). Concerning the correlation coefficient ρ, we find that the shape of the conditional and of the unconditional densities is very strongly influenced by the magnitude of $|\rho|$, which we have chosen as $\rho = 0.75$ in figures 1 and 2 above. For larger values of $|\rho|$ we get similar but more pronounced phenomena. As $|\rho|$ gets smaller, however, these phenomena tend to be less pronounced. For example, if we plot the unconditional densities as in Figure 1 but with $\rho = 0.25$, we get four rather similar curves which altogether roughly resemble a Gaussian density except for some skewness. This is in line with the observation made in Section 3.3.1 that the unconditional distributions are Gaussian in the special case where $\tilde{\theta}_p(p)$ is uncorrelated with $A\tilde{\theta}(p)$ for each $p = \mathcal{O}+1, \ldots, P$. In the simple setting considered here, we have, in particular, that the distribution of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ is Gaussian in the special case where $\rho = 0$.

4. An approximation result

In Theorem 4.2 below, we show that $G^*_{n,\theta,\sigma}(t)$ is close to $G_{n,\theta,\sigma}(t)$ in large samples, uniformly in the underlying parameters, where closeness is with respect to the total variation distance. (A similar result is provided in Leeb [1] for the conditional cdfs under slightly stronger assumptions.) Theorem 4.2 will be instrumental in the large-sample analysis in Section 5, because the large-sample behavior of $G^*_{n,\theta,\sigma}(t)$ is significantly easier to analyze. The total variation distance of two cdfs G and G^* on $\mathbb{R}^k$ will be denoted by $||G-G^*||_{TV}$ in the following. (Note that the relation $|G(t)-G^*(t)| \leq ||G-G^*||_{TV}$ always holds for each $t \in \mathbb{R}^k$. Thus, if G and G^* are close with respect to the total variation distance, then $G(t)$ is close to $G^*(t)$, uniformly in t. We shall use the total variation distance also for distribution functions G and G^* which are not necessarily normalized, i.e., in the case where G and G^* are the distribution functions of finite measures with total mass possibly different from one.)

Since the unconditional cdfs $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$ can be linearly expanded in terms of $G_{n,\theta,\sigma}(t|p)\pi_{n,\theta,\sigma}(p)$ and $G^*_{n,\theta,\sigma}(t|p)\pi^*_{n,\theta,\sigma}(p)$, respectively, a key step for the results in this section is the following lemma.

Lemma 4.1. *For each p, $\mathcal{O} \leq p \leq P$, we have*

$$\sup_{\substack{\theta\in\mathbb{R}^P \\ \sigma>0}} \left|\left|G_{n,\theta,\sigma}(\cdot|p)\pi_{n,\theta,\sigma}(p) - G^*_{n,\theta,\sigma}(\cdot|p)\pi^*_{n,\theta,\sigma}(p)\right|\right|_{TV} \overset{n\to\infty}{\longrightarrow} 0. \tag{4.1}$$

This lemma immediately leads to the following result.

Theorem 4.2. *For the unconditional cdfs $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$ we have*

$$\sup_{\substack{\theta\in\mathbb{R}^P\\ \sigma>0}} \left|\left|G_{n,\theta,\sigma} - G^*_{n,\theta,\sigma}\right|\right|_{TV} \overset{n\to\infty}{\longrightarrow} 0. \tag{4.2}$$

Moreover, for each p satisfying $\mathcal{O} \le p \le P$, the model selection probabilities $\pi_{n,\theta,\sigma}(p)$ and $\pi^*_{n,\theta,\sigma}(p)$ satisfy

$$\sup_{\substack{\theta\in\mathbb{R}^P\\ \sigma>0}} \left|\pi_{n,\theta,\sigma}(p) - \pi^*_{n,\theta,\sigma}(p)\right| \overset{n\to\infty}{\longrightarrow} 0.$$

By Theorem 4.2 we have, in particular, that

$$\sup_{\substack{\theta\in\mathbb{R}^P\\ \sigma>0}} \sup_{t\in\mathbb{R}^k} \left|G_{n,\theta,\sigma}(t) - G^*_{n,\theta,\sigma}(t)\right| \overset{n\to\infty}{\longrightarrow} 0;$$

that is, the cdf $G_{n,\theta,\sigma}(t)$ is closely approximated by $G^*_{n,\theta,\sigma}(t)$ if n is sufficiently large, uniformly in the argument t and uniformly in the parameters θ and σ. The result in Theorem 4.2 does not depend on the scaling factor $\sqrt{n}$ and on the centering constant $A\theta$ that are used in the definitions of $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$, cf. (2.4) and (2.5), respectively. In fact, that result continues to hold for arbitrary measurable transformations of $\tilde{\theta}$ and $\tilde{\theta}^*$. (See Corollary A.1 below for a precise formulation.)

Leeb [1] gives a result paralleling (4.2) for the conditional distributions of $A\tilde{\theta}$ and $A\tilde{\theta}^*$, conditional on the outcome of the model selection step. That result establishes closeness of the corresponding conditional cdfs uniformly not over the whole parameter space but over a slightly restricted set of parameters; cf. Theorem 4.1 in Leeb [1]. This restriction arose from the need to control the behavior of ratios of probabilities which vanish asymptotically. (Indeed, the probability of selecting the model of order p converges to zero as $n \to \infty$ if the selected model is incorrect; cf. (5.7) below.) In the unconditional case considered in Theorem 4.2 above, this difficulty does not arise, allowing us to avoid this restriction.

5. Asymptotic results for the unconditional distributions and for the selection probabilities

We now analyze the large-sample limit behavior of $G_{n,\theta,\sigma}(t)$ and $G^*_{n,\theta,\sigma}(t)$, both in the fixed parameter case where θ and σ are kept fixed while n goes to infinity, and along sequences of parameters $\theta^{(n)}$ and $\sigma^{(n)}$. The main result in this section is Proposition 5.1 below. Inter alia, this result gives a complete characterization of all accumulation points of the unconditional cdfs (with respect to weak convergence) along sequences of parameters; cf. Remark 5.5. Our analysis also includes the model selection probabilities, as well as the case of local-alternative and fixed-parameter asymptotics.

The following conventions will be employed throughout this section: For p satisfying $0 < p \le P$, partition $Q = \lim_{n\to\infty} X'X/n$ as

$$Q = \begin{pmatrix} Q[p:p] & Q[p:\neg p] \\ Q[\neg p:p] & Q[\neg p:\neg p] \end{pmatrix},$$

where $Q[p:p]$ is a $p \times p$ matrix. Let $\Phi_{\infty,p}(t)$ be the cdf of a k-variate centered Gaussian random vector with covariance matrix $\sigma^2 A[p]Q[p:p]^{-1}A[p]'$, $0 < p \le P$,

and let $\Phi_{\infty,0}(t)$ denote the cdf of point-mass at zero in $\mathbb{R}^k$. Note that $\Phi_{\infty,p}(t)$ has a density with respect to Lebesgue measure on $\mathbb{R}^k$ if $p > 0$ and the matrix $A[p]$ has rank k; in this case, we denote the Lebesgue density of $\Phi_{\infty,p}(t)$ by $\phi_{\infty,p}(t)$. Finally, for $p = 1, \ldots, P$, define the quantities

$$\begin{aligned}
\xi^2_{\infty,p} &= (Q[p:p]^{-1})_{p,p}, \\
\zeta^2_{\infty,p} &= \xi^2_{\infty,p} - C^{(p)\prime}_{\infty}(A[p]Q[p:p]^{-1}A[p]')^{-}C^{(p)}_{\infty}, \text{ and} \\
b_{\infty,p} &= C^{(p)\prime}_{\infty}(A[p]Q[p:p]^{-1}A[p]')^{-},
\end{aligned}$$

where $C^{(p)}_{\infty} = A[p]Q[p:p]^{-1}e_p$, with e_p denoting the p-th standard basis vector in $\mathbb{R}^p$. As the notation suggests, $\Phi_{\infty,p}(t)$ is the large-sample limit of $\Phi_{n,p}(t)$, $C^{(p)}_{\infty}$, $\xi_{\infty,p}$ and $\zeta_{\infty,p}$ are the limits of $C^{(p)}_{n}$, $\xi_{n,p}$ and $\zeta_{n,p}$, respectively, and $b_{n,p}z \to b_{\infty,p}z$ for each z in the column-space of $A[p]$; cf. Lemma A.2 in Leeb [1]. With these conventions, we can characterize the large-sample limit behavior of the unconditional cdfs along sequences of parameters.

Proposition 5.1. *Consider sequences of parameters $\theta^{(n)} \in \mathbb{R}^P$ and $\sigma^{(n)} > 0$, such that $\sqrt{n}\theta^{(n)}$ converges to a limit $\psi \in (\mathbb{R} \cup \{-\infty, \infty\})^P$, and such that $\sigma^{(n)}$ converges to a (finite) limit $\sigma > 0$ as $n \to \infty$. Let p_* denote the largest index p, $\mathcal{O} < p \leq P$, for which $|\psi_p| = \infty$, and set $p_* = \mathcal{O}$ if no such index exists. Then $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t)$ and $G_{n,\theta^{(n)},\sigma^{(n)}}(t)$ both converge weakly to a limit cdf which is given by*

$$\begin{aligned}
&\Phi_{\infty,p_*}(t - A\delta^{(p_*)}) \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\delta^{(q)}_q + \psi_q, c_q\sigma\xi_{\infty,q}) \\
&+ \sum_{p=p_*+1}^{P} \int_{z \leq t - A\delta^{(p)}} \Big(1 - \Delta_{\sigma\zeta_{\infty,p}}(\delta^{(p)}_p + \psi_p + b_{\infty,p}z,\, c_p\sigma\xi_{\infty,p})\Big)\, \Phi_{\infty,p}(dz) \\
&\qquad \times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\delta^{(q)}_q + \psi_q, c_q\sigma\xi_{\infty,q}),
\end{aligned} \tag{5.2}$$

where

$$\delta^{(p)} = \begin{pmatrix} Q[p:p]^{-1}Q[p:\neg p] \\ -I_{P-p} \end{pmatrix} \psi[\neg p], \tag{5.3}$$

$p_* \leq p \leq P$ (with the convention that $\delta^{(P)}$ is the zero-vector in $\mathbb{R}^P$ and, if necessary, that $\delta^{(0)} = -\psi$). Note that $\delta^{(p)}$ is the limit of the bias of $\tilde{\theta}(p)$ scaled by $\sqrt{n}$, i.e., $\delta^{(p)} = \lim_{n\to\infty} \sqrt{n}(\eta_n(p) - \theta^{(n)})$, with $\eta_n(p)$ given by (3.1) with $\theta^{(n)}$ replacing θ; also note that $\delta^{(p)}$ is always finite, $p_* \leq p \leq P$.

The above statement continues to hold with convergence in total variation replacing weak convergence in the case where $p_* > 0$ and the matrix $A[p_*]$ has rank k, and in the case where $p_* < P$ and $\sqrt{n}A[\neg p_*]\theta^{(n)}[\neg p_*]$ is constant in n.

Remark 5.2. Observe that the limit cdf in (5.2) is of a similar form as the finite-sample cdf $G^*_{n,\theta,\sigma}(t)$ as given in (3.7) (the only difference being that the right-hand side of (3.7) is the sum of $P - \mathcal{O} + 1$ terms while (5.2) is the sum of $P - p_* + 1$ terms, that quantities depending on the regressor matrix through $X'X/n$ in (3.7) are replaced by their corresponding limits in (5.2), and that the bias and mean of $\sqrt{n}\tilde{\theta}(p)$ in (3.7) are replaced by the appropriate large-sample limits in (5.2)). Therefore, the discussion of the finite-sample cdf $G^*_{n,\theta,\sigma}(t)$ given in Section 3.3

applies, mutatis mutandis, also to the limit cdf in (5.2). In particular, the cdf in (5.2) has a density with respect to Lebesgue measure on $\mathbb{R}^k$ if (and only if) $p_* > 0$ and $A[p_*]$ has rank k; in that case, this density can be obtained from (5.2) by differentiation. Moreover, we stress that the limit cdf is typically non-Gaussian. A notable exception where (5.2) reduces to the Gaussian cdf $\Phi_{\infty,P}(t)$ occurs in the special case where $\tilde{\theta}_q(q)$ and $A\tilde{\theta}(q)$ are asymptotically uncorrelated for each $q = p_* + 1, \ldots, P$.

Inspecting the proof of Proposition 5.1, we also obtain the large-sample limit behavior of the conditional cdfs weighted by the model selection probabilities, e.g., of $G_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$ (weak convergence of not necessarily normalized cdfs H_n to a not necessarily normalized cdf H on $\mathbb{R}^k$ is defined as follows: $H_n(t)$ converges to $H(t)$ at each continuity point t of the limit cdf, and $H_n(\mathbb{R}^k)$, i.e., the total mass of H_n on $\mathbb{R}^k$, converges to $H(\mathbb{R}^k)$).

Corollary 5.3. *Assume that the assumptions of Proposition* 5.1 *are met, and fix p with $\mathcal{O} \le p \le P$. In case $p = p_*$, $G_{n,\theta^{(n)},\sigma^{(n)}}(t|p_*)\pi_{n,\theta^{(n)},\sigma^{(n)}}(p_*)$ converges to the first term in* (5.2) *in the sense of weak convergence. If $p > p_*$, $G_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges weakly to the term with index p in the sum in* (5.2)*. Finally, if $p < p_*$, $G_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to zero in total variation. The same applies to $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$. Moreover, weak convergence can be strengthened to convergence in total variation in the case where $p > 0$ and $A[p]$ has rank k (in that case, the weighted conditional cdf also has a Lebesgue density), and in the case where $p < P$ and $\sqrt{n}A[\neg p]\theta^{(n)}[\neg p]$ is constant in n.*

Proposition 5.4. *Under the assumptions of Proposition* 5.1*, the large-sample limit behavior of the model selection probabilities $\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$, $\mathcal{O} \le p \le P$, is as follows: For each p satisfying $p_* < p \le P$, $\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to*

$$(1 - \Delta_{\sigma\xi_{\infty,p}}(\delta_p^{(p)} + \psi_p,\, c_p\sigma\xi_{\infty,p})) \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)} + \psi_q, c_q\sigma\xi_{\infty,q}). \tag{5.4}$$

For $p = p_*$, $\pi_{n,\theta^{(n)},\sigma^{(n)}}(p_*)$ converges to

$$\prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)} + \psi_q, c_q\sigma\xi_{\infty,q}). \tag{5.5}$$

For each p satisfying $\mathcal{O} \le p < p_*$, $\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to zero. The above statements continue to hold with $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ replacing $\pi_{n,\theta^{(n)},\sigma^{(n)}}(p)$.

Remark 5.5. With Propositions 5.1 and 5.4 we obtain a complete characterization of all possible accumulation points of the unconditional cdfs (with respect to weak convergence) and of the model selection probabilities, along arbitrary sequences of parameters $\theta^{(n)}$ and $\sigma^{(n)}$, provided that $\sigma^{(n)}$ is bounded away from zero and infinity: Let $\theta^{(n)}$ be any sequence in $\mathbb{R}^P$ and let $\sigma^{(n)}$ be a sequence satisfying $\sigma_* \le \sigma^{(n)} \le \sigma^*$ with $0 < \sigma_* \le \sigma^* < \infty$. Since the set $(\mathbb{R} \cup \{-\infty, \infty\})^P$ as well as the set $[\sigma_*, \sigma^*]$ is compact, each subsequence contains a further subsequence for which the assumptions of Propositions 5.1 and 5.4 are satisfied. For example, each accumulation point of $G_{n,\theta^{(n)},\sigma^{(n)}}(t)$ (with respect to weak convergence) is of the form (5.2), where here ψ and σ are accumulation points of $\sqrt{n}\theta^{(n)}$ and $\sigma^{(n)}$, respectively (and where p_* and the quantities $\delta^{(p)}$, $p_* \le p \le P$, are derived from

ψ as in Proposition 5.1). Of course, the same is true for $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t)$. The same considerations apply, mutatis mutandis, to the weighted conditional cdfs considered in Corollary 5.3.

