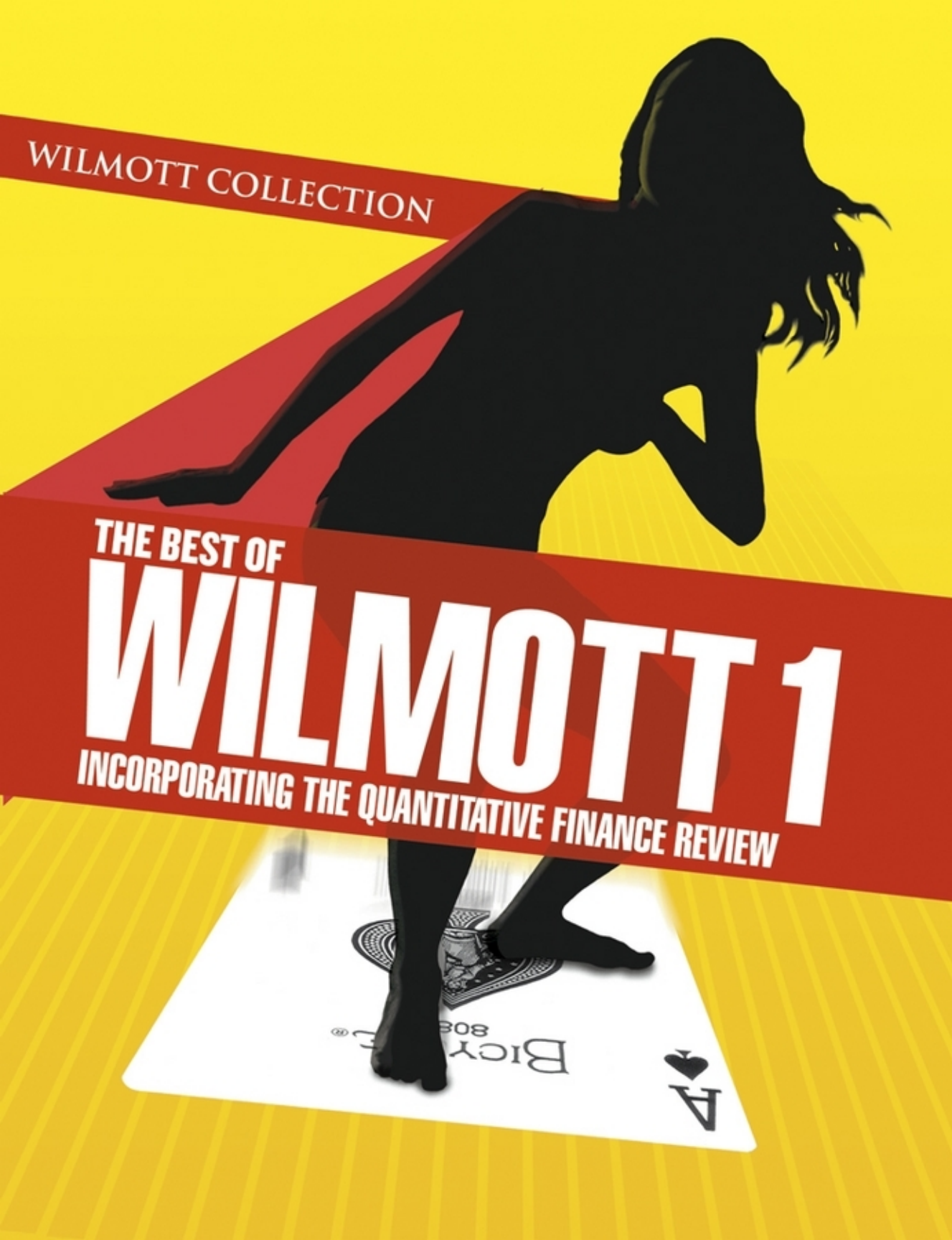
WILMOTT COLLECTION



THE BEST OF WILMOTT 1

INCORPORATING THE QUANTITATIVE FINANCE REVIEW

The Best of Wilmott

Volume 1

Incorporating the Quantitative Finance Review

Edited by

Paul Wilmott

John Wiley & Sons, Ltd

The Best of Wilmott

Volume 1

The Best of Wilmott

Volume 1

Incorporating the Quantitative Finance Review

Edited by

Paul Wilmott

John Wiley & Sons, Ltd

Published in 2005 by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-02351-1

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Introduction

Paul Wilmott

In September 2002 a small, keen group working for a small (but perfectly formed) website, serving a very niche financial market, joined forces with a book publisher to create a new magazine, *Wilmott*, aimed at mathematicians and scientists working in investment banks. The cunning plan was to bring together cutting-edge content, incisive articles and fab design; to combine the logic of the research papers for the left-brainers with an easy-on-the-eye look for the right-brainers. Can't be done, you say. Well, it can, and it was. And it struck a chord with quants everywhere.

The backbone of the magazine is the editor, Dan Tudball. But he does far, far more than just "edit." He writes, coordinates, plans, sketches, designs ... his background as an editor of *FHM* – not quite a "top-shelf mag" but certainly beyond the reach of the children – ensures that the appeal of the magazine is not purely cerebral. The legs and arms of the magazine must be the regular contributors: Aaron Brown, Alan Lewis, Bill Ziemba, Ed Thorp, Espen Haug, Gustavo Bamberger, Henriette Prast, Kent Osband, Mike Staunton and Rudi Bogni. They give the magazine its solid foundation, and when necessary grab the readers by the shoulders and give them a good shaking. The flattering garb would be down to Liam Larkin, our principal designer. His eye-catching layouts and covers are one of the reasons why magazines tend to "disappear" – he has unwittingly highlighted the "quant as jackdaw" effect. Graham Russel of John Wiley & Sons is the eyes and ears of the magazine, monitoring the final product and liaising with subscribers. Enough with the analogies and body parts already. I must also thank all of the technical writers. You submitted excellent research material, and by submitting to a start-up publication you showed great faith in us. Thank you for that faith. This book contains a selection of the research papers from the magazine's first year.

As I said, a chord was struck and we have not looked back. A natural development for a magazine in this field is to run events. We started small with our "Finance Focus" events in the Financial World Bookshop, London. These are free, open to members of wilmott.com and magazine subscribers. They now run every month, attracting a crowd of quants for a good lecture, a spot of casual networking and free food and drink. The success of the Finance Focus events gave us the confidence to run our first conference, the Quantitative Finance Review 2003. You can read about this event, and then see the write-up of the lectures themselves inside this book.

E-mail: paul@wilmott.com

I would like to thank magazine and event sponsors and advertisers: Algorithmics, BRODA, Commodity Appointments, d-fine, FinAnalytica, Fleet Search and Selection, GARP, GFI, Harry's Bar, Investment Analytics, ITO33, ITWM, London Business School, Millar Associates, Murex, SciComp, Shepherd Little, Statman Consulting and Wolfram. Special thanks go to Tamara Jacobs and Owen Walsh of FinancialCAD, who sponsored the Quantitative Finance Review 2003. Owen's perspective on 2003 and the Review may be read inside.

Finally, I would like to thank our partners in crime, 7city Learning. As well as being Europe's most successful financial training company they are also our partners in the Finance Focuses and the Certificate in Quantitative Finance, and hosted the Quantitative Finance Review 2003. Thank you, gentlemen, it's been emotional.

Paul Wilmott
2004

I
Education in Quantitative Finance

Riaz Ahmad

Quantitative Finance Review 2003

"Quantitative Finance" as a branch of modern banking is one of the fastest growing areas within the corporate arena. This, together with the sophistication of modern and complex financial products, has acted as the motivation for new mathematical models and the subsequent development of mathematical and computational techniques. Investment decisions for predicting risk and return are being increasingly based on principles taken from the Quantitative Finance arena, providing a challenge for both academics and practitioners. Consequently, a solid command of the fundamentals and techniques of mathematical finance is essential for a responsible approach to the trading, asset management, and risk control of complex financial products.

Although relatively young, financial mathematics has developed rapidly into a substantial body of knowledge. Seventy-three years before the pioneering days of Fischer Black, Myron Scholes and Robert Merton, Louis Bachelier had derived the price of an option where the share price movement was modelled by a Wiener process and derived the price of what is now called a barrier option.