To study, say, the large-sample limit minimal coverage probability of confidence sets for $A\theta$ centered at $A\tilde{\theta}$, a description of all possible accumulation points of $G_{n,\theta^{(n)},\sigma^{(n)}}(t)$ with respect to weak convergence is useful; here $\theta^{(n)}$ can be any sequence in $\mathbb{R}^P$ and $\sigma^{(n)}$ can be any sequence bounded away from zero and infinity. In view of Remark 5.5, we see that each individual accumulation point can be reached along a particular sequence of regression parameters $\theta^{(n)}$, chosen such that the $\theta^{(n)}$ are within an $O(1/\sqrt{n})$ neighborhood of one of the models under consideration, say, M_{p_*} for some $\mathcal{O} \le p_* \le P$. In particular, in order to describe all possible accumulation points of the unconditional cdf, it suffices to consider local alternatives to θ.

Corollary 5.6. *Fix $\theta \in \mathbb{R}^P$ and consider local alternatives of the form $\theta + \gamma/\sqrt{n}$, where $\gamma \in \mathbb{R}^P$. Moreover, let $\sigma^{(n)}$ be a sequence of positive real numbers converging to a (finite) limit $\sigma > 0$. Then Propositions* 5.1 *and* 5.4 *apply with $\theta + \gamma/\sqrt{n}$ replacing $\theta^{(n)}$, where here p_* equals $\max\{p_0(\theta), \mathcal{O}\}$ and $\psi[\neg p_*]$ equals $\gamma[\neg p_*]$ (in case $p_* < P$). In particular, $G^*_{n,\theta+\gamma/\sqrt{n},\sigma^{(n)}}(t)$ and $G_{n,\theta+\gamma/\sqrt{n},\sigma^{(n)}}(t)$ converge in total variation to the cdf in* (5.2) *with $p_* = \max\{p_0(\theta), \mathcal{O}\}$.*

In the case of fixed-parameter asymptotics, the large-sample limits of the model selection probabilities and of the unconditional cdfs take a particularly simple form. Fix $\theta \in \mathbb{R}^P$ and $\sigma > 0$. Clearly, $\sqrt{n}\theta$ converges to a limit ψ, whose $p_0(\theta)$-th component is infinite if $p_0(\theta) > 0$ (because the $p_0(\theta)$-th component of θ is non-zero in that case), and whose p-th component is zero for each $p > p_0(\theta)$. Therefore, Propositions 5.1 and 5.4 apply with $p_* = \max\{p_0(\theta), \mathcal{O}\}$, and either with $p_* < P$ and $\psi[\neg p_*] = (0, \ldots, 0)'$, or with $p_* = P$. In particular, $p_* = \max\{p_0(\theta), \mathcal{O}\}$ is the order of the smallest correct model for θ among the candidate models $M_{\mathcal{O}}, \ldots, M_P$. We hence obtain that $G^*_{n,\theta,\sigma}(t)$ and $G_{n,\theta,\sigma}(t)$ converge in total variation to the cdf

$$\begin{aligned} &\Phi_{\infty,p_*}(t) \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q}) \\ &\quad + \sum_{p=p_*+1}^{P} \int_{z\le t} (1 - \Delta_{\sigma\zeta_{\infty,p}}(b_{\infty,p}z, c_p\sigma\xi_{\infty,p}))\Phi_{\infty,p}(dz) \\ &\quad \times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q}), \end{aligned} \tag{5.6}$$

and the large-sample limit of the model selection probabilities $\pi_{n,\theta,\sigma}(p)$ and $\pi^*_{n,\theta,\sigma}(p)$ for $\mathcal{O} \le p \le P$ is given by

$$\begin{cases} (1 - \Delta_{\sigma\xi_{\infty,p}}(0, c_p\sigma\xi_{\infty,p})) \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q}) & \text{if } p > p_*, \\ \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q}) & \text{if } p = p_*, \\ 0 & \text{if } p < p_* \end{cases} \tag{5.7}$$

with $p_* = \max\{p_0(\theta), \mathcal{O}\}$.

Remark 5.7. (i) In defining the cdf $G_{n,\theta,\sigma}(t)$, the estimator has been centered at θ and scaled by $\sqrt{n}$; cf. (2.4). For the finite-sample results in Section 3, a different choice of centering constant (or scaling factor) of course only amounts to a translation (or rescaling) of the distribution and is hence inconsequential. Also, the results in Section 4 do not depend on the centering constant and on the scaling factor, because the total variation distance of two cdfs is invariant under a shift or rescaling of the argument. More generally, Lemma 4.1 and Theorem 4.2 extend to the distribution of arbitrary (measurable) functions of $\tilde{\theta}$ and $\tilde{\theta}^*$; cf. Corollary A.1 below.

(ii) We are next concerned with the question to which extent the limiting results given in the current section are affected by the choice of the centering constant. Let $d_{n,\theta,\sigma}$ denote a $P \times 1$ vector which may depend on n, θ and σ. Then centering at $d_{n,\theta,\sigma}$ leads to

$$P_{n,\theta,\sigma}\left(\sqrt{n}A(\tilde{\theta} - d_{n,\theta,\sigma}) \leq t\right) = G_{n,\theta,\sigma}\left(t + \sqrt{n}A(d_{n,\theta,\sigma} - \theta)\right). \tag{5.8}$$

The results obtained so far can now be used to describe the large-sample behavior of the cdf in (5.8). In particular, assuming that $\sqrt{n}A(d_{n,\theta,\sigma}-\theta)$ converges to a limit $\nu \in \mathbb{R}^k$, it is easy to verify that the large-sample limit of the cdf in (5.8) (in the sense of weak convergence) is given by the cdf in (5.6) with $t + \nu$ replacing t. If $\sqrt{n}A(d_{n,\theta,\sigma} - \theta)$ converges to a limit $\nu \in (\mathbb{R} \cup \{-\infty, \infty\})^k$ with some component of ν being either ∞ or $-\infty$, then the limit of (5.8) will be degenerate in the sense that at least one marginal distribution mass will have escaped to ∞ or $-\infty$. In other words, if i is such that $|\nu_i| = \infty$, then the i-th component of $\sqrt{n}A(\tilde{\theta} - d_{n,\theta,\sigma})$ converges to $-\nu_i$ in probability as $n \to \infty$. The marginal of (5.8) corresponding to the finite components of ν converges weakly to the corresponding marginal of (5.6) with the appropriate components of $t + \nu$ replacing the appropriate components of t. This shows that, for an asymptotic analysis, any reasonable centering constant typically must be such that $Ad_{n,\theta,\sigma}$ coincides with $A\theta$ up to terms of order $O(1/\sqrt{n})$. If $\sqrt{n}A(d_{n,\theta,\sigma} - \theta)$ does not converge, accumulation points can be described by considering appropriate subsequences. The same considerations apply to the cdf $G^*_{n,\theta,\sigma}(t)$, and also to asymptotics along sequences of parameters $\theta^{(n)}$ and $\sigma^{(n)}$.

Acknowledgments

I am thankful to Benedikt M. Pötscher for helpful remarks and discussions.

Appendix A: Proofs for Section 4

Proof of Lemma 4.1. Consider first the case where $p > \mathcal{O}$. In that case, it is easy to see that $G_{n,\theta,\sigma}(t|p)\pi_{n,\theta,\sigma}(p)$ does not depend on the critical values c_q for $q < p$ which are used by the model selection procedure $\hat{p}$ (cf. formula (3.9) above for $\pi_{n,\theta,\sigma}(p)$ and the expression for $G_{n,\theta,\sigma}(t|p)$ given in (16)–(18) of Leeb [1]). As a consequence, we conclude for $p > \mathcal{O}$ that $G_{n,\theta,\sigma}(t|p)\pi_{n,\theta,\sigma}(p)$ follows the same formula irrespective of whether $\mathcal{O} = 0$ or $\mathcal{O} > 0$. The same applies, mutatis mutandis, to $G^*_{n,\theta,\sigma}(t|p)\pi^*_{n,\theta,\sigma}(t)$. We hence may assume that $\mathcal{O} = 0$ in the following.

In the special case where A is the $p \times P$ matrix $(I_p : 0)$ (which is to be interpreted as I_P in case $p = P$), (4.1) follows from Lemma 5.1 of Leeb and Pötscher [3]. (In that result the conditional cdfs are such that the estimators are centered at $\eta_n(p)$ instead of θ. However, this different centering constant does not affect the

total variation distance; cf. Lemma A.5 in Leeb [1].) For the case of general A, write μ as shorthand for the conditional distribution of $\sqrt{n}(I_p : 0)(\tilde{\theta} - \theta)$ given $\hat{p} = p$ multiplied by $\pi_{n,\theta,\sigma}(p)$, μ^* as shorthand for the conditional distribution of $\sqrt{n}(I_p : 0)(\tilde{\theta}^* - \theta)$ given $\hat{p}^* = p$ multiplied by $\pi^*_{n,\theta,\sigma}(p)$, and let Ψ denote the mapping $z \mapsto ((A[p]z)' : (-\sqrt{n}A[\neg p]\theta[\neg p])')'$ in case $p < P$ and $z \mapsto Az$ in case $p = P$. It is now easy to see that Lemma A.5 of Leeb [1] applies, and (4.1) follows.

It remains to show that (4.1) also holds with $\mathcal{O}$ replacing p. Having established (4.1) for $p > \mathcal{O}$, it also follows, for each $p = \mathcal{O} + 1, \ldots, P$, that

$$\sup_{\substack{\theta \in \mathbb{R}^P \\ \sigma > 0}} \left| \pi_{n,\theta,\sigma}(p) - \pi^*_{n,\theta,\sigma}(p) \right| \overset{n \to \infty}{\longrightarrow} 0, \tag{A.1}$$

because the modulus in (A.1) is bounded by

$$||G_{n,\theta,\sigma}(\cdot|p)\pi_{n,\theta,\sigma}(p) - G^*_{n,\theta,\sigma}(\cdot|p)\pi^*_{n,\theta,\sigma}(p)||_{TV}.$$

Since the model selection probabilities sum up to one, we have $\pi_{n,\theta,\sigma}(\mathcal{O}) = 1 - \sum_{p=\mathcal{O}+1}^{P} \pi_{n,\theta,\sigma}(p)$, and a similar expansion holds for $\pi^*_{n,\theta,\sigma}(\mathcal{O})$. By this and the triangle inequality, we see that (A.1) also holds with $\mathcal{O}$ replacing p. Now (4.1) with $\mathcal{O}$ replacing p follows immediately, because the conditional cdfs $G_{n,\theta,\sigma}(t|\mathcal{O})$ and $G^*_{n,\theta,\sigma}(t|\mathcal{O})$ are both equal to $\Phi_{n,\mathcal{O}}(t - \sqrt{n}A(\eta_n(\mathcal{O}) - \theta))$, cf. (10) and (14) of Leeb [1], which is of course bounded by one. □

Proof of Theorem 4.2. Relation (4.2) follows from Lemma 4.1 by expanding $G^*_{n,\theta,\sigma}(t)$ as in (3.4), by expanding $G_{n,\theta,\sigma}(t)$ in a similar fashion, and by applying the triangle inequality. The statement concerning the model selection probabilities has already been established in the course of the proof of Lemma 4.1; cf. (A.1) and the attending discussion. □

Corollary A.1. *For each n, θ and σ, let $\Psi_{n,\theta,\sigma}(\cdot)$ be a measurable function on $\mathbb{R}^P$. Moreover, let $R_{n,\theta,\sigma}(\cdot)$ denote the distribution of $\Psi_{n,\theta,\sigma}(\tilde{\theta})$, and let $R^*_{n,\theta,\sigma}(\cdot)$ denote the distribution of $\Psi_{n,\theta,\sigma}(\tilde{\theta}^*)$. (That is, say, $R_{n,\theta,\sigma}(\cdot)$ is the probability measure induced by $\Psi_{n,\theta,\sigma}(\tilde{\theta})$ under $P_{n,\theta,\sigma}(\cdot)$.) We then have*

$$\sup_{\substack{\theta \in \mathbb{R}^P \\ \sigma > 0}} \left|\left| R_{n,\theta,\sigma}(\cdot) - R^*_{n,\theta,\sigma}(\cdot) \right|\right|_{TV} \overset{n \to \infty}{\longrightarrow} 0. \tag{A.2}$$

Moreover, if $R_{n,\theta,\sigma}(\cdot|p)$ and $R^*_{n,\theta,\sigma}(\cdot|p)$ denote the distributions of $\Psi_{n,\theta,\sigma}(\tilde{\theta})$ conditional on $\hat{p} = p$ and of $\Psi_{n,\theta,\sigma}(\tilde{\theta}^*)$ conditional on $\hat{p}^* = p$, respectively, then

$$\sup_{\substack{\theta \in \mathbb{R}^P \\ \sigma > 0}} \left|\left| R_{n,\theta,\sigma}(\cdot|p)\pi_{n,\theta,\sigma}(p) - R^*_{n,\theta,\sigma}(\cdot|p)\pi^*_{n,\theta,\sigma}(p) \right|\right|_{TV} \overset{n \to \infty}{\longrightarrow} 0. \tag{A.3}$$

Proof. Observe that the total variation distance of two cdfs is unaffected by a change of scale or a shift of the argument. Using Theorem 4.2 with $A = I_P$, we hence obtain that (A.2) holds if $\Psi_{n,\theta,\sigma}$ is the identity map. From this, the general case follows immediately in view of Lemma A.5 of Leeb [1]. In a similar fashion, (A.3) follows from Lemma 4.1. □

Appendix B: Proofs for Section 5

Under the assumptions of Proposition 5.1, we make the following preliminary observation: For $p \geq p_*$, consider the scaled bias of $\tilde{\theta}(p)$, i.e., $\sqrt{n}(\eta_n(p) - \theta^{(n)})$, where $\eta_n(p)$ is defined as in (3.1) with $\theta^{(n)}$ replacing θ. It is easy to see that

$$\sqrt{n}(\eta_n(p) - \theta^{(n)}) \;=\; \begin{pmatrix} (X[p]'X[p])^{-1}X[p]'X[\neg p] \\ -I_{P-p} \end{pmatrix} \sqrt{n}\theta^{(n)}[\neg p],$$

where the expression on the right-hand side is to be interpreted as $\sqrt{n}\theta^{(n)}$ and as the zero vector in $\mathbb{R}^P$ in the cases $p = 0$ and $p = P$, respectively. For p satisfying $p_* \leq p < P$, note that $\sqrt{n}\theta^{(n)}[\neg p]$ converges to $\psi[\neg p]$ by assumption, and that this limit is finite by choice of $p \geq p_*$. It hence follows that $\sqrt{n}(\eta_n(p) - \theta^{(n)})$ converges to the limit $\delta^{(p)}$ given in (5.3). From this, we also see that $\sqrt{n}\eta_{n,p}(p)$ converges to $\delta_p^{(p)} + \psi_p$, which is finite for each $p > p_*$; for $p = p_*$, this limit is infinite in case $|\psi_{p_*}| = \infty$. Note that the case where the limit of $\sqrt{n}\eta_{n,p_*}(p_*)$ is finite can only occur if $p_* = \mathcal{O}$. It will now be convenient to prove Proposition 5.4 first.

Proof of Proposition 5.4. In view of Theorem 4.2, it suffices to consider $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$. This model selection probability can be expanded as in (3.5)–(3.6) with $\theta^{(n)}$ and $\sigma^{(n)}$ replacing θ and σ, respectively. Consider first the individual Δ-functions occurring in these formulas, i.e.,

$$\Delta_{\sigma^{(n)}\xi_{n,q}}(\sqrt{n}\eta_{n,q}(q), c_q\sigma^{(n)}\xi_{n,q}), \tag{B.1}$$

$\mathcal{O} < q \leq P$. For $q > p_*$, recall that $\sqrt{n}\eta_{n,q}(q)$ converges to the finite limit $\delta_q^{(q)} + \psi_q$ as we have seen above, and it is elementary to verify that the expression in (B.1) converges to $\Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)} + \psi_q, c_q\sigma\xi_{\infty,q})$. For $q = p_*$ and $p_* > \mathcal{O}$, we have seen that the limit of $\sqrt{n}\eta_{n,p_*}(p_*)$ is infinite, and it is easy to see that (B.1) with p_* replacing q converges to zero in this case.

From the above considerations, it immediately follows that $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to the limit in (5.4) if $p > p_*$, and to the limit in (5.5) if $p = p_*$. To show that $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to zero in case p satisfies $\mathcal{O} \leq p < p_*$, it suffices to observe that here $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ is bounded by the expression in (B.1) with p_* replacing q. As we have seen above, $\sqrt{n}|\eta_{n,p_*}(p_*)|$ converges to infinity, such that this upper bound converges to zero as $n \to \infty$. □

Proof of Proposition 5.1. Again, it suffices to consider $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t)$ in view of Theorem 4.2. Recall that this cdf can be written as in (3.4). We first consider the individual terms $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ for $p = \mathcal{O}, \ldots, P$. In case p satisfies $\mathcal{O} \leq p < p_*$, note that $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p) \to 0$ by Proposition 5.4. Hence, $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges to zero in total variation.

In the remaining cases, i.e., for p satisfying $p_* \leq p \leq P$, it is elementary to verify that Proposition 5.1 of Leeb [1] applies to $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)$, where the quantity β in that paper equals $A\delta^{(p)}$ in our setting. In particular, that result gives the limit of the conditional cdf in the sense of weak convergence (because $\delta^{(p)}$ is finite). Consider first the case $p > p_*$. From Proposition 5.4, we obtain the limit of $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$. Combining the resulting limit expression with the limit expression for $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)$ as obtained by Proposition 5.1 of Leeb [1], we see that

$G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$ converges weakly to

$$\int_{\substack{z\in\mathbb{R}^k \\ z\le t-A\delta^{(p)}}} \Big(1-\Delta_{\sigma\zeta_{\infty,p}}(\delta_p^{(p)}+\psi_p+b_{\infty,p}z,\, c_p\sigma\xi_{\infty,p})\Big)\,\Phi_{\infty,p}(dz) \times \prod_{q=p+1}^{P}\Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)}+\psi_q, c_q\sigma\xi_{\infty,q}). \tag{B.2}$$

In case $p=p_*$ and $p_*>\mathcal{O}$, we again use Proposition 5.1 of Leeb [1] and Proposition 5.4 to obtain that the weak limit of $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p_*)\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p_*)$ is of the form (B.2) with p_* replacing p. Since $|\psi_{p_*}|$ is infinite, the integrand in (B.2) reduces to one, i.e., the limit is given by

$$\Phi_{\infty,p_*}(t-A\delta^{(p_*)})\prod_{q=p_*+1}^{P}\Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)}+\psi_q, c_q\sigma\xi_{\infty,q}).$$

Finally, consider the case $p=p_*$ and $p_*=\mathcal{O}$. Arguing as above, we see that $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|\mathcal{O})\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(\mathcal{O})$ converges weakly to

$$\Phi_{\infty,\mathcal{O}}(t-A\delta^{(\mathcal{O})})\prod_{q=\mathcal{O}+1}^{P}\Delta_{\sigma\xi_{\infty,q}}(\delta_q^{(q)}+\psi_q, c_q\sigma\xi_{\infty,q}).$$

Because the individual model selection probabilities $\pi^*_{n,\theta^{(n)},\sigma^{(n)}}(p)$, $\mathcal{O}\le p\le P$, sum up to one, the same is true for their large-sample limits. In particular, note that (5.2) is a convex combination of cdfs, and that all the weights in the convex combination are positive. From this, we obtain that $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t)$ converges to the expression in (5.2) at each continuity point t of the limit expression, i.e., $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t)$ converges weakly. (Note that a convex combination of cdfs on $\mathbb{R}^k$ is continuous at a point t if each individual cdf is continuous at t; the converse is also true, provided that all the weights in the convex combination are positive.) To establish that weak convergence can be strengthened to convergence in total variation under the conditions given in Proposition 5.1, it suffices to note, under these conditions, that $G^*_{n,\theta^{(n)},\sigma^{(n)}}(t|p)$, $p_*\le p\le P$, converges not only weakly but also in total variation in view of Proposition 5.1 of Leeb [1]. □

References

[1] Leeb, H., (2005). The distribution of a linear predictor after model selection: conditional finite-sample distributions and asymptotic approximations. *J. Statist. Plann. Inference* **134**, 64–89.

[2] Leeb, H. and Pötscher, B. M., Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, to appear.

[3] Leeb, H. and Pötscher, B. M., (2003). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory* **19**, 100–142.

[4] Leeb, H. and Pötscher, B. M., (2005). Can one estimate the unconditional distribution of post-model-selection estimators? Manuscript.

[5] Leeb, H. and Pötscher, B. M., (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**, 21–59.

[6] Pötscher, B. M., (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163–185.

[7] Rao, C. R., (1973). *Linear Statistical Inference and Its Applications*, 2nd edition. John Wiley & Sons, New York.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 312–321
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000527

Local asymptotic minimax risk bounds in a locally asymptotically mixture of normal experiments under asymmetric loss

Debasis Bhattacharya[1] **and A. K. Basu**[2]

Visva-Bharati University and Calcutta University

Abstract: Local asymptotic minimax risk bounds in a locally asymptotically mixture of normal family of distributions have been investigated under asymmetric loss functions and the asymptotic distribution of the optimal estimator that attains the bound has been obtained.

1. Introduction

There are two broad issues in the asymptotic theory of inference: (i) the problem of finding the limiting distributions of various statistics to be used for the purpose of estimation, tests of hypotheses, construction of confidence regions etc., and (ii) problems associated with questions such as: how good are the estimation and testing procedures based on the statistics under consideration and how to define 'optimality', etc. Le Cam [12] observed that the satisfactory answers to the above questions involve the study of the asymptotic behavior of the likelihood ratios. Le Cam [12] introduced the concept of 'Limit Experiment', which states that if one is interested in studying asymptotic properties such as local asymptotic minimaxity and admissibility for a given sequence of experiments, it is enough to prove the result for the limit of the experiment. Then the corresponding limiting result for the sequence of experiments will follow.

One of the many approaches which are used in asymptotic theory to judge the performance of an estimator is to measure the risk of estimation under an appropriate loss function. The idea of comparing estimators by comparing the associated risks was considered by Wald [19, 20]. Later this idea has been discussed by Hájek [8], Ibragimov and Has'minskii [9] and others. The concept of studying asymptotic efficiency based on large deviations has been recommended by Basu [4] and Bahadur [1, 2]. In the above context it is an interesting problem to obtain a lower bound for the risk in a wide class of competing estimators and then find an estimator which attains the bound. Le Cam [11] obtained several basic results concerning asymptotic properties of risk functions for LAN family of distributions. Jeganathan [10], Basawa and Scott [5], and Le Cam and Yang [13] have extended the results of Le Cam for Locally Asymptotically Mixture of Normal (LAMN) experiments. Basu and Bhattacharya [3] further extended the result for Locally Asymptotically Quadratic (LAQ) family of distributions. A symmetric loss structure (for example,

[1]Division of Statistics, Institute of Agriculture, Visva-Bharati University, Santiniketan, India, Pin 731236

[2]Department of Statistics, Calcutta University, 35 B. C. Road, Calcutta, India, Pin 700019

AMS 2000 subject classifications: primary 62C20, 62F12; secondary 62E20, 62C99.

Keywords and phrases: locally asymptotically mixture of normal experiment, local asymptotic minimax risk bound, asymmetric loss function

squared error loss) has been used to derive the results in the above mentioned references. But there are situations where the loss can be different for equal amounts of over-estimation and under-estimation, e. g., there exists a natural imbalance in the economic results of estimation errors of the same magnitude and of opposite signs. In such cases symmetric losses may not be appropriate. In this context Bhattacharya et al. [7], Levine and Bhattacharya [15], Rojo [16], Zellner [21] and Varian [18] may be referred to. In these works the authors have used an asymmetric loss, known as the LINEX loss function. Let $\Delta = \hat{\theta} - \theta$, $a \neq 0$ and $b > 0$. The LINEX loss is then defined as:

$$l(\Delta) = b[\exp(a\Delta) - a\Delta - 1]. \tag{1.1}$$

Other types of asymmetric loss functions that can be found in the literature are as follows:

$$l(\Delta) = \begin{cases} C_1\Delta, \text{ for } \Delta \geq 0 \\ -C_2\Delta, \text{ for } \Delta < 0, \end{cases} \qquad C_1, C_2 \text{ are constants,}$$

or

$$l(\Delta) = \begin{cases} \lambda w(\theta)L(\Delta), \text{ for } \Delta \geq 0, & \text{(over-estimation)} \\ w(\theta)L(\Delta), \text{ for } \Delta < 0, & \text{(under-estimation),} \end{cases}$$

where 'L' is typically a symmetric loss function, λ is an additional loss (in percentage) due to over-estimation, and $w(\theta)$ is a weight function.

The problem of finding the lower bound for the risk with asymmetric loss functions under the assumption of LAN was discussed by Lepskii [14] and Takagi [17]. In the present work we consider an asymmetric loss function and obtain the local asymptotic minimax risk bounds in a LAMN family of distributions.

The paper is organized as follows: Section 2 introduces the preliminaries and the relevant assumptions required to develop the main result. Section 3 is dedicated to the derivation of the main result. Section 4 contains the concluding remarks and directions for future research.

2. Preliminaries

Let $X_1, \ldots, X_n$ be n random variables defined on the probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$ and taking values in $(S, \mathcal{S})$, where S is the Borel subset of a Euclidean space and $\mathcal{S}$ is the σ-field of Borel subsets of S. Let the parameter space be Θ, where Θ is an open subset of R^1. It is assumed that the joint probability law of any finite set of such random variables has some known functional form except for the unknown parameter θ involved in the distribution. Let $\mathcal{A}_n$ be the σ-field generated by $X_1, \ldots, X_n$ and let $P_{\theta,n}$ be the restriction of P_θ to $\mathcal{A}_n$. Let θ_0 be the true value of θ and let $\theta_n = \theta_0 + \delta_n h$ $(h \in R^1)$, where $\delta_n \to 0$ as $n \to \infty$. The sequence δ_n may depend on θ but is independent of the observations. It is further assumed that, for each $n \geq 1$, the probability measures $P_{\theta_o,n}$ and $P_{\theta_n,n}$ are mutually absolutely continuous for all θ_0 and θ_n. Then the sequence of likelihood ratios is defined as

$$L_n(\mathbf{X}_n; \theta_0, \theta_n) = L_n(\theta_0, \theta_n) = \frac{dP_{\theta_n,n}}{dP_{\theta_0,n}},$$

where $\mathbf{X}_n = (X_1, \ldots, X_n)$ and the corresponding log-likelihood ratios are defined as

$$\Lambda_n(\theta_0, \theta_n) = \log L_n(\theta_0, \theta_n) = \log \frac{dP_{\theta_n,n}}{dP_{\theta_0,n}}.$$

Throughout the paper the following notation is used: $\phi_y(\mu, \sigma^2)$ represents the normal density with mean μ and variance σ^2; the symbol '$\Longrightarrow$' denotes convergence in distribution, and the symbol '$\rightarrow$' denotes convergence in $P_{\theta_0,n}$ probability.

Now let the sequence of statistical experiments $E_n = \{\mathcal{X}_n, \mathcal{A}_n, P_{\theta,n}\}_{n\geq 1}$ be a locally asymptotically mixture of normals (LAMN) at $\theta_0 \in \Theta$. For the definition of a LAMN experiment the reader is referred to Bhattacharya and Roussas [6]. Then there exist random variables Z_n and W_n ($W_n > 0$ a.s.) such that

$$\Lambda_n(\theta_0, \theta_n) = \log \frac{dP_{\theta_0+\delta_n h,n}}{dP_{\theta_0,n}} - hZ_n + \frac{1}{2}h^2 W_n \rightarrow 0, \tag{2.1}$$

and

$$(Z_n, W_n) \Rightarrow (Z, W) \text{ under } P_{\theta_0,n}, \tag{2.2}$$

where $Z = W^{1/2}G$, G and W are independently distributed, $W > 0$ a.s. and $G \sim N(0,1)$. Moreover, the distribution of W does not depend on the parameter h (Le Cam and Yang [13]).

The following examples illustrate the different quantities appearing in equations (2.1) and (2.2) and in the subsequent derivations.

Example 2.1 (An explosive autoregressive process of first order). Let the random variables $X_j, j = 1, 2, \ldots$ satisfy a first order autoregressive model defined by

$$X_j = \theta X_{j-1} + \epsilon_j, X_0 = 0, |\theta| > 1, \tag{2.3}$$

where ϵ_j's are *i.i.d.* $N(0,1)$ random variables. We consider the explosive case where $|\theta| > 1$. For this model we can write

$$f_j(\theta) = f(x_j | x_1, \ldots, x_{j-1}; \theta) \propto e^{-\frac{1}{2}(x_j - \theta x_{j-1})^2}.$$

Let θ_0 be the true value of θ. It can be shown that for the model described in (2.3) we can select the sequence of norming constants $\delta_n = \frac{(\theta_0^2-1)}{\theta_0^n}$ so that (2.1) and (2.2) hold. Clearly $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. We can also obtain $W_n(\theta_0), Z_n(\theta_0)$ and their asymptotic distributions, as $n \rightarrow \infty$, as follows:

$$W_n(\theta_0) = \frac{(\theta_0^2 - 1)^2}{\theta_0^{2n}} \sum_{j=1}^{n} X_{j-1}^2 \Rightarrow W \text{ as } n \rightarrow \infty, \text{ where } W \sim \chi_1^2 \text{ and}$$

$$G_n(\theta_0) = (\sum_{j=1}^{n} X_{j-1}^2)^{-\frac{1}{2}} (\sum_{j=1}^{n} X_{j-1}\epsilon_j) = (\sum_{j=1}^{n} X_{j-1}^2)^{\frac{1}{2}} (\hat{\theta}_n - \theta) \Rightarrow G,$$

where $G \sim N(0,1)$ and $\hat{\theta}_n$ is the m.l.e. of θ. Also

$$Z_n(\theta_0) = W_n^{\frac{1}{2}}(\theta_0) G_n(\theta_0) = \frac{(\theta_0^2 - 1)}{\theta_0^n} (\sum_{j=1}^{n} X_{j-1}\epsilon_j) \Rightarrow W^{\frac{1}{2}} G = Z,$$

where W is independent of G. It also holds that

$$(Z_n(\theta_0), W_n(\theta_0)) \Rightarrow (Z, W).$$

Hence $Z|W \sim N(0, W)$. In general Z is a mixture of normal distributions with W as the mixing variable.

Example 2.2 (A super-critical Galton–Watson branching process). Let $\{X_0=1, X_1, \ldots, X_n\}$ denote successive generation sizes in a super-critical Galton–Watson process with geometric offspring distribution given by

$$P(X_1 = j) = \theta^{-1}(1-\theta^{-1})^{j-1},\ j = 1, 2, \ldots, 1 < \theta < \infty. \tag{2.4}$$

Here $E(X_1) = \theta$ and $V(X_1) = \sigma^2(\theta) = \theta(\theta-1)$. For this model we can write

$$f_j(\theta) = f(x_j|x_1, \ldots, x_{j-1}; \theta) = (1-\frac{1}{\theta})^{x_j - x_{j-1}}(\frac{1}{\theta})^{x_{j-1}}.$$

Let θ_0 be the true value of θ. Here δ_n can be chosen as $\frac{\sqrt{\theta_0}(\theta_0-1)}{\theta_0^{n/2}}$. For this model the random variables $W_n(\theta_0), Z_n(\theta_0)$ and their asymptotic distributions are:

$$W_n(\theta_0) = \frac{(\theta_0-1)}{\theta_0^n}\sum_{j=1}^{n} X_{j-1} \Rightarrow W \text{ as } n \to \infty,$$

where W is an exponential random variable with unit mean. Here

$$G_n(\theta_0) = [\theta_0(\theta_0-1)]^{-\frac{1}{2}}(\sum_{j=1}^{n} X_{j-1})^{-\frac{1}{2}}\sum_{j=1}^{n}(X_j - \theta_0 X_{j-1}) \Rightarrow G,$$

where $G \sim N(0,1)$, and for W independent of G,

$$Z_n(\theta_0) = W_n^{\frac{1}{2}}(\theta_0)G_n(\theta_0) \Rightarrow W^{\frac{1}{2}}G.$$

It also holds that

$$(Z_n(\theta_0), W_n(\theta_0)) \Rightarrow (W^{\frac{1}{2}}G, W).$$

The decision problem considered here is the risk in the estimation of a parameter $\theta \varepsilon R^1$ using an asymmetric loss function $l(.)$. Throughout the rest of the manuscript the following assumptions apply:

A1 $l(z) \geq 0$ for all $z, z = \hat{\theta} - \theta$

A2 $l(z)$ is non increasing for $z < 0$, non-decreasing for $z > 0$ and $l(0) = 0$.