Quantitative Finance encompasses the complete range of pure and applied mathematical subjects, which include probability and statistics, partial differential equations, mathematical physics, numerical analysis and operational research. The result has been an extraordinary number of quantitative-based scientists from a wide variety of backgrounds moving into this area of research. In addition, the interdisciplinary nature of this subject matter has meant successful collaborative work being conducted by economists, finance professionals, theoretical physicists, mathematicians and computer scientists.

Mathematical finance has the attraction of being one of a few areas of mathematics that plays a central role in current developments in its domain of application. It has a reciprocal relationship with the "real world" while it both draws from and has direct implications upon everyday financial practice in the commercial arena. Its numerous applications have become an integral and visible part of the daily functioning of global financial institutions.

E-mail: r.ahmad@7city.com

While there continues to be a great demand for education in quantitative finance, the delivery of quality-based training in this area remains a premium. An education that is both demanding in mathematics and related to practice, concurrently, has become a joint concern and a success factor for both educational bodies and the capital markets. In addressing these concerns, *Wilmott* and 7city Learning have created a most successful partnership.

The Certificate in Quantitative Finance (CQF) designed by Dr Paul Wilmott is a six-month intensive course offering advanced instruction in the mathematical/quantitative methods applied to investment banking and finance. Delegates (who can also follow the course using the distance-learning programme) working through weekly problem sheets and monthly exams further develop these skills. Additional classes in the form of mathematics lectures, tutorials and computer workshops are arranged throughout to further support and complement the core teaching.

A consequence of delegates returning to mathematics-based education was the need to offer "refresher-type" courses in calculus, differential equations, linear algebra and probability – giving rise to the Mathematics Primer. These have been extremely popular for prospective CQF delegates who have felt "rusty" due to a long period away from the mathematics learning/application environment. Current plans are to also develop this two/three-day primer as a separate entity for individuals in industry wishing to take a crash course in basic mathematics – the type covered in the first term of a university mathematics undergraduate course.

With increased computing speed and the need for efficient and economical computation in the financial markets, together with recognizing C++ as the primary mode of technology in the banking arena, a beginner's course in C++ was launched. This provides an introduction to programming concepts with applications to modelling in quantitative finance. Delegates with no previous background in C++ are taken through the rudiments of this OO-based language for problem solving in areas such as Monte Carlo and Finite Difference methods and various other computational techniques of use in derivative pricing.

The year culminated with the Quantitative Finance Review (QFR). This one-day conference, headlined by Ed Thorp and held at 7city's facility in London, was a meeting designed for quantitative analysts, by like-minded professionals, to allow speakers and delegates the opportunity to meet and discuss current ideas in the field of Quantitative Finance. The QFR managed to draw together many strands, reminding us where we have come from and the need to always reach for the next innovation – as epitomized by the work of Ed Thorp. It provided us with a focus on the quandaries surrounding many tools we take for granted, Peter Jäckel's talk on Monte Carlo methods, Aaron Brown's discussion of the chord of association, Ephraim Clark's examination of sovereign issues. And we were also presented with new ways of thinking, behavioural finance being given a showcase through Henriette Prast's presentation. The QFR truly represented the trends and currents within this community. Much of the material presented that memorable day can be found in this volume.

The first formal year of *Wilmott* and 7city as providers of quality-based mathematical finance education has been a great success, exceeding by far all initial expectations. It is currently a very exciting period within the quantitative finance field, no more so than in the education of such a dynamic area of applicable mathematics. We are looking forward to building on current programmes and initiatives to further develop the range of both education/training

related products and finance meetings to continue to offer a superior range of products for the banking and finance community.

If you are interested in obtaining further information, please contact me on r.ahmad@7city.com

Dr Riaz Ahmad
CQF Course Director

II FinancialCAD®

Owen Walsh

Quantitative Finance Review 2003

FinancialCAD Corporation, a leading provider of derivative risk management software and services, is a proud sponsor of *Wilmott* magazine and *Wilmott* events such as the Quantitative Finance Review, held November 2003. We believe in our partnership with *Wilmott*, not only because it is a high calibre magazine that has quickly captured significant readership in the past year, but because FinancialCAD shares the vision of *Wilmott* – encouraging the exchange of ideas between quantitative finance professionals to facilitate progress in the field.

For those of you not familiar with FinancialCAD, we have built our success on encouraging an open dialogue with our clients that is fed directly into our products. But it's not as easy as it sounds, as meeting our customer's needs for new or latest industry-proven ways of modelling and measuring the risk of derivatives takes a consistent and concerted effort. The result, however, is that our industry-standard financial analytics and technology have been helping over 25,000 end-users, during a period of 12 years, add value to their businesses.

Like *Wilmott*, FinancialCAD also sees its role in the industry as bridging the gap between financial academics and the financial industry to create the best possible ideas and solutions for the finance practitioner today. The Quantitative Financial Review 2003 was another example of the kind of strategic discussion and dialogue that can occur from this exchange of academic and industry ideas. I left the seminar stimulated by the discussion on credit derivatives modelling, but my thoughts were quickly overtaken by how I might still be able to profit from Ed Thorp's gambling models.

With the world of derivatives in 2003 seeing ever-increasing derivatives transactions in the market-place, while at the same time, ever-increasing regulations to govern these transactions, the year 2004 should prove to be interesting.

Now I must finish reading the latest issue of *Wilmott*!

Cheers!

Owen Walsh

Vice President of Analytics, FinancialCAD®

Contact address: VP Analytics, FinancialCAD Corp, 7455, 132nd Street, Suite 100, Surrey BC, Canada V3W 1J8
E-mail: walsh@fincad.com Telephone: 604 507 2763 www.financialCAD.com

III
Quantitative Finance Review 2003

Dan Tudball

Quantitative Finance Review 2003

The first Wilmott Quantitative Finance review gathered together some of the industry's leading lights.

On 11 November 2003 the first conference designed for quants by quants took place in London. Rather than taking the approach of a three-ring circus, which seems to be the norm these days, the QFR 2003 was designed for a relatively small number of delegates over the course of one day. The structure of the Review, held at the headquarters of 7city Learning in the city of London, allowed speakers and delegates the opportunity to meet and discuss ideas without the clamour of hundreds of vendors trying to grab your attention. With a special focus on credit derivatives, given the phenomenal development of the market over the past three years, a stellar group of speakers were gathered together for what proved to be both enlightening and entertaining presentations. In attendance to chair the proceedings was Owen Walsh of Fincad, the Canadian-based software company whose technical perspective on the needs of the market had them ideally placed as platinum sponsors for the event.

Headlining the event was Ed Thorp, a folk hero of the quantitative finance community. It is rare to have Thorp speak at such events, and *Wilmott* was particularly honoured to play host to this most influential of thinkers. Speakers were drawn from various sectors to provide their perspective on the year's activities. Aaron Brown and Henriette Prast are both well known to readers of this magazine, Philipp Schönbucher, author of what will come to be known as a seminal work on credit derivatives, Ephraim Clark from the University of Middlesex, Elie Ayache of ITO33, and Hugues Pirotte Spéder from the Solvay Business School, and the Monte Carlo man himself, Peter Jäckel.

E-mail: dan@wilmott.com

Ed Thorp

"Worth the price of admission," one gushing delegate (who will remain nameless) informed us after Thorp's review of his career. The conference could not have started on a better note, as Thorp discussed his famous cracking of blackjack before taking us through a history of his involvement in the markets, which served to provide those present in the room with a context which is sometimes overlooked. Thorp's commitment to scientific advance and integrity shone through, taking in his work with Sheen Kassouf on warrants in the 1960s, the setting up of the first market-neutral hedge fund in the early 1970s, his preempting of the Black–Scholes formula, statistical arbitrage techniques: up–down, industry clusters, factor models. He wrapped up with some reflections on the current market before being mobbed by journalists.

Prof Dr Philipp Schönbucher

Currently teaching at ETH in Zurich, Schönbucher has quickly made a name for himself in the world of credit derivatives theory. His book *Credit Derivatives Pricing Models: Models, Pricing and Implementation* has been widely praised as a landmark work in the field. Aside from a little slapstick work with the LCD projector and some good-natured heckling from Peter Jäckel, Schönbucher was able to deliver his paper on portfolio credit risk models for CDs in relative peace. His discussion took in single tranche products and tracers as hedge instruments. He examined current benchmarks: Gauss copula, Vasicek model, and remarked on some strange properties of Gaussian and t-copulae.