A3 $\int_{-\infty}^{\infty}\int_{0}^{\infty} l(w^{-\frac{1}{2}}z)e^{-\frac{1}{2}cwz^2}g(w)dwdz < \infty$ for any $c > 0$, where $g(w)$ is the p.d.f. of the random variable W.

A4 $\int_{-\infty}^{\infty}\int_{0}^{\infty} w^{\frac{1}{2}}z^2 l(w^{-\frac{1}{2}}d - z)e^{-\frac{1}{2}cwz^2}g(w)dwdz < \infty$, for any $c, d > 0$.
Define $l_a(y) = \min(l(y), a)$, for $0 < a \leq \infty$. This truncated loss makes $l(y)$ bounded if it is not so.

A5 For given $W = w > 0$, $h(\beta, w) = \int_{-\infty}^{\infty} l(w^{-\frac{1}{2}}\beta - y)\phi_y(0, w^{-1})dy$ attains its minimum at a unique $\beta = \beta_0(w)$, and $E\beta_0(W))$ is finite.

A6 For given $W = w > 0$, any large $a, b > 0$ and any small $\lambda > 0$ the function $\tilde{h}(\beta, w) = \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\beta - y)\phi_y(0, ((1+\lambda)w)^{-1})dy$
attains its minimum at $\tilde{\beta}(w) = \tilde{\beta}(a, b, \lambda, w)$, and $E\tilde{\beta}(a, b, \lambda, W) < \infty$.

A7 $\lim_{a\to\infty, b\to\infty, \lambda\to 0}\tilde{\beta}(a, b, \lambda, w) = \beta_0(w)$.

A8 $E(W^{-\frac{1}{2}}) < \infty$.

Note the following:

1. Assumptions A3 and A4 are general assumptions made to ensure the finiteness of the expected loss and other functions. Assumptions A5 and A6 are satisfied, for example, by convex loss functions.

2. If $l(.)$ is symmetric, then $\beta_0(w) = 0 = \tilde{\beta}(a, b, \lambda, w)$.
3. If $l(.)$ is unbounded, then the assumption A8 is replaced by A8′ as $E(W^{-1} \times Z^2 l(W^{-\frac{1}{2}} Z)) < \infty$.

Here we will consider a randomized estimator $\xi(Z, W)$ which can be written as $\xi(Z, W, U)$, where U is uniformly distributed on $[0, 1]$ and independent of Z and W. The introduction of randomized estimators is justified since the loss function $l(.)$ may not be convex.

3. Main result

Under the set of assumptions and notations stated in Section 2 we can have the following generalization of Hájek's result for LAMN experiments under an asymmetric loss structure.

Lemma 3.1. *Let $l(.)$ satisfy assumptions* A1 – A7 *and let $Z|W$ be a normal random variable with mean $\theta\sqrt{w}+\beta_0(w)$ and variance* 1. *Further let W be a random variable with p.d.f. $g(w)$, then for any $\epsilon > 0$ there is an $\alpha = \alpha(\epsilon) > 0$ and a prior density $\pi(\theta)$ so that for any estimator $\xi(Z, W, U)$ satisfying*

$$P_{\theta=0}(|\xi(Z, W, U) - W^{-\frac{1}{2}} Z| > \epsilon) > \epsilon \tag{3.1}$$

the Bayes risk $R(\pi, \xi)$ is

$$\begin{aligned} R(\pi, \xi) &= \int \pi(\theta) R(\theta, \xi) d\theta \\ &= \int \pi(\theta) E(l_a(\xi(Z, W, U) - \theta)|\theta) d\theta \\ &\geq \int l(w^{-\frac{1}{2}} \beta_0(w) - y) \phi_y(0, w^{-1}) g(w) dy dw + \alpha. \end{aligned} \tag{3.2}$$

Proof. Let the prior distribution of θ be given by $\pi(\theta) = \pi_\sigma(\theta) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} e^{-\frac{\theta^2}{2\sigma^2}}$, $\sigma > 0$, where the variance σ^2, which depends on ϵ as defined in (3.1), will be appropriately chosen later. As $\sigma^2 \longrightarrow \infty$, the prior distribution becomes diffuse. The joint distribution of Z, W and θ is given by

$$f(z|w) g(w) \pi(\theta) = (2\pi)^{-1} \sigma^{-1} e^{-\frac{1}{2}(z - (\theta w^{\frac{1}{2}} + \beta_0(w)))^2 - \frac{1}{2}\frac{\theta^2}{\sigma^2}} g(w). \tag{3.3}$$

The posterior distribution of θ given (W, Z) is given by $\psi(\theta|w, z)$, where $\psi(\theta|w, z)$ is $N(\frac{w^{\frac{1}{2}}(z - \beta_0(w))}{r(w,\sigma)}, \frac{1}{r(w,\sigma)})$ and the marginal joint distribution of (Z, W) is given by

$$f(z, w) = \phi_z(\beta_0(w), \sigma^2 r(w, \sigma)) g(w), \tag{3.4}$$

where the function $r(s, t) = s + 1/t^2$. Note that the Bayes' estimator of θ is $\frac{W^{\frac{1}{2}}(Z - \beta_0(W))}{r(W,\sigma)}$ and when the prior distribution is sufficiently diffused, the Bayes' estimator becomes $W^{-\frac{1}{2}}(Z - \beta_0(W))$.

Now let $\epsilon > 0$ be given and consider the following events:

$$\left|\frac{W^{\frac{1}{2}}(Z - \beta_0(W))}{r(W, \sigma)}\right| \leq b - \sqrt{b}, \;\; |\xi(Z, W, U) - W^{-\frac{1}{2}} Z| > \epsilon,$$

$$|W^{-\frac{1}{2}}(Z - \beta_0(W))| \leq M, \;\; \frac{1}{m} = \frac{1}{\sigma^2}\left(\frac{2M}{\epsilon} - 1\right) \leq W \leq m.$$

Then

$$
\begin{aligned}
|W^{-\frac{1}{2}}(Z-\beta_0(W)) - \frac{W^{\frac{1}{2}}(Z-\beta_0(W))}{r(W,\sigma)}| &= |\frac{\frac{W^{-\frac{1}{2}}(Z-\beta_0(W))}{\sigma^2}}{r(W,\sigma)}| \\
&\leq \frac{M}{\sigma^2 r(W,\sigma)} = \frac{M}{\sigma^2 W+1} \leq \frac{\epsilon}{2}.
\end{aligned} \tag{3.5}
$$

Now, for any large $a, b > 0$, we have

$$
\begin{aligned}
&\int_{-b}^{b} l_a(\xi(z,w,u)-\theta)\psi(\theta|z,w)d\theta \\
&\quad = \int_{-b}^{b} l_a(\xi(z,w,u) - y - \frac{w^{\frac{1}{2}}(z-\beta_0(w))}{r(w,\sigma)})\phi_y(0,\frac{1}{r(w,\sigma)})dy,
\end{aligned} \tag{3.6}
$$

where $y = \theta - \frac{w^{\frac{1}{2}}(z-\beta_0(w))}{r(w,\sigma)}$. Now, since $\theta|z,w \sim N(\frac{w^{\frac{1}{2}}(z-\beta_0(w))}{r(w,\sigma)}, \frac{1}{r(w,\sigma)})$, we have $y|z,w \sim N(0,\frac{1}{r(w,\sigma)})$. It can be seen that $|\xi(z,w,u) - \frac{w^{\frac{1}{2}}(z-\beta_0(w))}{r(w,\sigma)} - w^{-\frac{1}{2}}\beta_0(w)| > \frac{\epsilon}{2}$. Hence due to the nature of the loss function, for a given $w > 0$, we can have, from (3.6),

$$
\begin{aligned}
&\int_{-b}^{b} l_a(\xi(z,w,u) - \frac{w^{\frac{1}{2}}(z-\beta_0(w))}{r(w,\sigma)} - y)\phi_y(0,\frac{1}{r(w,\sigma)})dy \\
&\quad \geq \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\beta_0(w) - y)\phi_y(0,\frac{1}{r(w,\sigma)})dy \\
&\quad \geq \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\tilde{\beta}(a,b,\lambda,w) - y)\phi_y(0,\frac{1}{r(w,\sigma)})dy + \delta \\
&\quad = \tilde{h}(\tilde{\beta}(a,b,\lambda,w)) + \delta,
\end{aligned} \tag{3.7}
$$

where $\delta > 0$ depends only on ϵ but not on a, b, σ^2 and (3.7) holds for sufficiently large a, b, σ^2 (here $\lambda = \frac{1}{w\sigma^2} \to 0$ as $\sigma^2 \to \infty$ and $\frac{1}{m} \leq w \leq m$).

A simple calculation yields

$$
\begin{aligned}
&\tilde{h}(\tilde{\beta}(a,b,\lambda,w)) \\
&\quad = \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\tilde{\beta}(a,b,\lambda,w) - y)\phi_y(0,\frac{1}{r(w,\sigma)})dy \\
&\quad \geq \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\tilde{\beta}(a,b,\lambda,w) - y)\phi_y(0,\frac{1}{w})(1-\frac{y^2}{\sigma^2})dy \\
&\quad = h(\beta_0(w)) - \int_{-\sqrt{b}}^{\sqrt{b}} l_a(w^{-\frac{1}{2}}\tilde{\beta}(a,b,\lambda,w) - y)\frac{y^2}{\sigma^2}\phi_y(0,\frac{1}{w})dy.
\end{aligned} \tag{3.8}
$$

Hence

$$
\begin{aligned}
R(\pi(\theta),\xi) &= \int_{-\infty}^{\infty} \pi(\theta) R(\theta,\xi) d\theta \\
&\geq \int_{-b}^{b} \pi(\theta) E(l_a(\xi(Z,W,U)-\theta)) d\theta \\
&= \int_{\theta=-b}^{b} \int_{u=0}^{1} \int_{w=0}^{\infty} \int_{z=-\infty}^{\infty} l_a(\xi(z,w,u)-\theta)\psi(\theta|z,w) f(z,w) d\theta du dw dz \qquad (3.9) \\
&\geq \int h(\beta_0(w)) g(w) dw \times P(|W^{-\frac{1}{2}}(Z-\beta_0(W))| \geq b-\sqrt{b}) - \frac{k}{\sigma^2} \\
&\quad + \delta P\{|\xi(Z,W,U) - W^{-\frac{1}{2}} Z| > \epsilon, |W^{-\frac{1}{2}}(Z-\beta_0(W))| \leq M, \frac{1}{m} \leq W \leq m\},
\end{aligned}
$$

using (3.7), (3.8) and assumption A4, where $k > 0$ does not depend on a, b, σ^2. Let

$$
\begin{aligned}
A = \{(z,w,u) \in (-\infty,\infty) \times (0,\infty) \times (0,1) : &|\xi(z,w,u) - w^{-\frac{1}{2}} z| > \epsilon, \\
&|w^{-\frac{1}{2}}(z-\beta_0(w))| \leq M, \frac{1}{m} \leq w \leq m\}.
\end{aligned}
$$

Then $P(A|\theta=0) > \frac{\epsilon}{2}$ for sufficiently large M due to (3.1). Now under $\theta = 0$ the joint density of Z and W is $\phi_z(\beta_0(w),1) g(w)$. The overall joint density of Z and W is given in (3.4). The likelihood ratio of the two densities is given by

$$
\frac{f(z,w)}{f(z,w|\theta=0)} = \sigma^{-1} r(w,\sigma)^{-\frac{1}{2}} e^{\frac{1}{2}(z-\beta_0(w))^2 \frac{w}{r(w,\sigma)}}
$$

and the ratio is bounded below on $\{(z,w) : |w^{-\frac{1}{2}}(z-\beta_0(w))| \leq M, \frac{1}{m} \leq w \leq m\}$ by $\sigma^{-1} r(m,\sigma)^{-\frac{1}{2}} = \frac{1}{(m\sigma^2+1)^{\frac{1}{2}}}$. Finally we have

$$
P(A) = \int_{u=0}^{1} \int_{A} f(z,w,u) dz dw du \geq \frac{1}{(m\sigma^2+1)^{\frac{1}{2}}} \frac{\epsilon}{2}. \qquad (3.10)
$$

Hence for sufficiently large m and M, from (3.9), we have

$$
R(\pi(\theta),\xi) \geq \int h(\beta_0(w)) g(w) dw [1 - \frac{\alpha}{2h(\beta_0(w))}] - \frac{k}{\sigma^2} + \delta \frac{\epsilon}{2} \frac{1}{(m\sigma^2+1)^{\frac{1}{2}}},
$$

assuming $P[|W^{-\frac{1}{2}}(Z-\beta_0(W))| \leq b-\sqrt{b}] \geq 1 - \frac{\alpha}{2h(\beta_0(w))}$. That is,

$$
R(\pi(\theta),\xi) \geq \int h(\beta_0(w)) g(w) dw - \frac{\alpha}{2} - \frac{k}{\sigma^2} + \delta \frac{\epsilon}{2} \frac{1}{(m\sigma^2+1)^{\frac{1}{2}}}.
$$

Putting $\delta \frac{\epsilon}{2} (m\sigma^2+1)^{-1/2} - \frac{k}{\sigma^2} = \frac{3\alpha}{2}$ we find $R(\pi(\theta),\xi) \geq \int h(\beta_0(w)) g(w) dw + \alpha$. Hence the proof of the result is complete. □

Theorem 3.1. *Suppose that the sequence of experiments $\{E_n\}$ satisfies LAMN conditions at $\theta \in \Theta$ and the loss function $l(.)$ meets the assumptions* A1–A8 *stated in Section* 2. *Then for any sequence of estimators $\{T_n\}$ of θ based on $X_1, \ldots, X_n$ the lower bound of the risk of $\{T_n\}$ is given by*

$$
\lim_{\delta \to 0} \liminf_{n \to \infty} sup_{|\theta - t| < \delta} E_\theta \{l(\delta_n^{-1}(T_n - \theta))\} \geq \int l(\beta_0(w) - y) \phi_y(0, w^{-1}) g(w) dy dw
$$

Furthermore, if the lower bound is attained, then

$$\delta_n^{-1}(T_n - \theta)) - \frac{W_n^{\frac{1}{2}}(Z_n - \beta_0(W))}{r(W_n, \sigma)} \to 0$$

or, as $\sigma^2 \to \infty$

$$\delta_n^{-1}(T_n - \theta) - W_n^{-\frac{1}{2}}(Z_n - \beta_0(W)) \to 0.$$

Proof. Since the upper bound of values of a function over a set is at least its mean value on that set, we may write, for sufficiently large n,

$$\sup_{|\theta - t| < \delta} E_\theta\{l(\delta_n^{-1}(T_n - \theta))\} \geq \int_{-b}^{b} \pi(h) E\{l_a(\delta_n^{-1}(T_n - t) - h) | \theta = t + \delta_n h\} dh,$$

whatever the values of constants a, b and the prior density $\pi(h)$ may be. Now, let $Z|w, \theta = t + \delta_n h \sim N(\theta w^{\frac{1}{2}} + \beta_0(w), 1)$. Then we can fix some $\delta > 0$ and choose a, b and $\pi(.)$ in such a way that

$$\int_{-b}^{b} \pi(h) E\{l_a(\xi(Z, W, U) - h) | t + \delta_n h\} dh \geq \int l(\beta_0(w) - y) \phi_y(0, w^{-1}) g(w) dy dw - \delta,$$

for any estimator $\xi(Z, W, U)$.

Next we use Lemmas 3.3 and 3.4 of Takagi [17], where we set

$$S_n = \delta_n^{-1}(T_n - t), \ \ \Delta_{n,t} = W_n^{-\frac{1}{2}}(Z_n - \beta_0(W)), \text{ and}$$
$$S_n(\Delta_{n,t} = x, U = u) = \inf\{y : P(S_n \leq y | \Delta_{n,t} = x) \geq u\}.$$

Let $F_{n,h}$ = distribution of S_n under $P_{n,h}$, $F^*_{n,h}$ = distribution of $S_n(\Delta_{n,t}, U) = \xi_n(Z_n, W, U)$ under $P_{n,h}$, where $U \sim$ Uniform $(0, 1)$ and is independent of $\Delta_{n,t}$; $G_{n,h}$ is the distribution of $\Delta_{n,t}$ and $G^*_{n,h}$ is the distribution of $\Delta_t = W^{\frac{1}{2}}(Z - \beta_0(W))$. As a consequence of this we have (Takagi [17], p.44)

$$\lim_{n \to \infty} ||F_{n,h} - F^*_{n,h}|| = 0 \text{ and } \lim_{n \to \infty} ||G_{n,h} - G^*_{n,h}|| = 0.$$

Now for any estimator $\xi_n(Z_n, W_n, U) = S_n(\Delta_{n,t}, U)$ and for every $h \varepsilon R^1$ we have

$$|E[l_a(\delta_n^{-1}(T_n - t) - h) | t + \delta_n h] - E[l_a(S_n(\Delta_{n,t}, U) - h) | t + \delta_n h]| \longrightarrow 0$$

and

$$|E[l_a(S_n(\Delta_{n,t}, U) - h) | t + \delta_n h] - E[l_a(\xi_n(Z, W, U) - h) | t + \delta_n h]| \longrightarrow 0.$$

Finally

$$\int_{-b}^{b} \pi(h) E\{l(\delta_n^{-1}(T_n - t) - h) | \theta = t + \delta_n h\} dh$$
$$\geq \int_{-b}^{b} \pi(h) E\{l_a(\xi_n(Z, W, U) - h) | t + \delta_n h\} dh$$
$$\geq \int l(\beta_0(w) - y) \phi_y(0, w^{-1}) g(w) dy dw, \text{ for } n \geq n(a, b, \delta_n, \pi)$$

which proves the result. □

Example 3.1. Consider the LINEX loss function as defined by (1.1). It can be seen that $l(\triangle)$ satisfies all the assumptions A1–A7 stated in Section 2. Here a simple calculation will yield

$$h(\beta, w) = b(e^{aw^{-\frac{1}{2}}(\beta + \frac{1}{2}\frac{a}{w^{1/2}})} - aw^{-\frac{1}{2}}\beta - 1),$$

and $h(\beta, w)$ attains its minimum at $\beta_0(w) = -\frac{1}{2}\frac{a}{w^{1/2}}$ and $h(\beta_0, w) = -\frac{b}{2}\frac{a^2}{w}$.

4. Concluding remarks

From the results discussed in Le Cam and Yang [13] and Jeganathan [10] it is clear that under symmetric loss structure the results derived in Theorem 3.1 hold with respect to the estimator $W_n^{-\frac{1}{2}}(\theta_0)Z_n(\theta_0)$ and its asymptotic counterpart $W^{-\frac{1}{2}}Z$. Here due to the presence of asymmetry in the loss structure the results derived in Theorem 3.1 hold with respect to the estimator $W_n^{-\frac{1}{2}}(\theta_0)(Z_n(\theta_0) - \beta_0(W)) + \beta_0(W)$ and $W^{-\frac{1}{2}}(Z - \beta_0(W)) + \beta_0(W)$.

Now $W_n^{-\frac{1}{2}}(\theta_0)(Z_n(\theta_0) - \beta_0(W)) \Rightarrow W^{-\frac{1}{2}}(Z - \beta_0(W))$. Hence the asymptotic bias of the estimator under asymmetric loss would be $E(W^{-\frac{1}{2}}(\theta_0)(Z - \beta_0(W)) + \beta_0(W) - \theta) = E(\theta + \beta_0(W) - \theta) = E(\beta_0(W))$.

Consider the model described in Example 2.1. Under the LINEX loss we have $\beta_0(w) = -\frac{a}{2}\frac{1}{w^{1/2}}$ (vide Example 3.1). Here the asymptotic bias of the estimator would be $E(\beta_0(W)) = -\frac{a}{2}E(W^{-\frac{1}{2}})$, which is finite due to Assumption A8.

The results obtained in this paper can be extended in the following two directions: (1) To investigate the case when the experiment is Locally Asymptotically Quadratic (LAQ), and (2) To find the asymptotic minimax lower bound for a sequential estimation scheme under the conditions of LAN, LAMN and LAQ considering asymmetric loss function.

Acknowledgments. The authors are indebted to the referees, whose comments and suggestions led to a significant improvement of the paper. The first author is also grateful to the Editor for his support in publishing the article.

References

[1] Bahadur, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhya* **22**, 229–252.

[2] Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38**, 303–324.

[3] Basu, A. K. and Bhattacharya, D. (1999). Asymptotic minimax bounds for sequential estimators of parameters in a locally asymptotically quadratic family, *Braz. J. Probab. Statist.* **13**, 137–148.

[4] Basu, D. (1956). The concept of asymptotic efficiency. *Sankhyā* **17**, 193–196.

[5] Basawa, I. V. and Scott, D. J. (1983). *Asymptotic Optimal Inference for Nonergodic Models*. Lecture Notes in Statistics. Springer-Verlag.

[6] Bhattacharya, D. and Roussas, G. G. (2001). Exponential approximation for randomly stopped locally asymptotically mixture of normal experiments. *Stochastic Modeling and Applications* **4**, 2, 56–71.

[7] Bhattacharya, D., Samaniego, F. J. and Vestrup, E. M. (2002). On the comparative performance of Bayesian and classical point estimators under asymmetric loss. *Sankhyā* Ser. B **64**, 230–266.

[8] Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* Univ. California Press, Berkeley, 175–194.

[9] Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York.

[10] Jeganathan, P. (1983). Some asymptotic properties of risk functions when the limit of the experiment is mixed normal. *Sankhyā* Ser. A **45**, 66–87.

[11] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood and Bayes' estimates. *Univ. California Publ. Statist.* **1**, 277–330.

[12] Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3**, 37–98.

[13] Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics, Some Basic Concepts.* Lecture Notes in Statistics. Springer, Verlag.

[14] Lepskii, O. V. (1987). Asymptotic minimax parameter estimator for non symmetric loss function. *Theo. Probab. Appl.* **32**, 160–164.

[15] Levine, R. A. and Bhattacharya, D. (2000). Bayesion estimation and prior selection for AR(1) model using asymmetric loss function. Technical report 353, Department of Statistics, University of California, Davis.

[16] Rojo, J. (1987). On the admissibility of $c\overline{X} + d$ with respect to the LINEX loss function. *Commun. Statist. Theory Meth.* **16**, (12), 3745–3748.

[17] Takagi, Y. (1994). Local asymptotic minimax risk bounds for asymmetric loss functions. *Ann. Statist.* **22**, 39–48.

[18] Varian, H. R. (1975). A Bayesian approach to real estate assessment; In *Studies in Bayesian Econometrics and Statistics, in Honor of Leonard J. Savage* (eds. S. E. Feinberg and A. Zellner). North Holland, 195–208.

[19] Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* **10**, 299–326.

[20] Wald, A. (1947). An essentially complete class of admissible decision functions. *Ann. Math. Statist.* **18**, 549–555.

[21] Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* **81**, 446–451.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 322–333
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000536

On moment-density estimation in some biased models

Robert M. Mnatsakanov[1] and Frits H. Ruymgaart[2]

West Virginia University and Texas Tech University

Abstract: This paper concerns estimating a probability density function f based on iid observations from $g(x) = W^{-1}\, w(x)\, f(x)$, where the weight function w and the total weight $W = \int w(x)\, f(x)\, dx$ may not be known. The length-biased and excess life distribution models are considered. The asymptotic normality and the rate of convergence in mean squared error (MSE) of the estimators are studied.

1. Introduction and preliminaries

It is known from the famous "moment problem" that under suitable conditions a probability distribution can be recovered from its moments. In Mnatsakanov and Ruymgaart [5, 6] an attempt has been made to exploit this idea and estimate a cdf or pdf, concentrated on the positive half-line, from its empirical moments.

The ensuing density estimators turned out to be of kernel type with a convolution kernel, provided that convolution is considered on the positive half-line with multiplication as a group operation (rather than addition on the entire real line). This does not seem to be unnatural when densities on the positive half-line are to be estimated; the present estimators have been shown to behave better in the right hand tail (at the level of constants) than the traditional estimator (Mnatsakanov and Ruymgaart [6]).

Apart from being an alternative to the usual density estimation techniques, the approach is particularly interesting in certain inverse problems, where the moments of the density of interest are related to those of the actually sampled density in a simple explicit manner. This occurs, for instance, in biased sampling models. In such models the pdf f (or cdf F) of a positive random variable X is of actual interest, but one observes a random sample $Y_1, \ldots, Y_n$ of n copies of a random variable Y with density

$$g(y) = \frac{1}{W}\, w(y)\, f(y)\,,\ y \geq 0, \tag{1.1}$$

where the weight function w and the total weight

$$W = \int_0^\infty w(x)\, f(x)\, dx\,, \tag{1.2}$$

[1]Department of Statistics, West Virginia University, Morgantown, WV 26506, USA, e-mail: rmnatsak@stat.wvu.edu

[2]Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX 79409, USA, e-mail: h.ruymgaart@ttu.edu

AMS 2000 subject classifications: primary 62G05; secondary 62G20.

Keywords and phrases: moment-density estimator, weighted distribution, excess life distribution, renewal process, mean squared error, asymptotic normality.

may not be known. In this model one clearly has the relation

$$\mu_{k,F} = \int_0^\infty x^k f(x)\,dx = W \int_0^\infty y^k \frac{1}{w(y)}\, g(y)\,dy,\ k = 0, 1, \ldots, \tag{1.3}$$

and unbiased $\sqrt{n}$-consistent estimators of the moments of F are given by

$$\widehat{\mu}_k = \frac{W}{n} \sum_{i=1}^{n} Y_i^k \frac{1}{w(Y_i)}. \tag{1.4}$$

If w and W are unknown they have to be replaced by estimators to yield $\widehat{\widehat{\mu}}_k$, say. In Mnatsakanov and Ruymgaart [7] moment-type estimators for the cdf F of X were constructed in biased models. In this paper we want to focus on estimating the density f and related quantities. Following the construction pattern in Mnatsakanov and Ruymgaart [6], substitution of the empirical moments $\widehat{\mu}_k$ in the inversion formula for the density yields the estimators

$$\hat{f}_\alpha(x) = \frac{W}{n} \sum_{i=1}^{n} \frac{1}{w(Y_i)} \cdot \frac{\alpha - 1}{x \cdot (\alpha - 1)!} \cdot \left(\frac{\alpha}{x} Y_i\right)^{\alpha - 1} \exp(-\frac{\alpha}{x} Y_i)\,,\ x \geq 0, \tag{1.5}$$

after some algebraic manipulation, where α is positive integer with $\alpha = \alpha(n) \to \infty$, as $n \to \infty$, at a rate to be specified later. If W or w are to be estimated, the empirical moments $\widehat{\widehat{\mu}}_k$ are substituted and we arrive at $\hat{\hat{f}}_\alpha$, say.

A special instance of model (1.1) to which this paper is devoted for the most part is length-biased sampling, where

$$w(y) = y,\ y \geq 0. \tag{1.6}$$

Bias and MSE for the estimator (1.5) in this particular case are considered in Section 3 and its asymptotic normality in Section 4. Although the weight function w is known, its mean W still remains to be estimated in most cases, and an estimator of W is also briefly discussed. The literature on length-biased sampling is rather extensive; see, for instance Vardi [9], Bhattacharyya *et al.* [1] and Jones [4].

Another special case of (1.1) occurs in the study of the distribution of the excess of a renewal process; see, for instance, Ross [8] for a brief introduction. In this situation, it turns out that the sampled density satisfies (1.1) with

$$w(y) = \frac{1 - F(y)}{f(y)} = \frac{1}{h_F(y)}\,,\ y \geq 0, \tag{1.7}$$

where h_F is the hazard rate of F. Although apparently w and hence W are not known here, they depend exclusively on f. In Section 5 we will briefly discuss some estimators for f, h_F and W and in particular show that they are all related to estimators of g and its derivative. Estimating this g is a "direct" problem and can formally be considered as a special case of (1.1) with $w(y) = 1$, $y \geq 0$ and $W = 1$. Investigating rates of convergence of the corresponding estimators is beyond the scope of this paper. Finally, in Section 6 we will compare the mean squared errors of the moment-density estimator $\widehat{f}_\alpha^*$ introduced in the Section 2 and the kernel-density estimator f_h studied by Jones [4] for the length-biased model. Throughout the paper let us denote by $G(a, b)$ a gamma distribution with shape and scale parameters a and b, respectively. We carried out simulations for length-biased model (1.1) with g as the gamma $G(2, 1/2)$ density and constructed corresponding graphs for $\widehat{f}_\alpha^*$ and f_h. Also we compare the performance of the moment-type and kernel-type estimators for the model with excess life-time distribution when the target distribution F is gamma $G(2, 2)$.

2. Construction of moment-density estimators and assumptions

Let us consider the general weighted model (1.1) and assume that the weight function w is known. The estimated total weight $\hat{W}$ can be defined as follows:

$$\hat{W} = \Big(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{w(Y_j)}\Big)^{-1}.$$

Substitution of the empirical moments

$$\widehat{\widehat{\mu}}_k = \frac{\hat{W}}{n}\sum_{i=1}^{n} Y_i^k\,\frac{1}{w(Y_i)}$$

in the inversion formula for the density (see, Mnatsakanov and Ruymgaart [6]) yields the construction

$$\hat{\hat{f}}_\alpha(x) = \frac{\hat{W}}{n}\sum_{i=1}^{n}\frac{1}{w(Y_i)}\cdot\frac{\alpha-1}{x\cdot(\alpha-1)!}\cdot\Big(\frac{\alpha}{x}Y_i\Big)^{\alpha-1}\exp(-\frac{\alpha}{x}Y_i)\,. \tag{2.1}$$

Here α is positive integer and will be specified later. Note that the estimator $\hat{\hat{f}}_\alpha$ is the probability density itself. Note also that

$$\hat{W} = W + O_p(\frac{1}{\sqrt{n}}),\ n\to\infty\,,$$

(see, Cox [2] or Vardi [9]). Hence one can replace $\hat{W}$ in (2.1) by W.

Investigating the length-biased model, modify the estimator $\hat{\hat{f}}_\alpha$ and consider

$$\begin{aligned}\widehat{f}^{*}_\alpha(x) &= \frac{W}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\cdot\frac{\alpha}{x\cdot(\alpha-1)!}\cdot\Big(\frac{\alpha}{x}Y_i\Big)^{\alpha-1}\exp(-\frac{\alpha}{x}Y_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{W}{Y_i^2}\cdot\frac{1}{\Gamma(\alpha)}\cdot\Big(\frac{\alpha Y_i}{x}\Big)^{\alpha}\cdot\exp(-\frac{\alpha}{x}Y_i) = \frac{1}{n}\sum_{i=1}^{n}M_i.\end{aligned}$$

In Sections 3 and 4 we will assume that the density f satisfies

$$f\in C^{(2)}([0,\infty)),\quad \text{with}\quad \sup_{t\ge0}\big|f''(t)\big| = M<\infty. \tag{2.2}$$

In Section 5 we will consider the estimator of the unknown survival function S given the sample $Y_1,\dots,Y_n$ from pdf (1.1) where $w=(1-F)/f$. Namely, we will use the moment-density estimator proposed in Mnatsakanov and Ruymgaart [6] which yields the estimator of $S=1-F$:

$$\hat{S}_\alpha(x) = \frac{W}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\cdot\frac{1}{(\alpha-1)!}\cdot\Big(\frac{\alpha}{x}Y_i\Big)^{\alpha}\exp(-\frac{\alpha}{x}Y_i)\,.$$

We will assume that F satisfies the conditions

$$F\in C^{(2)}([0,\infty)),\quad \text{with}\quad \sup_{t\ge0}\big|F''(t)\big| = L<\infty. \tag{2.3}$$

Throughout the paper the moment-type estimators will be considered at a fixed point $x>0$, where $f(x)>0$.