Ephraim Clark

Director of Countrymetrics, Clark is the authority on sovereign credit risk. His presentation provided a fascinating look at how sovereign risk could be more analogous to corporate credit risk than we might first assume. Despite having rushed up from Paris that morning, Clark was unruffled and quickly had the audience engrossed. Of particular note was his analysis of sovereign debt and ratings migrations, new modelling techniques and new techniques for parameter estimation.

Hugues Pirotte Spéder

Responsible for the corporate finance module at the Swiss Accounting Academy, Spéder is also a co-founder of Finmetrics, which specializes in advisory and programming development of risk management for banks, pension funds and corporate treasuries. Spéder's presentation covered some of the intricacies currently at play in the credit derivatives market. Spéder's talk covered the development of technical analysis in credit derivatives from the early days of the first Basel Accord. Of particular note was his focus on the need for sophistication in this burgeoning area; stochastic interest rates, firm-exogenous processes, credit risk portfolio creation.

Aaron Brown

The vice-president of risk architecture at Citigroup was in the unenviable position of bringing everyone back to earth after lunch. A tough job for even the most accomplished speaker, but it was difficult not to be engaged by Brown's fascinating look at 'association modelling'. In his usual inimitable style, Brown quickly refuted common misconceptions before turning to recently available liquid market data in order to discuss how we should account for complex association in correlation-based trading.

Elie Ayache

Wilmott's newest columnist is no stranger to the website, where he is numbersix. Ayache's firm ITO33 produces cutting-edge convertible bond software. The equity-to-credit problem provided the focus of Ayache's talk, which provided a rigorous examination of traditional models, where equity level determines the intensity of default. Utilizing active spreadsheets and real contracts, Ayache demonstrated his approach to optimal hedging.

Peter Jäckel

The man from Monte Carlo was on fine form, having provided a number of belly laughs from the back of the room. When it came time to step up to the pulpit Jäckel was well warmed up. Jäckel provided an absorbing overview of the past and present approaches in stochastic volatility modelling before considering future developments in the light of the demands created by new markets in credit derivatives and increasingly complex instruments.

Henriette Prast

Readers of this magazine are familiar with Dr Prast's column, "Emotionomics". Behavioural finance still produces knee-jerk reactions at the mention of its name; people are either attuned to it or not. Prast began the final lecture, it seemed, speaking to a room that was roughly divided on the subject. However, by the end of her exposition of this newest of approaches in finance, the room seemed converted to her cause, not a hint of cognitive dissonance to be seen. Covering prospect theory as an alternative to expected utility theory and explaining anomalies in the financial markets through behavioural perspectives, Prast's presentation was a fascinating endnote to the Review.

Papers from the QFR 2003 will be published in 2004.

1
Rewind

Dan Tudball

Wilmott magazine, January 2003

Dan Tudball reviews the major events of the year just gone.
Aaron Brown, Vice President of risk architecture at Citigroup, provides perspective.

"Thematic resonance", that's what the literati call it; the consistent and thereby satisfying emergence and re-emergence of the same refrain delivering the moral and political caution of the work. Explicit, implicit – it matters not; their presence differentiates art from mere reportage. At this end of the scriptorial spectrum, however, there just isn't much opportunity to exercise the theme bone on a regular basis. When it's actually happening life rarely imitates art; defining moments are by definition not two-a-penny. But occasionally a poor hack is thrown a bone in the shape of an annual review. There is nothing like the luxury of hindsight to allow a person to discover "meaning" amidst the ebb and flow of events. Post-rationalization is a cakewalk; ask any divorce lawyer. So here we are at a new beginning, 2004 still a babe in arms, and 2003 just so last year and ready to be consigned to history once we've neatly summed it up. But a little like those poor urchins in *A Christmas Carol* who were disappointed to find that the giant bird they had masochistically gathered to view in the poulterer's window each day was gone, we find there is no discernible theme to bear witness to in the end-of-year shop display. And no, Ebenezer is not just around the corner to deliver the sweet in swift succession to the bitter.

The preceding year, 2002, was a gift to theme-hounds. Credit derivatives provided a neat form of industry closure following Enron, Worldcom and other major defaults. The CD market was the success story of the year, the new friendly face of derivatives, and the saviour of many a bank's potentially embarrassed balance sheet. 2002 ended with a huge collective sigh of relief, quickly followed by a gasp of awe at how huge the CD market now was. Nice and tidy. But what of this year just gone? No such luck, sure we may have over-egged the pudding a little just now – there are worthwhile stories to tell and reflect upon, but the experience has been somewhat granular.

E-mail: dan@wilmott.com

Exchanges: Unfinished business

The business of running exchanges has rarely seemed more interesting than this year. Carla Furse had come to the London Stock Exchange the previous spring only to see a carefully negotiated merger with Liffe snatched away from her by Euronext. After much speculation, Clara Furse had been announced Chief Executive of the LSE in 2002. Her coming was trumpeted far and wide as the beginning of a new age for the exchange, which under her steerage would awake like a giant from long slumber and stretch its arms to encircle the financial globe. Acquisitions had been the order of the day and a natural bedfellow surely would be London's Liffe exchange, and from there who knew? Unfortunately, Liffe did not share LSE's views on whose corner it belonged to and struck a deal with Euronext, the conglomeration of the Dutch, Belgian and French exchanges. Following this slap in the face, LSE was left sniffing around for another derivatives partner and lighted upon OM Gruppen, the operator of Sweden's main exchange in Stockholm.

In January LSE paid a hefty £18.2 million for 76 per cent of OM London Exchange, the derivatives branch of OM in Britain. Furse was eager to get a slice of the exchange based derivatives trading market which had, in the preceding four years, seen over 20 per cent growth in Europe. Statistics like this loom large against a greater than 15 per cent drop in normal share trading volumes. If anything, 2003 began in much the way it would continue, with the major themes being consolidation and control. And a certain amount of dragging, kicking and screaming.

Good times

According to the Bank of International Settlements (BIS), the organized exchanges posted much stronger six-month growth to June 2003 than the OTC market, with a 61 per cent increase in notional amounts outstanding (versus 20 per cent in OTC) in the first half of 2003. "This contrasts with the second half of 2002, when the stock of OTC contracts had increased by 11 per cent whilst the exchanges stagnated. The stronger increase of activity at exchanges occurred both in the foreign exchange and interest rate segments with exchange-based contracts showing increases of 42 per cent and 65 per cent, respectively." LSE's bugbears, Euronext.liffe and Eurex, had already made successful plays into the US markets by the end of 2002. LSE had meanwhile seen an ill-fated partnership with NASDAQ hit the rocks. Eurex began the year livid with its erstwhile American partner CBOT, with whom the German-Swiss bourse had begun an electronic trading venture. The Board of Trade had announced, on 9 January, that they would not be going any further with Eurex after the expiry of their contract at the end of 2003.

Eurex swiftly responded on the same day with the announcement that upon expiry of that contract it would launch its own American exchange. CBOT had opted to switch to Euronext.liffe, who promised to match Eurex technology, but on a flat fee basis instead of the potentially lucrative revenue-sharing arrangement, which Eurex had insisted on. The two European exchanges have shaken things up to some extent in the States. Fully electronic exchanges expose the inefficiencies and unfairness of open-outcry, making the system look anachronistic. Dealing spreads tend to be wider in open-outcry, which is great for the exchange who will be able to profit from larger commissions – and looking at Chicago in particular, where the three major exchanges are mutually owned by the traders on the floor, it's no surprise that open-outcry has lasted so long. They have all had to become hybridized exchanges, providing

both physical and electronic trading – the sheer volumes of heavily traded contracts, such as government bond futures, demand the speed and fairness that electronic trading can provide.

In credit?

By the end of 2002 credit derivatives were inescapable. Everyone who wasn't already in was talking about getting in. If you were talking this way this time last year you were already too late. After the tech bubble and the muted performance of equities since 2000, credit derivatives provided an exciting and potentially massive new market. In January 2003 the number on everyone's lips was $2 trillion, this being the estimated size of the global credit derivatives market at that point; the forecast was for a doubling of this figure by the beginning of 2004. There was a lot of mutual backslapping going on within investment banking, too. Credit derivatives were, ahem, credited with being the instruments that allowed the banking system to ride out Enron, Worldcom and other major corporate defaults; it looked like £44,000 dinners for six wouldn't be too far off. It's no secret that banks have become increasingly disenchanted with their role of lender, and over the past half-decade have made moves to reduce the amount of default risk they are exposed to. The credit derivatives market has continued to grow apace through 2003, the American market is nearing the $1 trillion mark, whilst London still represents over 50 per cent of the global market. But the environment is changing.