3. The bias and MSE of $\hat{f}^*_\alpha$

To study the asymptotic properties of $\widehat{f}^*_\alpha$ let us introduce for each $k \in \mathbb{N}$ the sequence of gamma $G(k(\alpha-2)+2, x/k\alpha)$ density functions

$$h_{\alpha,x,k}(u) = \frac{1}{\{k(\alpha-2)+1\}!}\left(\frac{k\alpha}{x}\right)^{k(\alpha-2)+2} u^{k(\alpha-2)+1} \times \exp(-\frac{k\alpha}{x}u), \quad u \geq 0, \tag{3.1}$$

with mean $\{k(\alpha-2)+2\}x/(k\alpha)$ and variance $\{k(\alpha-2)+2\}x^2/(k\alpha)^2$. For each $k \in \mathbb{N}$, moreover, these densities form as well a delta sequence. Namely,

$$\int_0^\infty h_{\alpha,x,k}(u)\, f(u)\, du \to f(x) \ \text{, as } \alpha \to \infty,$$

uniformly on any bounded interval (see, for example, Feller [3], vol. II, Chapter VII). This property of $h_{\alpha,x,k}$, when $k=2$ is used in (3.10) below. In addition, for $k=1$ we have

$$\int_0^\infty u\, h_{\alpha,x,1}(u)\, du = x, \tag{3.2}$$

$$\int_0^\infty (u-x)^2\, h_{\alpha,x,1}(u)\, du = \frac{x^2}{\alpha}. \tag{3.3}$$

Theorem 3.1. *Under the assumptions (2.2) the bias of $\widehat{f}^*_\alpha$ satisfies*

$$E\widehat{f}^*_\alpha(x) - f(x) = \frac{x^2 f''(x)}{2\cdot\alpha} + \mathrm{o}\left(\frac{1}{\alpha}\right), \quad \text{as } \alpha \to \infty. \tag{3.4}$$

For the Mean Squared Error (MSE) *we have*

$$MSE\{\widehat{f}^*_\alpha(x)\} = n^{-4/5}\left[\frac{W\cdot f(x)}{2\sqrt{\pi}x^2} + \frac{x^4\{f''(x)\}^2}{4}\right] + o(1), \tag{3.5}$$

provided that we choose $\alpha = \alpha(n) \sim n^{2/5}$.

Proof. Let $M_i = W \cdot Y_i^{-1} \cdot h_{\alpha,x,1}(Y_i)$. Then

$$\begin{aligned} E\, M_i^k &= \int_0^\infty W^k \cdot Y_i^{-k}\, h^k_{\alpha,x,1}(y)\, g(y)\, dy \\ &= \int_0^\infty \frac{W^k}{\{y\cdot(\alpha-1)!\}^k}\left(\frac{\alpha}{x}\right)^{k\alpha} y^{k(\alpha-1)} \exp\left(-\frac{k\alpha}{x}y\right)\frac{y\cdot f(y)}{W}dy \\ &= W^{k-1}\int_0^\infty \frac{1}{\{(\alpha-1)!\}^k}\left(\frac{\alpha}{x}\right)^{k\alpha} y^{k(\alpha-2)+1}\exp\left(-\frac{k\alpha}{x}y\right) f(y)\, dy \\ &= W^{k-1}\left(\frac{\alpha}{x}\right)^{2(k-1)} \frac{\{k(\alpha-2)+1\}!}{\{(\alpha-1)!\}^k}\,\frac{1}{k^{k(\alpha-2)+2}}\int_0^\infty h_{\alpha,x,k}(y) f(y) dy. \end{aligned} \tag{3.6}$$

In particular, for $k=1$:

$$\begin{aligned} E\widehat{f}^*_\alpha(x) &= f_\alpha(x) = W\int_0^\infty \frac{1}{y^2}\cdot\frac{1}{\Gamma(\alpha)}\cdot\left(\frac{\alpha}{x}y\right)^\alpha \exp(-\frac{\alpha}{x}y)\,\frac{y f(y)}{W}\,dy \\ &= \int_0^\infty h_{\alpha,x,1}(y) f(y) dy = E\, M_i. \end{aligned} \tag{3.7}$$

This yields for the bias ($\mu = x, \sigma^2 = x^2/\alpha$)

$$
\begin{aligned}
f_\alpha(x) - f(x) &= \int_0^\infty h_{\alpha,x,1}(y)\{f(y) - f(x)\}du \\
&= \int_0^\infty h_{\alpha,x,1}(y)\{f(x) + (y-x)f'(x) \\
&\quad + \frac{1}{2}\int_0^\infty (y-x)^2\{f''(\tilde{y}) - f(x)\}dy \\
&= \frac{1}{2}\int_0^\infty (y-x)^2 h_{\alpha,x,1}(y) f''(x)du \\
&\quad + \frac{1}{2}\int_0^\infty (y-x)^2 h_{\alpha,x,1}(y)\{f''(\tilde{y}) - f''(x)\}dy \\
&= \frac{1}{2}\frac{x^2}{\alpha}f''(x) + o\left(\frac{1}{\alpha}\right), \text{ as } \alpha \to \infty.
\end{aligned} \tag{3.8}
$$

For the variance we have

$$
\operatorname{Var}\widehat{f}^*_\alpha(x) = \frac{1}{n}\operatorname{Var} M_i = \frac{1}{n}\{E\, M_i^2 - f_\alpha^2(x)\}. \tag{3.9}
$$

Applying (3.6) for $k = 2$ yields

$$
\begin{aligned}
E\, M_i^2 &= W\,\frac{\alpha^2}{x^2}\frac{(2\alpha-3)!}{\{(\alpha-1)!\}^2}\frac{1}{2^{2\alpha-2}}\int_0^\infty h_{\alpha,x,2}(u)f(u)du \\
&\sim \frac{\alpha^2}{x^2}\frac{W}{\sqrt{2\pi}}\frac{e^{-(2\alpha-3)}\{(2\alpha-3)\}^{(2\alpha-3)+1/2}}{e^{-2(\alpha-1)}\{(\alpha-1)\}^{2(\alpha-1)+1}}\frac{1}{2^{2(\alpha-1)}} \\
&\quad \times \int_0^\infty h_{\alpha,x,2}(u)f(u)du = \frac{W}{2\sqrt{\pi}}\frac{\sqrt{\alpha}}{x^2}\int_0^\infty h_{\alpha,x,2}(u)f(u)du \\
&= \frac{W}{2\sqrt{\pi}}\frac{\sqrt{\alpha}}{x^2}\{f(x) + o(1)\} = \frac{W}{2\sqrt{\pi}}\frac{\sqrt{\alpha}}{x^2}f(x) + o(\sqrt{\alpha})
\end{aligned} \tag{3.10}
$$

as $\alpha \to \infty$. Now inserting this in (3.9) we obtain

$$
\begin{aligned}
\operatorname{Var}\widehat{f}^*_\alpha(x) &= \frac{1}{n}\Big[\frac{W}{2\sqrt{\pi}}\frac{\sqrt{\alpha}}{x^2}f(x) + o(\sqrt{\alpha}) - \left\{f(x) + O\left(\frac{1}{\alpha}\right)\right\}^2\Big] \\
&= \frac{W\sqrt{\alpha}}{2n\sqrt{\pi}}\frac{f(x)}{x^2} + o\left(\frac{\sqrt{\alpha}}{n}\right).
\end{aligned} \tag{3.11}
$$

Finally, this leads to the MSE of $\widehat{f}^*_\alpha(x)$:

$$
MSE\{\widehat{f}^*_\alpha(x)\} = \frac{W\sqrt{\alpha}}{2n\sqrt{\pi}}\frac{f(x)}{x^2} + \frac{1}{4}\frac{x^4}{\alpha^2}\{f''(x)\}^2 + o\left(\frac{\sqrt{\alpha}}{n}\right) + o\left(\frac{1}{\alpha^2}\right). \tag{3.12}
$$

For optimal rate we may take

$$
\alpha = \alpha_n \sim n^{2/5}, \tag{3.13}
$$

assuming that n is such that α_n is an integer. By substitution (3.13) in (3.12) we find (3.5). □

Corollary 3.1. *Assume that the parameter $\alpha = \alpha(x)$ is chosen locally for each $x > 0$ as follows*

$$\alpha(x) = n^{2/5} \cdot \{\frac{\pi}{4 \cdot W^2}\}^{1/5} \left[\frac{x^3 \cdot f''(x)}{\sqrt{f(x)}}\right]^{4/5}, \; f''(x) \neq 0. \tag{3.14}$$

Then the estimator $\widetilde{f}^*_\alpha(x) = \widehat{f}^*_\alpha(x)$ satisfies

$$\widetilde{f}^*_\alpha(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{W}{Y_i^2} \cdot \frac{1}{\Gamma(\alpha(x))} \cdot \left(\frac{\alpha(x)}{x} Y_i\right)^{\alpha(x)} \exp\{-\frac{\alpha(x)}{x} Y_i\} \tag{3.15}$$

$$MSE\,\{\widetilde{f}^*_\alpha(x)\} = n^{-4/5} \left[\frac{W^2 \cdot f''(x) \cdot f^2(x)}{\pi \cdot x^2 \sqrt{2}}\right]^{2/5} + o(1), \text{ as } n \to \infty. \tag{3.16}$$

Proof. Assuming the first two terms in the right hand side of (3.12) are equal to each other one obtains that for each n the function $\alpha = \alpha(x)$ can be chosen according to (3.14). This yields the proof of Corollary 1. □

4. The asymptotic normality of $\widehat{f}^*_\alpha$

Now let us derive the limiting distributions of $\widehat{f}^*_\alpha$. The following statement is valid.

Theorem 4.1. *Under the assumptions* (2.2) *and $\alpha = \alpha(n) \sim n^\delta$, for any $0 < \delta < 2$, we have, as $\alpha \to \infty$,*

$$\frac{\widehat{f}^*_\alpha(x) - f_\alpha(x)}{\sqrt{Var \widehat{f}^*_\alpha(x)}} \to_d \; Normal\,(0,1)\,. \tag{4.1}$$

Proof. Let $0 < C < \infty$ denote a generic constant that does not depend on n but whose value may vary from line to line. Note that for arbitrary $k \in \mathbb{N}$ the "c_r-inequality" entails that $\mathrm{E}\big|M_i - f_\alpha(x)\big|^k \leq C\,\mathrm{E}\,M_i^k$, in view of (3.6) and (3.7). Now let us choose the integer $k > 2$. Then it follows from (3.6) and (3.11) that

$$\begin{aligned} \frac{\sum_{i=1}^n \mathrm{E}\big|\frac{1}{n}\{M_i - f_\alpha(x)\}\big|^k}{\{\mathrm{Var}\hat{f}_\alpha(x)\}^{k/2}} &\leq C \frac{n^{1-k} k^{-1/2} \alpha^{k/2-1/2}}{(n^{-1}\alpha^{1/2})^{k/2}} \\ &= C \frac{1}{\sqrt{k}} \frac{\alpha^{k/4-1/2}}{n^{k/2-1}} \to 0, \text{ as } n \to \infty, \end{aligned} \tag{4.2}$$

for $\alpha \sim n^\delta$. Thus the Lyapunov's condition for the central limit theorem is fulfilled and (4.1) follows for any $0 < \delta < 2$. □

Theorem 4.2. *Under the assumptions* (2.2) *we have*

$$\frac{n^{1/2}}{\alpha^{1/4}} \{\widehat{f}^*_\alpha(x) - f(x)\} \to_d \; Normal\left(0, \frac{W \cdot f(x)}{2\,x^2\,\sqrt{\pi}}\right), \tag{4.3}$$

as $n \to \infty$, provided that we take $\alpha = \alpha(n) \sim n^\delta$ for any $\frac{2}{5} < \delta < 2$.

Proof. This is immediate from (3.11) and (4.1), since combined with (3.8) entails that $n^{1/2}\,\alpha^{-1/4}\,\{f_\alpha(x) - f(x)\} = O(n^{1/2}\,\alpha^{-5/4}) = O(n^{-\frac{5\delta-2}{4}}) = o(1)$, as $n \to \infty$, for the present choice of α. □

Corollary 4.1. *Let us assume that* (2.2) *is valid. Consider* $\widetilde{f}^*_\alpha(x)$ *defined in* (3.15) *with* $\alpha(x)$ *given by* (3.14). *Then*

$$\frac{n^{1/2}}{\alpha(x)^{1/4}}\,\{\widetilde{f}^*_\alpha(x)-f(x)\}\to_d \mathit{Normal}\left([\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}]^{1/2},\,\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}\right), \tag{4.4}$$

as $n\to\infty$ *and* $f''(x)\neq 0$.

Proof. From (4.1) and (3.11) with $\alpha=\alpha(x)$ defined in (3.13) it is easy to see that

$$\frac{n^{1/2}}{\alpha(x)^{1/4}}\,\{\widetilde{f}^*_\alpha(x)-E\widetilde{f}^*_\alpha(x)\}=\text{Normal}\left(0,\,\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}\right)+o_P(\frac{1}{n^{2/5}}), \tag{4.5}$$

as $n\to\infty$. Application of (3.4) where $\alpha=\alpha(x)$ is defined by (3.14) yields (4.4). □

Corollary 4.2. *Let us assume that* (2.2) *is valid. Consider* $\widetilde{f}^*_{\alpha^*}(x)$ *defined in* (3.15) *with* $\alpha^*(x)$ *given by*

$$\alpha^*(x)=n^\delta\cdot\{\frac{\pi}{4\cdot W^2}\}^{1/5}\left[\frac{x^3\cdot f''(x)}{\sqrt{f(x)}}\right]^{4/5},\ \frac{2}{5}<\delta<2. \tag{4.6}$$

Then when $f''(x)\neq 0$, *and letting* $n\to\infty$, *it follows that*

$$\frac{n^{1/2}}{\alpha^*(x)^{1/4}}\,\{\widetilde{f}^*_{\alpha^*}(x)-f(x)\}\to_d \mathit{Normal}\left(0,\,\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}\right). \tag{4.7}$$

Proof. Again from (4.1) and (3.11) with $\alpha=\alpha^*(x)$ defined in (4.6) it is easy to see that

$$\frac{n^{1/2}}{\alpha^*(x)^{1/4}}\,\{\widetilde{f}^*_{\alpha^*}(x)-E\widetilde{f}^*_{\alpha^*}(x)\}=\text{Normal}\left(0,\,\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}\right)+o_P(1), \tag{4.8}$$

as $n\to\infty$. On the other hand application of (3.4) where $\alpha=\alpha^*(x)$ is defined by (4.6) yields

$$\frac{n^{1/2}}{\alpha^*(x)^{1/4}}\,\{E\widetilde{f}^*_{\alpha^*}(x)-f(x)\}=O\left(\frac{C(x)}{n^{(5\delta-2)/4}}\right), \tag{4.9}$$

as $n\to\infty$. Here $C(x)=\{\frac{W\,f(x)}{2\,x^2\,\sqrt{\pi}}\}^{1/2}$. Combining (4.8) and (4.9) yields (4.7). □

5. An application to the excess life distribution

Assume that the random variable X has cdf F and pdf f defined on $[0,\infty)$ with $F(0)=0$. Denote the hazard rate function $h_F=f/S$, where $S=1-F$ is the corresponding survival function of X. Assume also that the sampled density g satisties (1.1) and (1.7). It follows that

$$g(y)=\frac{1}{W}\,\{1-F(y)\},\ \ y\geq 0\,. \tag{5.1}$$

It is also immediate that $W=1/g(0)$ and, $f(y)=-W\,g'(y)=-\frac{g'(y)}{g(0)}$, $y\geq 0$, so that

$$h_F(y)=-\frac{g'(y)}{g(y)},\ y\geq 0\,.$$

Suppose now that we are given n independent copies $Y_1, \ldots, Y_n$ of a random variable Y with cdf G and density g from (5.1).

To recover F or S from the sample $Y_1, \ldots, Y_n$ use the moment-density estimator from Mnatsakanov and Ruymgaart [6], namely

$$\hat{\hat{S}}_\alpha(x) = \frac{\hat{W}}{n} \sum_{i=1}^{n} \frac{1}{Y_i} \cdot \frac{1}{(\alpha-1)!} \cdot \left(\frac{\alpha}{x} Y_i\right)^\alpha \exp(-\frac{\alpha}{x} Y_i)\,. \tag{5.2}$$

Where the estimator $\hat{W}$ can be defined as follows:

$$\hat{W} = \frac{1}{\hat{g}(0)}\,.$$

Here $\hat{g}$ is any estimator of g based on the sample $Y_1, \ldots, Y_n$.

Remark 5.1. As has been noted at the end of Section 1, estimating g from $Y_1, \ldots, Y_n$ is a "direct" problem and an estimator of g can be constructed from (1.5) with W and $w(Y_i)$ both replaced by 1. This yields

$$\hat{g}_\alpha(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{y} \cdot \frac{\alpha-1}{(\alpha-1)!} \cdot \left(\frac{\alpha}{y} Y_i\right)^{\alpha-1} \exp(-\frac{\alpha}{y} Y_i)\,,\ y \geq 0. \tag{5.3}$$

The relations above suggest the estimators

$$\hat{f}(y) = -\frac{\hat{g}'_\alpha(y)}{\hat{g}_\alpha(0)}\,,\quad y \geq 0\,,$$

$$\hat{h}_F(y) = -\frac{\hat{g}'_\alpha(y)}{\hat{g}_\alpha(y)}\,,\quad \hat{w}(y) = -\frac{\hat{g}_\alpha(y)}{\hat{g}'_\alpha(y)}\,,\ y \geq 0\,.$$

Here let us assume for simplicity that W is known and construct the estimator of survival function S as follows:

$$\hat{S}_\alpha(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{W}{Y_i} \cdot \frac{1}{\Gamma(\alpha)} \cdot \left(\frac{\alpha}{x} Y_i\right)^\alpha \exp(-\frac{\alpha}{x} Y_i) = \frac{1}{n} \sum_{i=1}^{n} L_i\,. \tag{5.4}$$

Theorem 5.1. *Under the assumptions* (2.3) *the bias of* $\hat{S}_\alpha$ *satisfies*

$$E\hat{S}_\alpha(x) - S(x) = -\frac{x^2\, f'(x)}{2 \cdot \alpha} + \mathrm{o}\left(\frac{1}{\alpha}\right),\quad \textit{as } \alpha \to \infty. \tag{5.5}$$

For the Mean Squared Error (MSE) *we have*

$$MSE\{\hat{S}_\alpha(x)\} = n^{-4/5} \left[\frac{W \cdot S(x)}{2 \cdot x\sqrt{\pi}} + \frac{x^4\{f'(x)\}^2}{4}\right] + o(1), \tag{5.6}$$

provided that we choose $\alpha = \alpha(n) \sim n^{2/5}$.

Proof. By a similar argument to the one used in (3.8) and (3.10) it can be shown that

$$\begin{aligned} E\hat{S}_\alpha(x) - S(x) &= \int_0^\infty h_{\alpha,x,1}(u)\{S(u) - S(x)\}du \\ &= -\frac{1}{2}\frac{x^2}{\alpha} f'(x) + o\left(\frac{1}{\alpha}\right),\ \text{as}\, \alpha \to \infty \end{aligned} \tag{5.7}$$

and

$$E\,L_i^2 = \frac{W}{2\sqrt{\pi}}\frac{\sqrt{\alpha}}{x}S(x) + o(\sqrt{\alpha}),\ \text{as}\,\alpha \to \infty\,, \tag{5.8}$$

respectively. So that combining (5.7) and (5.8) yields (5.6). □

Corollary 5.1. *If the parameter $\alpha = \alpha(x)$ is chosen locally for each $x > 0$ as follows*

$$\alpha(x) = n^{2/5} \cdot \{\frac{\pi}{4 \cdot W^2}\}^{1/5} \cdot x^2 \cdot \left[\frac{f'(x)}{\sqrt{1-F(x)}}\right]^{4/5},\ f'(x) \neq 0. \tag{5.9}$$

then the estimator (5.4) *with $\alpha = \alpha(x)$ satisfies*

$$MSE\{\hat{S}_\alpha(x)\} = n^{-4/5}\left[\frac{W^2 \cdot f'(x) \cdot (1-F(x))^2}{\pi\sqrt{2}}\right]^{2/5} + o(1),\ \text{as}\ n \to \infty.$$

Theorem 5.2. *Under the assumptions* (2.3) *and $\alpha = \alpha(n) \sim n^\delta$ for any $0 < \delta < 2$ we have, as $n \to \infty$,*

$$\frac{\hat{S}_\alpha(x) - E\hat{S}_\alpha(x)}{\sqrt{Var\,\hat{S}_\alpha(x)}} \to_d \ Normal\,(0,1)\,. \tag{5.10}$$

Theorem 5.3. *Under the assumptions* (2.3) *we have*

$$\frac{n^{1/2}}{\alpha^{1/4}}\,\{\hat{S}_\alpha(x) - S(x)\} \to_d \ Normal\left(0, \frac{W \cdot S(x)}{2\,x\,\sqrt{\pi}}\right), \tag{5.11}$$

as $n \to \infty$, provided that we take $\alpha = \alpha(n) \sim n^\delta$ for any $\frac{2}{5} < \delta < 2$.

Corollary 5.2. *If the parameter $\alpha = \alpha(x)$ is chosen locally for each $x > 0$ according to* (5.9) *then for $\hat{S}_\alpha(x)$ defined in* (5.4) *we have*

$$\frac{n^{1/2}}{\alpha(x)^{1/4}}\,\{\hat{S}_\alpha(x) - S(x)\} \to_d \ Normal\left(-[\frac{W\,S(x)}{2\,x\,\sqrt{\pi}}]^{1/2}, \frac{W\,S(x)}{2\,x\,\sqrt{\pi}}\right),$$

provided $f'(x) \neq 0$ and $n \to \infty$.

Corollary 5.3. *If the parameter $\alpha = \alpha^*(x)$ is chosen locally for each $x > 0$ according to*

$$\alpha^*(x) = n^\delta \cdot \{\frac{\pi}{4 \cdot W^2}\}^{1/5} \cdot x^2 \cdot \left[\frac{f'(x)}{\sqrt{1-F(x)}}\right]^{4/5},\ \frac{2}{5} < \delta < 2\,, \tag{5.12}$$

then for $\hat{S}_{\alpha^*}(x)$ defined in (5.4) *we have*

$$\frac{n^{1/2}}{\alpha^*(x)^{1/4}}\,\{\hat{S}_{\alpha^*}(x) - S(x)\} \to_d \ Normal\left(0, \frac{W\,S(x)}{2\,x\,\sqrt{\pi}}\right),$$

provided $f'(x) \neq 0$ and $n \to \infty$.

Note that the proofs of all statements from Theorems 5.2 and 5.3 are similar to the ones from Theorems 4.1 and 4.2, respectively.

6. Simulations

At first let us compare the graphs of our estimator $\widehat{f}^*_\alpha$ and the kernel-density estimator f_h proposed by Jones [4] in the length-biased model:

$$f_h(x) = \frac{\hat{W}}{n\,h} \sum_{i=1}^{n} \frac{1}{Y_i} \cdot K\Big(\frac{x - Y_i}{h}\Big),\ x > 0\,. \tag{6.1}$$

Assume, for example, that the kernel $K(x)$ is a standard normal density, while the bandwidth $h = O(n^{-\beta})$, with $0 < \beta < 1/4$. Here $\hat{W}$ is defined as follows

$$\hat{W} = \Big(\frac{1}{n} \sum_{j=1}^{n} \frac{1}{Y_j}\Big)^{-1}.$$

In Jones [4] under the assumption that f has two continuous derivatives, it was shown that as $n \to \infty$

$$\begin{aligned} MSE\{f_h(x)\} &= \operatorname{Var} f_h(x) + \operatorname{bias}^2\{f_h\}(x) \\ &\sim \frac{W f(x)}{n\,h\,x} \int_0^\infty K^2(u)du + \frac{1}{4}\,h^4\{f''(x)\}^2 \Big[\int_0^\infty u^2 K(u)du\Big]^2. \end{aligned} \tag{6.2}$$

Comparing (6.2) with (3.12), where $\alpha = h^{-2}$, one can see that the variance term $\operatorname{Var} \widehat{f}^*_\alpha(x)$ for the moment-density estimator could be smaller for large values of x than the corresponding $\operatorname{Var}\{f_h(x)\}$ for the kernel-density estimator. Near the origin the variability of f_h could be smaller than that of $\widehat{f}^*_\alpha$. The bias term of $\widehat{f}^*_\alpha$ contains the extra factor x^2, but as the simulations suggest this difference is compensated by the small variability of the moment-density estimator.

We simulated $n = 300$ copies of length-biased r.v.'s from gamma $G(2, 1/2)$. The corresponding curves for f (solid line) and its estimators $\widehat{f}^*_\alpha$ (dashed line), and f_h (dotted line), respectively are plotted in Figure 1. Here we chose $\alpha = n^{2/5}$ and

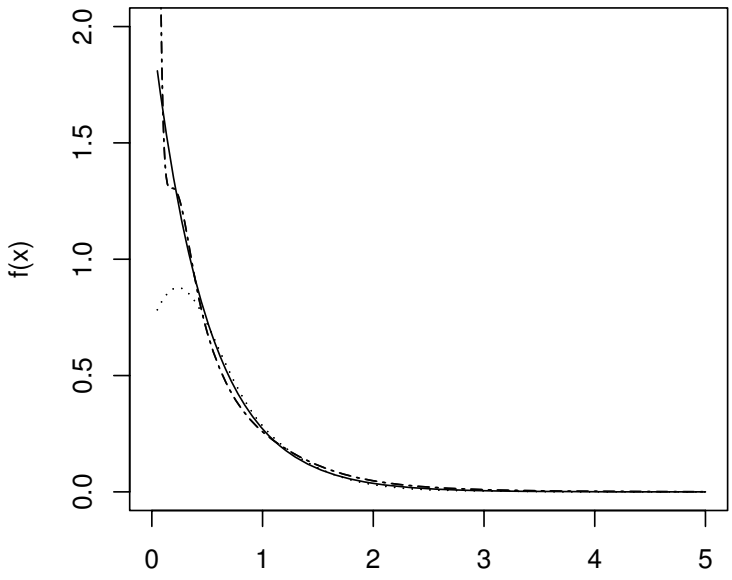


Figure 1.

FIG 1.

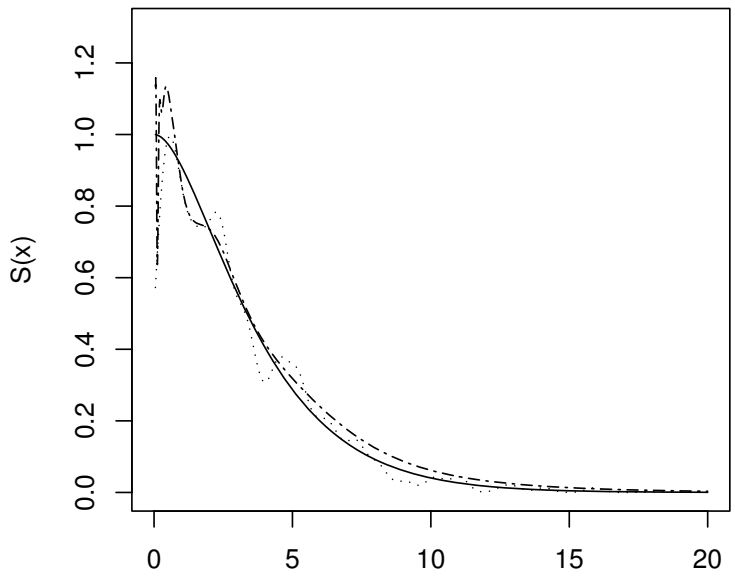


Figure 2.

FIG 2.

$h = n^{-1/5}$, respectively. To construct the graphs for the moment-type estimator $\hat{S}_\alpha$ defined by (5.4) and the kernel-type estimator S_h defined in a similar way as the one given by (6.1) let us generate $n = 400$ copies of r.v.'s $Y_1, \ldots, Y_n$ with pdf g from (5.1) with $W = 4$ and

$$1 - F(x) = e^{-\frac{x}{2}} + \frac{x}{2}\, e^{-\frac{x}{2}}\,, x \geq 0\,.$$

We generated $Y_1, \ldots, Y_n$ as a mixture of two gamma $G(1,2)$ and $G(2,2)$ distributions with equal proportions. In the Figure 2 the solid line represents the graph of $S = 1 - F$ while the dashed and dotted lines correspond to $\hat{S}_\alpha$ and S_h, respectively. Here again we have $\alpha = n^{2/5}$ and $h = n^{-1/5}$.

References

[1] BHATTACHARYYA, B. B., KAZEMPOUR, M. K. AND RICHARDSON, G. D. (1991). Length biased density estimation of fibres. *J. Nonparametr. Statist.* **1**, 127–141.

[2] COX, D.R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (Johnson, N.L. and Smith, H. Jr., eds.). Wiley, New York, 506–527.

[3] FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. II. Wiley, New York.

[4] JONES, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511–519.

[5] MNATSAKANOV, R. AND RUYMGAART, F. H. (2003). Some properties of moment-empirical cdf's with application to some inverse estimation problems. *Math. Meth. Stat.* **12**, 478–495.

[6] MNATSAKANOV, R. AND RUYMGAART, F. H. (2004). Some properties of moment-density estimators. *Math. Meth. Statist.*, to appear.

[7] Mnatsakanov, R. and Ruymgaart, F.H. (2005). Some results for moment-empirical cumulative distribution functions. *J. Nonparametr. Statist.* **17**, 733–744.
[8] Ross, S. M. (2003). *Introduction to Probability Models*. Acad. Press.
[9] Vardi, Y. (1985). Empirical distributions in selection bias models (with discussion). *Ann. Statist.* **13**, 178–205.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 334–339
© Institute of Mathematical Statistics, 2006
DOI: 10.1214/074921706000000545

A note on the asymptotic distribution of the minimum density power divergence estimator

Sergio F. Juárez[1] **and William R. Schucany**[2]

Veracruzana University and Southern Methodist University

Abstract: We establish consistency and asymptotic normality of the minimum density power divergence estimator under regularity conditions different from those originally provided by Basu et al.

1. Introduction

Basu et al. [1] and [2] introduce the *minimum density power divergence estimator* (MDPDE) as a parametric estimator that balances infinitesimal robustness and asymptotic efficiency. The MDPDE depends on a tuning constant $\alpha \geq 0$ that controls this trade-off. For $\alpha = 0$ the MDPDE becomes the maximum likelihood estimator, which under certain regularity conditions is asymptotically efficient, see chapter 6 of Lehmann and Casella [5]. In general, as α increases, the robustness (bounded influence function) of the MDPDE increases while its efficiency decreases. Basu et al. [1] provide sufficient regularity conditions for the consistency and asymptotic normality of the MDPDE. Unfortunately, these conditions are not general enough to establish the asymptotic behavior of the MDPDE in more general settings. Our objective in this article is to fill this gap. We do this by introducing new conditions for the analysis of the asymptotic behavior of the MDPDE.

The rest of this note is organized as follows. In Section 2 we briefly describe the MDPDE. In Section 3 we present our main results for proving consistency and asymptotic normality of the MDPDE. Finally, in Section 4 we make some concluding comments.

2. The MDPDE

Let G be a distribution with support $\mathcal{X}$ and density g. Consider a parametric family of densities $\{f(x;\theta) : \theta \in \Theta\}$ with $x \in \mathcal{X}$ and $\Theta \subseteq \mathbb{R}^p$, $p \geq 1$. We assume this family is identifiable in the sense that if $f(x;\theta_1) = f(x;\theta_2)$ a.e. in x then $\theta_1 = \theta_2$. The *density power divergence* (DPD) between an f in the family and g is defined as

$$d_\alpha(g,f) = \int_{\mathcal{X}} \left\{ f^{1+\alpha}(x;\theta) - \left(1 + \frac{1}{\alpha}\right) g(x) f^\alpha(x;\theta) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} \mathrm{d}x$$

[1]Facultad de Estadística e Informática, Universidad Veracruzana, Av. Xalapa esq. Av. Avila Camacho, CP 91020 Xalapa, Ver., Mexico, e-mail: sejuarez@uv.mx

[2]Department of Statistical Science, Southern Methodist University, PO Box 750332 Dallas, TX 75275-0332, USA, e-mail: schucany@smu.edu

AMS 2000 subject classifications: primary 62F35; secondary 62G35.