Ratings agencies, initially fundamental to the credibility of the new market, have throughout the year been sounding a cautionary (albeit far from consistent) note. Collateralized Debt Obligations, particularly, enjoyed huge growth in the years preceding 2003, with a profusion of AAA ratings and barely any CDOs rated below investment grade. For example, the $3.5 billion-worth of CDOs issued by Barclays between 1999 and 2001 included near $3 billion rated AAA by Fitch, the rating agency. Less than $130 million now retain that rating, whilst the original $196 million basket of bonds which fell below investment grade now has grown to cover over $1 billion worth of bonds. CDOs have benefited the investment banking industry through the fees it generates, whilst credit default swaps have, it is argued, helped reduce the concentration of default risk and so diffused the shock of major credit events. All very well, but like water in a length of leaky pipe, bad debt has to go somewhere and, although the banking industry has managed to pop a metaphorical finger on the fracture while it reaches for a more permanent fix, elsewhere the pipe's integrity has to be questioned. The worry is that it is highly likely that the credit markets harbour many disasters just waiting to happen.

The credit markets have risen in tandem with the increasing pace of the consultation toward the new Basel Accord, which potentially makes certain classes of lending unprofitable – say that to medium-sized enterprises – due to the capital margins required. Banks have been offloading large amounts of lending over the last three years, some nearly halving the amount of lending they have on their books. Even without the new capital margin requirements looming in the not too distant future, retail and corporate lending just will not yield the kind of return that shareholders and banks now require. Securitization, syndication and reinsuring seem the ideal way to shift the debts or the risks onto those whose appetites are better suited due to different regulatory limitations and capital costs.

Of course, the poor performance of the equity markets provides the other motivator in the growth of the credit market. Institutions that had depended upon more traditional approaches to investment growth have seen their margins whittled away over the last decade. The insurance industry in particular has bought into the credit markets and has suffered, often embarrassingly

so. Insurance companies are the biggest net sellers of protection in the credit derivatives market; they also constitute the largest single group of buyers of CDOs. Insurance regulations now require that firms disclose their mark to market positions on loans that they have guaranteed – all part of the new commitment to transparency that has been wielded like a blunt instrument since the demise of Enron and its sharp bean counters. The picture has not been a pretty one; firms specializing in credit insurance, like Financial Security Assurance, have seen considerable losses on securitized credits (in FSA's case, around 14 per cent on a $75 billion book). Reinsurers have backed out of the market (as in the case of Swiss Re and Chubb). The worry about the shifting of credit risk out into new hands is a reasonable one, although one can't help but comment on the irony that the insurance industry, whose very life depends upon the calculation and aggregation of risk, should feel so sorely done by this marketplace. There is speculation, for example, that much of the over $30 billion losses on loans to Enron and Worldcom (which have barely made a dent on the loan originator's books) have been buried deep amongst the overall (and non-itemized) losses of insurance companies.

The Basel effect

The realities of Basel II certainly became more tangible in 2003. The consultation process had already provided some strategic impetus to the growth of the credit markets, but the Basel committee has been masterfully neutral on developments in that sector. Perhaps a case of out-of-sight out-of-mind; after all, from a certain point of view it wasn't under their own carpet that the banks had been sweeping their dust. Joking aside, in its June annual review, the Bank of International Settlements highlighted the growth of the market, indicating concern over the reconcentration of risk beyond the view of the authorities. What provided for interesting viewing was the interesting development by which the uber-nanny of monetary policy Alan Greenspan, took to the defence of the oft-maligned instruments in the annual, "Derivatives: work of the devil?" debate.

To digress momentarily from the ongoing saga in the sleepy Swiss burg, in March the Sage of Omaha pronounced derivatives "financial weapons of mass destruction" and warned all off of their use. Berkshire Hathaway is no stranger to the occasional hedge, and most laughed the declaration off as the rantings of a man who had found himself lumbered with a new purchase whose positions were somewhat compromised. Most notable, and what will eventually bring us back to the land of the silver bears, is the fact that Mr Greenspan responded fairly swiftly (for the titans tussle in dramatic slow motion) in May at the 2003 Conference on Bank Structure and Competition, Chicago, Illinois, with the declaration that derivatives were indeed a healthy means of controlling risk.

> I do not say that the success of the OTC derivatives market in creating greater financial flexibility is due solely to the prevalence of private reputation rather than public regulation. Still, the success to date clearly could not have been achieved were it not for counterparties' substantial freedom from regulatory constraints on the terms of OTC contracts. This freedom allows derivatives counterparties to craft contracts that transfer risks in the most effective way to those most willing and financially capable of absorbing them.

The Basel Committee on Banking Supervision proposed adjustments to the contemplated new Capital Accord in response to industry comment. Basel II historically has contemplated requiring banking institutions to incorporate both expected and unexpected loss components

into the Internal Ratings-based (IRB) capital requirement. The Committee now intends the IRB measurement of risk-weighted assets to be based solely on the unexpected loss portion of the IRB calculations. As to expected losses, banks now would compare the IRB measurement of expected losses with the total amount of provisions (both specific and general) that they have made for such losses. If the comparison indicates a shortfall, the shortfall amount would be deducted equally from Tier 1 and Tier 2 capital. Any excess resulting from the comparison would be eligible as an element of Tier 2 capital (up to 20 per cent of the total Tier 2 capital). Comments on this proposed change were due to the Committee by 31 December 2003. The Committee also said generally that it is contemplating simplifying the treatment of asset securitizations, and revisiting the treatment of credit card commitments and of certain risk mitigation techniques. The Committee hopes to have all outstanding issues resolved by mid-2004.

Currency conundrums

Although not in the doldrums as such, and still representing by far the largest asset class traded in OTC markets, foreign exchange had, relative to the ascendant star of credit derivatives, been subdued. The year just past saw a resurgence of trading, however, driven entirely by the reflationary policies of the US administration and a controversial commitment to a weak dollar. Interest rate cuts have been par for the course for the longest time, first in an attempt to kick-start the markets, then in order to get consumers out buying American. Now the US is at loggerheads with China and other developing world economies in their calls for currencies to be taken off the dollar peg and allow the buck to drop. Worries abounded that knock-on to export competitiveness in Europe and Japan would send shockwaves through the markets. The bond markets shuddered at the thought of Asian central banks withdrawing from the Treasuries markets – ill-afforded after the mauling of the summer, and with 36 per cent of treasuries held by overseas investors not to be discounted.

But foreign exchange traders, denied any heavy leverage since the European currencies folded into one, have seen a rebirth. BIS remarked that:

> This is an area which had not seen double-digit growth since the BIS began collecting these statistics. However, in the first half of 2003, outstanding contracts rose by 20 per cent for this risk category. Currency options provided the prime source of momentum, with a surge of 42 per cent. This was particularly marked in non-financial counterparties, where the use of currency options grew by 91 per cent period-on-period.

Brownian motion

Aaron Brown on credit derivatives

"I'm a big fan of credit derivatives. I think credit risk was in the nineteenth century a few years ago and credit derivatives basically took all the stuff that worked for market risk and now credit risk; everyone used to say credit risk is so much harder to deal with, it's so much lumpier, so much more unpredictable, and it just wasn't true – we just didn't have the instruments to handle it. Now we do and we're finding that it's no worse than any other market risk."

"Credit derivatives will continue to grow. Early in the year, JP Morgan settled with Travelers Insurance and 12 other companies with their Enron performance bonds and Morgan took a $1.3

billion charge because of that settlement, which told people that insurance companies are not the place to go to get rid of this credit risk. The ratings agencies basically did not do a very good job predicting any of this stuff; one thing that worked was the credit derivatives, and they worked without a squawk. In all of the credit losses of 2002, we didn't hear one complaint that people had overbought credit derivatives, or that somebody had some they didn't understand, or that an investor was misled about them. The market worked seamlessly and perfectly, and then we had an unexpected credit crunch and it was the most painless one in history. I don't see how it could have worked any better; there are huge pools of capital there that use this stuff and the credit derivatives find the right place. I think a lot of it was insurance companies, and insurance companies are built, with experience of long-term life insurance and pension contracts, to take that kind of credit risk. It hits you really hard, say every five or six years, and you take a big charge, but in the long run it's a very profitable business. If you've got the long-term capital so that you can afford that, it's a much better way for insurance companies and pension funds to be investing than say equities. You know they need that credit risk because they need that return. So I think the system works very well and it's working better as people evolve better products."

"Ratings agencies really failed in the last round of credit problems, they didn't spot any of the telecommunications, they didn't tell anyone about ratings cliffs until after we found out about it the hard way and they were slow to downgrade. I think they did pull themselves together and have reacted well. What we're seeing, for example at Fitch, is that there still is a role for ratings, and ratings will be conservative and slow, and that's good for the market, because ratings are the only time when an analyst will look at the individual company and produce a public report. And people should look at the credit derivative spreads to see the ticking up and ticking down of the probability of default and expected recovery and so on, so I believe there's a natural synergy between the two. They're feeling that now. Another thing I've noticed is that they've started looking very hard at credit derivatives because they realize that's a very important part of credit. Are you hedging? Whose risk are you exposed to? The ratings agencies have gotten a lot more sophisticated about that now."

"Insurance companies are exactly the people to be doing this because it is so sporadic; I do agree you could see a potential for abuse where a hedge fund could be making steady profits for five years then going bust because it bet everything on there being no credit defaults, but again, hedge funds are supposed to be sophisticated and I don't see these things being marketed to retail investors. I suppose if anyone ever starts a credit derivative mutual fund, well to me that's the SEC's problem, a consumer protection issue and not a market issue."

Basel II: an impetus for credit derivatives market growth?

"A year ago a lot of people in a lot of organizations were running around trying to comply with Basel; no one's thinking that way now. They're thinking, we're making ourselves better and we're tightening up a lot of stuff we should have done a long time ago, putting more money into risk management and monitoring; credit derivatives are one aspect of that. We used to see a lot of credit risk build-up that was not very well managed, not closely monitored, not carefully costed out. Once you actually do that then you think you'd better trade this stuff, you have too much exposure to this stuff that wasn't aggregated in one place or monitor on a market basis before. Start charging trading desks with the counterparty risk and they start saying, 'Gee, those BBB counterparties generate big fees, are trading all the time and generating business but they're hitting me with a big capital charge.' When we didn't charge trading desks with

that, then of course all the trading entities were BBB, now of course all the trading desks are concerned about the credit rating of their counterparties. In insurance it used to be that you used to buy a pool, then syndicate it, and every insurance company would get the same fraction of premium whether they were AAA or A. Suddenly people are running around saying that doesn't make a lot of economic sense, we should pay more for AAA than A."

"These are the kinds of things that, through Basel, are rippling through the entire economy. When banks have to do it then they force people in business to think that way and pretty soon everyone thinks that way. I do think there is a general move toward rationality and Basel II is a big part of that, and credit derivatives are one of the major effects. It's misleading to say Basel II will only apply to 12 major US banks; it will essentially apply to every financial services company, large or small, in the world, these things trickle down, filter out, and if your bank is assessing you as a counterparty risk, suddenly you have to think about the way counterparty risk is thought of in Basel. It really does affect everybody; either your banker is going to come up to you and say, 'This is what you have to do,' or you read the documents yourself and start doing it. If you want a good credit rating, if you want to borrow money, if you want to trade with anybody, you must think like a Basel person."

FX markets: new lease of life and the weak dollar policy?

"Without the strengthening of the Euro we wouldn't have seen that. I've been wrong on this for several years but I've been predicting the gradual demise of foreign exchange trading, simply because the economies are closer linked. We're in a world where there are only three main currencies, not sure if the yen is going to be one of those for much longer, the euro and the dollar seem pretty stable, so now we see this huge move on dollar/euro and dollar/yen as well; a new lease of life, but I think that is just temporary. I think we're moving towards a world with one currency in all but name, so I don't think of foreign exchange trading as a long-term proposition.

"FX always does well when there are big moves that no one expects, but I don't know how many more of those we'll have. When we were moving towards the euro there was a lot of interest, but once you had the euro a lot of the inter-Europe FX trading went away, and now that it's stable and strong you can do business these days without worrying about currencies as much as you used to. There are plenty of currencies with wild currencies, but there was never a huge volume of business in there anyway. The volume was all in dollar/mark dollar/yen and so on. In a rational global world there's no need for currency fluctuation, but like I say, I've been saying that for a while and I've been wrong!"

"I would prefer to see a balanced budget, strong dollar and higher interest rates. But given the spending that they've been doing, I think a weak dollar and low interest rates are a reasonable response to that. Macroeconomics is being pushed at the service of politics. I would not want a strong dollar and higher interest rates with the kind of borrowing they've taken on. It doesn't matter so much in the long run. It's not the bread and butter issue it was 20 years ago, I don't think people are suffering, no one's going without their vacations in Europe – it's not an issue."

2
In for the Count

Dan Tudball

Wilmott magazine, September 2002

Ed Thorp cracked blackjack, used the first wearable computer to beat roulette, started the world's first quantitative hedge fund, anticipated the Black–Scholes formulae 5 years in advance, and has maintained consistently excellent returns through nearly 40 years in futures. Dan Tudball reviews the life of one of quantitative finance's great heroes, and speaks to the man himself.

The year is 1938. The place, about 45 miles out of Chicago. On the steps of a market, a boy of not quite six faces off against a perplexed looking local man who holds a heavy tome belonging to the kid, and studies it with some scepticism. "Egbert 802 to 839," the boy begins, quietly, and in a considered tone somewhat beyond his years he continues:

"Ethelwulf 839 to 857, Ethelbald 857 to 860, Ethelbert 860 to 866, Ethelred I 866 to 871, Alfred the Great 871 to 901, Edward I 901 to 924, Ethelstan 924 to 940, Edmund I 940 to 946, Edred 946 to 955, Edwig 955 to 959, Edgar I 959 to 972, Edward II 975 to 978, Ethelred II 978 to 1016, Sven 1013 to 1014, Canute the Great 1016 to 1035, Harold 1036 to 1039, Harthacnute 1039 to 1042, Edward III 1042 to 1066, Harold . . ."

The man's face sets in disbelief as the boy continues his litany; the book he refers to is *A Child's History of England* by Charles Dickens – every entry in the chronology of monarchs recited from this boy's memory to perfection. The boy's father is equally taken aback. Thorp senior had long known his son was a prodigy, but this display is shocking. Only moments before the man had questioned how such a young child could read such a weighty volume. He'd followed that with a challenge to name all the kings and queens of England in order and with the dates of their reigns. And here was the child, nearing the end of his recitation, suddenly looking puzzled himself; "Queen Victoria, I know when her reign began, but I don't know when it ended." But then again, neither did Dickens.

Edward Oakley Thorp was born on 14 August 1932 in Chicago, Illinois. His parents had met in Manila, when Thorp senior was stationed with the Philippine constabulary. That the

E-mail: dan@wilmott.com

child Thorp was different was already evident to his parents when, even at the age of two and a half, he had not yet uttered a single word. This difference was soon recognized as prodigy when the mute child, by then nearly three, was taken on a shopping trip to Montgomery Ward, a department store in Chicago. During a break from the shopping expedition, the young Thorp's parents and friends were sat down trying to induce the child to speak – still a popular pastime in the Thorp household. Some people stepped out of the elevator and someone asked, "Where's the man gone?". Thorp recounts the moment down the line from Newport Beach; " 'Oh, he's gone to buy a shirt,' so everybody's eyes popped out and the next question was, 'Where has the woman gone?' and I answered 'Oh, She's gone to the bathroom to do pee-pee' and their faces turned reddish and they started to ply me with questions."