Keywords and phrases: consistency, efficiency, M-estimators, minimum distance, large sample theory, robust.

for positive α, and for $\alpha = 0$ as

$$d_0(g, f) = \lim_{\alpha \to 0} d_\alpha(g, f) = \int_{\mathcal{X}} g(x) \log[g(x)/f(x;\theta)] \mathrm{d}x.$$

Note that when $\alpha = 1$, the DPD becomes

$$d_1(g, f) = \int_{\mathcal{X}} [g(x) - f(x;\theta)]^2 \mathrm{d}x.$$

Thus when $\alpha = 0$ the DPD is the Kullback–Leibler divergence, for $\alpha = 1$ it is the L^2 metric, and for $0 < \alpha < 1$ it is a smooth bridge between these two quantities. For $\alpha > 0$ fixed, we make the fundamental assumption that there exists a unique point $\theta_0 \in \Theta$ corresponding to the density f *closest* to g according to the DPD. The point θ_0 is defined as the target parameter. Let $X_1, \ldots, X_n$ be a random sample from G. The minimum density power estimator (MDPDE) of θ_0 is the point that minimizes the DPD between the probability mass function $\hat{g}_n$ associated with the empirical distribution of the sample and f. Replacing g by $\hat{g}_n$ in the definition of the DPD, $d_\alpha(g, f)$, and eliminating terms that do not involve θ, the MDPDE $\hat{\theta}_{\alpha,n}$ is the value that minimizes

$$\int_{\mathcal{X}} f^{1+\alpha}(x;\theta) \mathrm{d}x - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} f^\alpha(X_i;\theta)$$

over Θ. In this parametric framework the density $f(\cdot;\theta_0)$ can be interpreted as the *projection* of the true density g on the parametric family. If, on the other hand, g is a member of the family then $g = f(\cdot;\theta_0)$.

Consider the score function and the information matrix of $f(x;\theta)$, $S(x;\theta)$ and $i(x;\theta)$, respectively. Define the $p \times p$ matrices $K_\alpha(\theta)$ and $J_\alpha(\theta)$ by

$$K_\alpha(\theta) = \int_{\mathcal{X}} S(x;\theta) S^t(x;\theta) f^{2\alpha}(x;\theta) g(x) \mathrm{d}x - U_\alpha(\theta) U_\alpha^t(\theta), \tag{2.2}$$

where

$$U_\alpha(\theta) = \int_{\mathcal{X}} S(x;\theta) f^\alpha(x;\theta) g(x) \mathrm{d}x$$

and

$$\begin{aligned} J_\alpha(\theta) \;=\; & \int_{\mathcal{X}} S(x;\theta) S^t(x;\theta) f^{1+\alpha}(x;\theta) \mathrm{d}x \\ & + \int_{\mathcal{X}} \left\{ i(x;\theta) - \alpha S(x;\theta) S^t(x;\theta) \right\} \times [g(x) - f(x;\theta)] f^\alpha(x;\theta) \mathrm{d}x. \end{aligned} \tag{2.3}$$

Basu et al. [1] show that, under certain regularity conditions, there exists a sequence $\hat{\theta}_{\alpha,n}$ of MDPDEs that is consistent for θ_0 and the asymptotic distribution of $\sqrt{n}(\hat{\theta}_{\alpha,n} - \theta_0)$ is multivariate normal with mean vector zero and variance-covariance matrix $J_\alpha(\theta_0)^{-1} K_\alpha(\theta_0) J_\alpha(\theta_0)^{-1}$. The next section shows this result under assumptions different from those of Basu et al. [1].

3. Asymptotic Behavior of the MDPDE

Fix $\alpha > 0$ and define the function $m : \mathcal{X} \times \Theta \to \mathbb{R}$ as

$$m(x, \theta) = \left(1 + \frac{1}{\alpha}\right) f^\alpha(x;\theta) - \int_{\mathcal{X}} f^{1+\alpha}(x;\theta) \mathrm{d}x \tag{3.1}$$

for all $\theta \in \Theta$. Then the MDPDE is an M-estimator with criterion function given by (3.1) and it is obtained by *maximizing*

$$m_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)$$

over the parameter space Θ. Let $\Theta_G \subseteq \Theta$ be the set where

$$\int_{\mathcal{X}} |m(x,\theta)| g(x) \mathrm{d}x < \infty. \tag{3.2}$$

Clearly $\theta_0 \in \Theta_G$, but we assume Θ_G has more points besides θ_0. For $\theta \in \Theta_G$ consider the expected value of $m(X;\theta)$ in (3.1) under the true distribution G

$$M(\theta) = \left(1 + \frac{1}{\alpha}\right) \int_{\mathcal{X}} f^{\alpha}(x;\theta) g(x) \mathrm{d}x - \int_{\mathcal{X}} f^{1+\alpha}(x;\theta) \mathrm{d}x, \tag{3.3}$$

and define $M(\theta) = -\infty$ for $\theta \in \Theta \setminus \Theta_G$. Then the target parameter θ_0 is such that $-\infty < M(\theta_0) = \sup_{\theta\in\Theta} M(\theta) < \infty$. Furthermore, we assume that Θ may be endowed with a metric d. Heretofore it is assumed that (Θ, d) is compact. The next theorem establishes consistency of the MDPDE.

Theorem 1. *Suppose the following conditions hold.*

1. *The target parameter $\theta_0 = \arg\max M_{\theta\in\Theta}(\theta)$ exists and is unique.*
2. *For $\theta \in \Theta_G$, $\theta \mapsto m(x,\theta)$ is upper semicontinuous a.e. in x.*
3. *For all sufficiently small balls $B \subset \Theta$, $x \mapsto \sup_{\theta\in B} m(x,\theta)$ is measurable and satisfies*

$$\int_{\mathcal{X}} \sup_{\theta\in B} m(x,\theta) g(x)\, dx < \infty.$$

Then any sequence of MDPDEs $\hat{\theta}_{\alpha,n}$ that satisfies $m_n(\hat{\theta}_{\alpha,n}) \geq m_n(\theta_0) - o_p(1)$, is such that for any $\epsilon > 0$ and every compact set $K \subset \Theta$,

$$P(d(\hat{\theta}_{\alpha,n}, \theta_0) \geq \epsilon, \hat{\theta}_{\alpha,n} \in K) \to 0.$$

Proof. This is Theorem 5.14 of van der Vaart [6] page 48. □

The first condition is our assumption of existence of θ_0. It states that θ_0 is an element of the parameter space and it is unique (identifiable). Without this assumption there is no minimum density power estimation to do. Compactness of K is needed for $\{\theta \in K : d(\theta,\theta_0) \geq \epsilon\}$ to be compact; this is a technical requirement to prove the theorem. If Θ is not compact, one possibility is to compactify it. The third condition would follow if $f(x;\theta)$ is upper semicontinuous (trivially if it is continuous) in θ a.e. in x. Finally, the fourth condition is warranted by (3.2) in the interior of Θ_G. Thus we can claim the following result.

Theorem 2. *If condition* 1 *in Theorem* 1 *holds, and if $f^{\alpha}(x;\theta)$ is upper semicontinuous (continuous) in θ in the interior of Θ_G and for a.e. in x, then any sequence $\hat{\theta}_{\alpha,n}$ of MDPDEs such that $m_n(\hat{\theta}_{\alpha,n}) \geq m_n(\theta_0) - o_p(1)$, satisfies $d(\hat{\theta}_{\alpha,n}, \theta_0) \xrightarrow{p} 0$.*

The asymptotic normality of the MDPDE hinges on smoothness conditions that are not required for consistency. These conditions are provided in the two following results.

Lemma 3. *$M(\theta)$ as given by (3.3) is twice continuous differentiable in a neighborhood B of θ_0 with second derivative (Hessian matrix) $\mathsf{H}_\theta M(\theta) = -(1+\alpha)J_\alpha(\theta)$, if:*

1. *The integral $\int_{\mathcal{X}} f^{1+\alpha}(x;\theta)dx$ is twice continuously differentiable with respect to θ in B, and the derivative can be taken under the integral sign.*
2. *The order of integration with respect to x and differentiation with respect to θ can be interchanged in $M(\theta)$, for $\theta \in B$.*

Proof. Consider the (transpose) score function $S^t(x;\theta) = \mathsf{D}_\theta \log f(x;\theta)$ and the information matrix $i(x;\theta) = -\mathsf{H}_\theta \log f(x;\theta) = -\mathsf{D}_\theta S(x;\theta)$. Also note that $[\mathsf{D}_\theta f(x;\theta)]f^{\alpha-1}(x;\theta) = S^t(x;\theta)f^\alpha(x;\theta)$. Use the previous expressions and condition 1 to obtain the first derivative of $\theta \mapsto m(x;\theta)$

$$\mathsf{D}_\theta m(x,\theta) = (1+\alpha)S^t(x;\theta)f^\alpha(x;\theta) - (1+\alpha)\int_{\mathcal{X}} S^t(x;\theta)f^{1+\alpha}(x;\theta)\mathrm{d}x. \tag{3.4}$$

Proceeding in a similar way, the second derivative of $\theta \mapsto m(x;\theta)$ is

$$\begin{aligned}\mathsf{H}_\theta m(x,\theta) &= (1+\alpha)\{-i(x;\theta) + \alpha S(x;\theta)S^t(x;\theta)\}f^\alpha(x;\theta) - (1+\alpha) \\ &\quad \times \left\{\int_{\mathcal{X}} -i(x;\theta)f^{1+\alpha}(x;\theta) + (1+\alpha)S(x;\theta)S^t(x;\theta)f^{1+\alpha}(x;\theta)\mathrm{d}x\right\}.\end{aligned} \tag{3.5}$$

Then using condition 2 we can compute the second derivative of $M(\theta)$ under the integral sign and, after some algebra, obtain

$$\mathsf{H}_\theta M(\theta) = \int_{\mathcal{X}} \{\mathsf{H}_\theta m(x,\theta)\}g(x)\mathrm{d}x = -(1+\alpha)J_\alpha(\theta). \qquad \square$$

The second result is an elementary fact about differentiable mappings.

Proposition 4. *Suppose the function $\theta \mapsto m(x,\theta)$ is differentiable at θ_0 for x a.e. with derivative $\mathsf{D}_\theta m(x,\theta)$. Suppose there exists an open ball $B \in \Theta$ and a constant $M < \infty$ such that $\| \mathsf{D}_\theta m(x,\theta) \| \leq M$ for all $\theta \in B$, where $\| \cdot \|$ denotes the usual Euclidean norm. Then for every θ_1 and θ_2 in B and a.e. in x, there exist a constant that may depend on x, $\phi(x)$, such that*

$$|m(x,\theta_1) - m(x,\theta_2)| \leq \phi(x) \parallel \theta_1 - \theta_2 \parallel, \tag{3.7}$$

and

$$\int_{\mathcal{X}} \phi^2(x)g(x)dx < \infty.$$

We can now establish the asymptotic normality of the MDPDE.

Theorem 5. *Let the target parameter θ_0 be an interior point of Θ, and suppose the conditions of Lemma 3 and Proposition 4 hold. Then, any sequence of MDPDEs $\hat{\theta}_{\alpha,n}$ that is consistent for θ_0 is such that*

$$\sqrt{n}(\hat{\theta}_{n,\alpha} - \theta_0) \rightsquigarrow N_p(0, J_\alpha^{-1}(\theta_0)K_\alpha(\theta_0)J_\alpha^{-1}(\theta_0)),$$

where K_α and J_α are given in (2.2) and (2.3), respectively.

Proof. From Lemma 3, $M(\theta)$ admits the following expansion at θ_0

$$M(\theta) = M(\theta_0) + \frac{1}{2}(\theta - \theta_0)^t V_\alpha(\theta_0)(\theta - \theta_0) + o(\| \theta - \theta_0 \|^2),$$

where $V_\alpha(\theta) = \mathsf{H}_\theta M(\theta)$. Proposition 4 implies the Lipschitz condition (3.7). Then the conclusion follows from Theorem 5.23 of van der Vaart [6]. $\square$

So far we have not given explicit conditions for the existence of the matrices J_α and K_α as defined by (2.3) and (2.2), respectively. In order to complete the asymptotic analysis of the MDPDE we now do that. Condition 2 in Lemma 3 implicitly assumes the existence of J_α. This can be justified by observing that the condition that allows interchanging the order integration and differentiation in $M(\theta)$ is equivalent to the existence of J_α. For J_α to exist we need $i_{jk}(x;\theta)$, the jk-element of the information matrix $i(x;\theta)$, to be such that

$$\int_{\mathcal{X}} i_{jk}(x;\theta) f^{1+\alpha}(x;\theta)\mathrm{d}x < \infty, \text{ and } \int_{\mathcal{X}} i_{jk}(x;\theta) f^{\alpha}(x;\theta) g(x)\mathrm{d}x < \infty.$$

Regarding K_α, let $S_j(x;\theta)$ be the jth component of the score $S(x;\theta)$. If

$$S_j^2(x;\theta) f^{2\alpha}(x;\theta) < C_j, \quad j = 1, \ldots, p, \tag{3.8}$$

then

$$\int_{\mathcal{X}} S_j^2(x;\theta) f^{2\alpha}(x;\theta) g(x)\mathrm{d}x < \infty.$$

Thus, K_α exists. Furthermore, by (3.4) we see that the jth component of $\mathsf{D}_\theta m(x,\theta)$ is

$$A_j = (1+\alpha) S_j(x;\theta) f^{\alpha}(x;\theta) - (1+\alpha) \int_{\mathcal{X}} S_j(x;\theta) f^{1+\alpha}(x;\theta)\mathrm{d}x.$$

Then A_j^2 would be bounded by a constant $M_j < \infty$ if all the components $S_j(x;\theta)$ of the score vector $S(x;\theta)$ satisfy (3.8). This is true because in this case $S_j(x;\theta) f^{\alpha}(x;\theta)$ would be bounded by a constant too, and then

$$\int_{\mathcal{X}} S(x;\theta) f^{1+\alpha}(x;\theta)\mathrm{d}x < \infty.$$

Hence

$$\| \mathsf{D}_\theta m(x,\theta) \|^2 = \sum_{i=1}^{p} A_j^2 \leq \sum_{i=1}^{p} M_j < \infty.$$

Therefore, if (3.8) holds, then the Lipschitz condition in (3.7) follows.

From the previous analysis, we see that the conditions on $M(\theta)$ and $m(x,\theta)$ given in Theorem 5 can be established in terms of the density $f(x;\theta)$, its score vector $S(x;\theta)$, and its information matrix $i(x;\theta)$ as indicated in the next theorem.

Theorem 6. *The MDPDE is asymptotically normal, as in Theorem* 5, *if the following conditions hold in a neighborhood* $B \subseteq \Theta_G$ *of* θ_0*:*

1. *Condition* 1 *of Lemma* 3
2. *For each* $j = 1, \ldots, p$, $S_j^2(x;\theta) f^{2\alpha}(x;\theta) < C_j$ *a.e. in* x.
3. *Suppose there are functions* ϕ_{jk} *such that* $|i_{jk}(x;\theta) f^{\alpha}(x;\theta)| \leq \phi_{jk}(x)$ *for* $j,k = 1, \ldots, p$, *and*

$$\int_{\mathcal{X}} \phi_{jk}(x) f(x;\theta)\,dx < \infty, \text{ and } \int_{\mathcal{X}} \phi_{jk}(x) g(x)\,dx < \infty.$$

4. Concluding Remarks

We have obtained consistency of the MDPDE under rather general conditions on the criterion function m. Namely, integrability of $x \mapsto m(x,\theta)$, and upper semicontinuity of $\theta \mapsto m(x,\theta)$. However these are not necessary conditions; concavity or

asymptotic concavity of $m_n(\theta)$, would also give consistency of the MDPDE without requiring compactness of Θ, see Giurcanu and Trindade [3]. To decide which set of conditions are easier to verify seems to be more conveniently handled on a case by case basis.

Acknowledgements

The authors thank professor Javier Rojo for the invitation to present this work at the Second Symposium in Honor of Erich Lehmann held at Rice University. They are also indebted to the editor for his comments and suggestions which led to a substantial improvement of the article. Finally, the first author is deeply grateful to Professor Rojo for his proverbial patience during the preparation of this article.

References

[1] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1997). Robust and efficient estimation by minimising a density power divergence. Statistical Report No. 7, Department of Mathematics, University of Oslo.

[2] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85** (3), 549–559.

[3] Giurcanu, M., and Trindade, A. A. (2005). Establishing consistency of M-estimators under concavity with an application to some financial risk measures. Paper available at `http:www.stat.ufl.edu/ trindade/papers/concave.pdf`.

[4] Juárez, S. F. (2003). Robust and efficient estimation for the generalized Pareto distribution. Ph.D. dissertation. Statistical Science Department, Southern Methodist University. Available at `http://www.smu.edu/statistics/faculty/SergioDiss1.pdf`.

[5] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

[6] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.