This revelation motivated Thorp's father to see how much he could teach his young son. Reading primers led swiftly to more complex books and Ed was reading confidently by the age of three and a half; successively more complex books led to the showdown at the market. Between the ages of five and ten, Ed devoured every scrap of reading material he could get his hands on. Concerned that he was becoming too cerebral at the expense of other activities, Ed's parents were concerned he wasn't getting out enough – they started him on building model airplanes, and then bought him a mineral set when he was ten. This was followed by a chemistry set, which really set Ed off. He cordoned off a section of the garage for his "experiments", which allowed an outlet for his fascination with controlled explosions. These explosions led to an explosion of another kind; from chemistry to physics, electronics, astronomy and mathematics. Being most interested in chemistry, he sat for the All-Southern-California high school chemistry test, despite being a few years younger than other students sitting for the exam. He came fourth in that part of the state, and was very proud of that result, but he recalls that the reason why he had only achieved that position rather than coming first was down to a new section requiring slide rules. Ed only had a 10 cent slide rule, which was "a piece of junk" in his hands. He decided to avenge himself the following year by taking the analogous physics test and came out first, by a very large margin. It was this result that got him a scholarship to UC Berkeley – without it, Thorp may not have been able to advance further in education, money was so tight.

Austerity and reason

Ed had grown up in the Depression era, and that defining time had affected him just as deeply as his contemporaries. Even at six years old, Ed had begun formulating ways to assist with the household income.

> "It was a time when everybody was very poor, and I remember getting five cent packs of Kool Aid, making six glasses out of each pack and selling them at a penny a glass to WPA workers out on the streets who got hot and sweaty in the summer. I remember I saved every cent. Fortunately my father was in a moderately secure job and we always had food on the table, but I remember seeing pictures of homeless people in the newspapers, tattered clothes and that sort of thing. It's something that people of that era remember very vividly. Saving everything. And that had an impact on me, I was very frugal for the first twenty-five years of my life and this allowed me to make it through university on very limited means."

With money in short supply, and with a burgeoning interest in expensive experimentation, Ed would deliver newspapers at two to three in the morning in order to fund his science.

When speaking to Thorp about these formative years, a vivid picture is painted of a child whose critical and analytical faculties were highly developed; a child with a preternatural gift for reflection, and independence of thought – only prepared to commit to something, whether an idea or a course of action, after the very deepest consideration. Ed was fascinated with Morse code, and was a radio ham – this in itself deriving from a passion for structure and organization. Certain gifts, like a near-photographic memory, died away as life required the skill less, but even to this day Ed has a facility for two- and three-dimensional geometry, which allows him to mentally map any journey he makes and very quickly draw an accurate map from memory. Other lessons learnt in childhood still resonate with Ed today, such as his commitment to Reason and its values.

> "I think as far as the way I approach things, professionally and otherwise, I'm unusually rational as people go and I don't feel like I have any of the usual areas of irrationality that people have. I don't want to offend but I'll mention things like astrology and tales from olden times about things that allegedly happened. The place I may go wrong from time to time is that I may not have enough experience of some aspects of the world, for instance as I grew up I was very naïve about people. Until I was nine I believed that everything I saw in print was true – I found it impossible to believe otherwise. Until, that is, I saw two newspapers with conflicting information and that particular naïvety disappeared rather rapidly after that."

The deck re-stacked

Back in 1914 Ambrose Bierce wrote, "The gambling known as business looks with austere disfavor upon the business known as gambling." However, during Ed's formative years through to the late 1960s, the notion that the world of finance could derive valuable lessons from the world of casinos had not become the cliché that it is today. As far as gambling was concerned, Thorp's first brush with that world was under the tutelage of an older cousin, who would take his young relative to gas stations that housed illegal slot machines in their washrooms. There, Ed was shown how to jiggle the handles on the machines to pay off when they shouldn't. Naturally, money being scarce at the time, this was a delight – however, it was not until he was at university that Thorp got to seriously thinking about gambling in relation to his innate talents of mathematics and physics. And the challenge that interested him then was less getting the payoff when there shouldn't be one, but rather when there should.

Between 1955 and 1964 Ed was to work on two things that were to have a profound effect, not only on people in Nevada and Atlantic City who sported names such as "The Fish" or "Ax Handle" between their fore- and surnames, but also on every person with the slightest interest in reducing risk. The first was the development of the wearable computer for predicting the outcome on a roulette table. Success there primed him for his approach to blackjack.

By 1958, when Ed first started thinking about blackjack, he had married Vivian and had achieved his PhD at UCLA. Work on roulette had resulted from some idle banter with friends on how to make easy money, back in 1955. This time round it was a trip to Las Vegas for a cheap, non-gambling, vacation that got Ed thinking. At the time the prevailing assumption was that none of the major gambling games allowed for systems. The accepted thought was that because most games depended upon independent trials processes, i.e every spin or dice roll was unaffected by those that preceded it, then there was no way a mathematical system would allow you

to numerically track outcomes and reasonably predict future outcomes. Unless you used rigged dice or had some information on the croupier, you might as well bring along a rabbit's foot as a calculator. Ed had previously concluded (in 1955) that roulette was an exception to this rule, because he wasn't using a numerical system and instead relied on the physical properties of the mechanism. Prior to his trip to Vegas, Thorp had been given a paper, published in the *Journal of the American Statistical Association*, written by US Army mathematicians (Roger R. Baldwin, Wilbur E. Cantey, Herbert Maisel and James P. McDermott) on basic strategy in the game of blackjack. The contention of the paper was that the house edge on blackjack could fall as low as 0.0062 (somewhat later corrected by them to 0.0032). Ed made himself a little reference card to take to the table, purchased ten bucks worth of chips and prepared to test the methodology. Once at the table, he played the game for about 20 minutes – never having played it before, and this being the first time he'd set foot in a Casino – eventually losing the ten dollars, but the important observation he took with him was that he had been losing at a far slower rate than others on the table, and the realization that he could modify the methodology.

Like roulette, blackjack was in fact also an exception to the rule that gambling games couldn't be beaten by fair means. At that time, when a card was dealt it was put aside, thus shifting the composition of the now depleted deck in a set manner, a manner that would favour either the player or the casino. Independent trials processes were not a factor in this game and thus, Ed reasoned, all you needed was a decent frequency of favourable situations and adjustments in the betting spread in order to get the edge. Ed, whether he realized it or not, was on the edge of something himself.

In the fall of 1961 Ed was CLE Moore instructor at MIT and went to Washington to present a paper entitled 'Fortune's Formula' at the American Mathematical Association. After the trip to Nevada, Ed had tested some of his own theories on MIT's own IBM 704 mainframe computer (a far cry from 10 cent slide rules!) and duplicated in a few hours what would have taken over 10,000 man-years of labour on a hand-held calculator. It was these findings that he presented in Washington. At the end of the presentation, all of Ed's mimeographed copies of his report were snapped up as the 300 or more mathematicians in the room rushed the podium. When Thorp had arrived in Washington, he was already aware that the media had whipped up a small storm in advance. An AP reporter had been leafing through the Association's abstracts prior to the meeting and called Ed; this resulted in a story in the *Boston Globe* the next day. Ed recalls that the phone was ringing off the hook for the next four days, his wife Vivian filled an entire legal pad with messages before finally refusing to pick up the phone, and their daughter – for weeks after – would cry at the sound of a phone ringing. The following weeks at MIT, all six faculty secretaries were snowed under by tens of thousands of letters – until the university had to tell Ed to deal with the correspondence himself because the secretaries weren't able to deal with other faculty business.

Lady Luck RIP

Of course, it wasn't just letters. Offers to bankroll Thorp came apace, but Ed didn't take the bait. However, this situation was to change in early 1961. One of the people who had read about Ed's presentation was Emmanuel 'Manny' Kimmel, a professional (and very successful) gambler, whose own background could not be further from that of Thorp's. The story goes that Kimmel was kidnapped as a child and put to sea – he managed to jump ship somewhere in the Far East where he found work on a cattle boat; the work involved a shovel – and from

then on he raised himself. He was a well-known face in the demi-monde and one of the best proposition men in the USA. He had a good, uneducated, intuitive understanding of odds and "proposition bets". Not being a man of letters, Kimmel had Thorp checked out, phoned him, and made his way to Thorp's apartment outside Boston. Kimmel's proposal coincided with Ed's decision to try out his theories in practice, to show sceptics that his theory really worked and in preparation for a book he planned.

> "A lot of people said it was pie in the sky, a half-baked theory, and a few challenged me to actually do it. So, having a childhood experience of actually doing things in science as well as thinking about theories, I knew I had to do it."

Ed goes on to describe the day he met Kimmel:

> "One wintry afternoon in February 1961 we looked out our window and saw a midnight blue Cadillac pull up, but I didn't see the man I'd spoken to on the phone – I saw two young blondes in mink coats and they got out – and tucked as snug as could be between them in a long cashmere coat was Kimmel. He introduced the two blondes as his nieces; I took it on face value but my wife disagreed! She was much more aware of the ways of the world than I was, she was a literature major and very widely read. She's very perceptive about people, what makes them tick and what their hidden agendas are."

Whatever Kimmel's background was, Ed was blissfully unaware of it at the time. Kimmel was later to be immortalized as "Mr X" in *Beat the Dealer*, wherein Thorp goes into some detail about the trip to Reno that Kimmel and another flush gambler, Eddie Hand ("Mr Y"), bankrolled in 1961. It wasn't until the early 1990s that Ed learned about Kimmel's credentials, during a conversation with the author Connie Brook who was working on *Master of the Game*, a book about the creation of Time-Warner. Mr X was closely associated to Longie Zwillman – a.k.a. Mobster Number Two – and had made his money bootlegging and running numbers in the 1930s. Ed recalled that on their first meeting, on a bitter winter's day, Kimmel had told him that he owned 64 parking lots in New York, and due to the weather snowing them out for two days he had lost $1.5 million. Bruck explained to Ed that Kimmel had a controlling stake in Kinney National Services, whose 1960s SEC filings revealed that amongst their assets were indeed 64 parking lots in New York!

Hand and Kimmel had one very simple goal; they wanted to bankrupt Nevada. They were sure that with the help of their secret weapon, "The Professor", and a bankroll of $100,000 this would be a feat both achievable and worth savouring. However, ever the empiricist, Thorp declined the colossal stake and opted instead for the less imposing (but still sizeable) float of $10,000. Four nights into the experiment, Thorp's system was proving unstoppable. The Wagon Wheel in Lake Tahoe saw Thorp play against six dealers in a row, without a break and without losing a cent. Kimmel was also playing at the same table, and Thorp was so into the system that he was able to direct his backer as well as play his own hands. By the time Thorp decided to bring the slaughter to an end and retire to his room, he was $17,000 ahead. But superstition and luck are constant companions to the dyed-in-the-wool gambler; Kimmel could not equate the "streak" with anything but the "fact" that the cards were "hot". The system, The Professor, and good management were but a distant concept to him now. Thorp left, exasperated after trying to convince Kimmel to cash in his chips and return another day.

In less than an hour Kimmel squandered $11,000. After five days the group decided to call it quits. They discovered that, despite Kimmell's voodoo possession, they had still managed to return $21,000 on a capital outlay of $10,000. Thorp returned to Boston financially secure for the first time in his life.

A Question of finance

During the summer of 1964, Ed was at liberty to conduct whatever kind of research he wanted. He decided to spend it educating himself about the stockmarket and see if he could discover a system for giving himself an edge in the stockmarket over the kind of performance people attained by chance. He observed that on average everyone did well in the long run, barring short-term unpleasantries. The summer ended and it was back to work at the university, but his interest picked up again in the summer of 1965, which the Thorps spent in Los Angeles. Ed had sent away to the periodical *Barons* for some information on warrants and, on receiving the material, he noted that they could be mathematically analysed because they were so much simpler than stocks. He saw that most of the variables were captured by the stock, and most of the differential behaviour was between the stock and the warrant, thus he could eliminate most of the variables.

When Ed joined the new University of California campus which opened in Irvine that fall, he ran into Sheen Kassouff, who was also joining the new faculty as an economist. Kassouff had been working on the same idea, but far in advance of Ed, and had actually started trading on it. The two decided to collaborate and the result was the book *Beat the Market*. Kassouff recalled that time "100 years ago"

> "I think at the very first meeting when we went to talk about it, we met in this conference room in the Dean of Social Science's office. One of the people there, the Associate Dean, Julian Feldman, later told me that it was a battle for the chalk between Ed and I ... rat-a-tatting there on the blackboard over pricing and relationships and so on. I think he was very interested in finding some mathematical application to finance. Being a mathematician and being able to apply it to finance and make money from it, was a very interesting endeavor. He convinced me that was true."

Kassouff was impressed with Thorp's sophisticated approach to life.

> "I was a naïve, but not young, academic. That was my very first entry into the world of academics and the life of the mind and I hadn't thought of it in a commercial way, I was more interested in the theoretical underpinnings and of course I was practicing this for a number of years on my own accounts and friends and family, but I never thought of expanding it to a book until I met and talked with Ed. We got a $50,000 advance, which to me was staggering; my annual salary at the time was something like $10,000! I was an assistant professor. So it was going to be a sequel to *Beat the Dealer*. He also wanted to develop some expertise, he liked the impersonality of the financial markets rather than the one-on-one of Las Vegas – where you're actually dealing with the person whose money you're winning, leading sometimes to unhappy kinds of results, whereas in financial markets you don't really know who's on the other side, and that appealed to him a good deal I think."

In late 1967 or early 1968, Ed started trading OTC options. Prior to this he had sat down to figure out what they were actually worth, using integration. He saw there were a few unknown

parameters, so with very little to go on he applied Occam's razor, went for the simplest possible choice, and had a few other reasons for making the choice – actually what Ed had worked out was what would ultimately come to be known as the Black–Scholes formula. Modest to a fault, however, Ed had this to say about it:

> "I just happened to guess the right formula and put it to use some years before it was published. I was convinced it was right because all the tests that I applied to it worked. It did all the right things; it gave all the right values, and had all the right properties. The way you prove it is by using the arbitrage argument. Much later, in 1973, Black sent me a preprint of his paper and wrote that he admired my work, and said that his methods differ from mine in that they go one step further than simply hedging – they make an assumption that if you have a perfect hedge that you should get the same result as if you'd bought a riskless security – that was the key observation. I actually had a note I had made in 1970 saying I ought to pursue that line, but I was so busy trading securities and using the formula that I never took the opportunity. Black and Scholes found the formula in 1969; I was already trading using the formula in 1967/68, trading on OTC Options at the time."

As the 1960s came to a close, word of Ed's investment methods had spread around UCI. By November of 1969 he had a dozen or so individual accounts, which had anywhere between $25 and $100,000 in them. These were put into warrant hedges whilst Ed traded in options on his own account, using his own anticipation of Black–Scholes – Ed didn't apply this to the other accounts because although he had, as he puts it, "guessed" the method, he didn't have what he felt was definite proof. However, he was using that methodology amongst others to evaluate warrant hedges. It was at this time that he ran into the legendary Warren Buffett. The meeting occurred through the auspices of Ralph Gerrard, Dean of the graduate division at the University of California, and one of Buffett's original investors. Gerrard was a relative of Benjamin Graham, the man who single-handedly created modern security analysis and set the highest standards from the 1920s until his retirement in 1955; Graham in turn was Buffett's mentor.

Buffett had decided that stocks were overpriced in 1968 and decided to shut down his partnership, and return the money to all his very happy investors. Gerrard was looking for someone to invest with and had just read *Beat the Market*. Buffett had averaged 24% for the last twelve years, and Gerrard wanted him to take a look at Ed as a candidate for investment. The first meeting was at Gerrard's, where they played bridge and discussed finance; the second was dinner with their wives. After that, Thorp and Buffett never met again, but Gerrard invested. In a recent interview by journalist Ken Kurson, Buffett fondly remembered his meeting with Ed. 1969 saw Convertible Hedge Associates launched. It was the first market-neutral hedge fund utilizing OTC options, convertible bonds, warrants and preferreds. All the hedges were approximately delta-neutral, and all of these four years before either options were listed or the Black–Scholes formula was published.

Consistency to calamity

Between November 1969 and its dramatic demise in 1988 at the hands of Rudi Giulliani, Princeton Newport Partners (formerly Convertible Hedge Associates) demonstrated astounding consistency and growth. Over the 19 years in operation, the total percentage increase was

1382%, an annual compound rate of return of 15.1%. Compare this to the same period in the S&P 500, which saw an increase of 545% and an annualized rate of return of 10.2% or 3 month US Treasury Bills, 345% total increase, 8.1% annualized. Thorp worked on the theory from the West coast, while his associate Jay Regan did the selling and made the transactions in the East.

The days of the raider loomed large in the 1980s, and the poster boy of the period was Michael Milken of Drexel Burnham. His use of bonds to finance second-tier firms, and also the raiders who were proving the bane of the 'light shoe' directors of major companies, made him an obvious target for reprisals. Unfortunately, Milken was also committing a number of excesses and violations of securities laws, trading fast and loose. A close confederate of his was Thorp's partner, Jay Regan. Through Regan, Thorp had met the likes of Milken, who had always acted cordially to the professor. There was nothing to suggest any illegal activity whatsoever.

Rudi Giulliani was then US Attorney for the Southern District of New York I. He saw an opportunity to emulate Tom Dewey, who busted the bootleggers in the 1930s. Milken proved too difficult to get a grip on. The second in line was Robert Freeman at Goldman Sachs, who had been James Regan's roommate at college. Goldman was prime broker for PNP. Giulliani decided that if he applied pressure to Regan, he'd get Freeman and Milken. PNP became the number one target. In December 1987, the ATF, FBI and Treasury came pouring out the elevators at the Princeton offices of the partnership, and they seized hundreds of cartons of records. To Thorp it was all nonsense, but it turned out there were three tapes that would prove to be destructive.

"They found some incriminating stuff," Thorp recalls.

> "Someone at PNP and someone at Drexel were manipulating a new security that Drexel was issuing. They wanted to control the price at issue date, so there was an agreement about what we would buy, and how much and so on. Then there was a stock-parking issue. Someone at Drexel had used up his $25 million capital limit and wanted to put on more positions. Of course, Drexel had a capital limit and didn't want any more positions. What he did was sell part of his positions to Princeton-Newport and agree to buy them back at 20 per cent annualized profit. So this is parking – illegal because it conceals the true ownership of securities."

Giulliani invoked RICO, the first time it was used against a Securities Firm. Two incidents were needed over 10 years to prove a pattern:

> "They tried tax fraud, wire fraud, and mail fraud and so on to try to get us. Tax was a joke, because it turned out we paid taxes on $4 million twice – we made an accounting mistake, so we were owed money. It took ten years for us to get back some of the money, every individual partner had to file separately. I got my money back – but it cost a lot in legal and accounting fees."

The case was brought to trial. Thorp offered to take over the running of the partnership if Regan would step down until proven innocent. Everyone could return and reclaim their share once the trial was over. Regan declined. Thorp didn't want to continue in an atmosphere of suspicion, and the partnership was dissolved.

> "Five people, including Jay Regan, were convicted from the Princeton office, given fines and jail terms. One Drexel trader was given fines and jail terms. They appealed, and it

was found that the judge had given improper instruction to the jury, so it was brought to retrial. Giulliani had gone on to greater things and he couldn't care less. US Attorneys had lost interest because by then they'd gotten Milken, Freeman and so on, so didn't contest. The conclusion was that the defendants were 'Not found guilty' as opposed to 'Found not guilty'. So it's still open. The jail time was light, only three to six months. It was basically a vendetta."

The acrimony, the legal complications, the lack of direct communication decided Thorp to quit the business for a time. He decided that he wanted a smaller shop, a simpler life. In the early 1990s he had done some Japanese warrant trading and Nikkei put options. He shrunk the operation from 40 to 20, then he proceeded to leave warrants, and the staff shrunk to six. In 1991 Thorp was informed by one of his larger investors that one of his products, Statistical Arbitrage, was doing very well. Since 1992 Thorp has been running his Stat Arb operation, and a parallel hedge fund since 1994. When LTCM happened, Statistical Arbitrage positions were one of the few good positions left. Thorp profited as Hedge Funds suffered a run on the bank, liquidating good positions in order to hold on to the bad. Diligence and a supreme commitment to logic and empirical evidence once again proved Thorp right. The irony is that he had been offered a place at LTCM – he had turned it down flat.

"Because I knew two things. I knew Meriwether from Salamon, he was a big roller of the dice, and I believed Scholes didn't understand the risk. I'd had some interchange with Samuel ... and Merton over logarithmic utility; it's a particular prescription for approaching certain risk problems. They made some points that are true – that it's not all things to all people, but that's not an assertion I would ever make. But some of their arguments are wrong. And I could see that they didn't understand how it controlled the danger of extreme risk and the danger of fat tail distributions. So that was a theoretical place that we fundamentally disagreed. It came back to haunt them in a grand way."

The last year has been an interesting one as Thorp watches the flight from equities with great interest. His mind now turns to the future:

"People who run things like Statistical Arbitrage operations have gotten a lot of new money, and many of them have imprudently expanded, their returns have gone negative as a result. People have also found it easy to start up funds of that type, due to the demand – and those people may not be particularly qualified. More money is chasing the same opportunities, thus driving the value of the opportunity down. Our policy has been to stay moderate in size and allow size to fluctuate according to what we see as our near term performance in the market. We shrank to a third then expanded to a half of our peak size. That's where we sit now.

The outlook for equities is not quite as good as it has been over the last century. There are a number of excesses that need yet to be corrected, we seem to be reading daily, week after week, in the press about this. People are used to a high rate of return, now they've seen two–three down years in a row. They tend to overcorrect. They'll flee to other areas, market neutral hedge funds, property–where I live the rise in property prices has been near 20%, in California there's only a two month inventory left of properties at market. Seven months used to be the typical supply. Real estate will run its course, and a thundering herd of investors will run to the next asset class."

A fitting enough image to leave on. Ed Thorp has defined not just what it means to take a quantitative approach to finance over the past half century. His values remind us that it is the evidence of your own eyes, and the power of the intellect, which guard against the temptation to jump on the latest bandwagon until it rolls over a cliff.

The rules of blackjack or 21

The aim of the game for the player is to hold a card count greater than that of the dealer without exceeding 21 (going "bust").

Before any cards are dealt, the player must place his bet in front of his table position. The dealer deals two cards to each of the players, and two to himself (one of the dealer's cards is dealt face up and the other face down). Court cards (kings, queens and jacks) count as 10, the ace counts as either 1 or 11, and all other cards are counted at their face value. The value of the ace is chosen by the player.

If the player's first two cards are an ace and a 10-count card, he has what is known as "Blackjack". If he gets blackjack with his first two cards, the player wins unless the dealer also has a blackjack, in which case it is a standoff or tie (a "push") and no money changes hands. A winning blackjack pays the player 3 to 2.

"Hit" means to draw another card. "Stand" means no more cards are taken. If the player hits and busts, his wager is lost.

The player is also allowed to double the bet on his first two cards and draw one additional card only. This is called "doubling down".

If the first two cards a player is dealt are a pair, he may split them into two separate hands, bet the same amount on each and then play them as two distinct hands. This is called "splitting pairs". Aces can receive only one additional card. After splitting, ace + 10 counts as 21 and not as blackjack.

If the dealer's up card is an ace, the player may take insurance, a bet not exceeding one-half of his original bet. If the dealer's down card is a 10-count card, the player wins 2 to 1. Any other card means a win for the dealer.

It is sometimes permitted to "surrender" your bet. When permitted, a player may give up his first two cards and lose only one half of his original bet.

The dealer must draw on 16 and stand on 17. In some casinos, the dealer is required to draw on soft 17 (a hand in which an ace counts as 11, not one). Regardless of the total the player has, the dealer must play this way.

In a tie no money is won or lost.

Rules differ subtly from casino to casino, as do the number of decks used. The advantage to the dealer is that the player can go bust, losing his bet immediately, yet the dealer may later bust. This asymmetry is the key to the House's edge. The key to the player's edge is that he can vary both his bets and his strategy.

The world's first wearable computer

In spring 1955, Ed Thorp was in his second year of graduate physics at UCLA. At tea time one Sunday he got to chatting with colleagues about how to make "easy money". The conversation turned to gambling, and roulette in particular. Was it possible to predict, at least with some

exploitable degree of accuracy, the outcome of a spin of the wheel? Some of his colleagues, the ones in the know, were certain that the roulette wheels were manufactured so precisely that there were no imperfections that could be discerned, never mind exploited. But Ed's counter to that was simple. If the wheels are so perfect, you should be able to predict, using simple Newtonian principles, the path of the ball and its final resting place.

Ed got to work in the late 1950s, playing around with a cheap miniature roulette wheel, filming and timing the revolutions. He met up with Claude Shannon, the father of Information Theory in 1959, originally to discuss his blackjack results, but the conversation soon turned to other games and roulette in particular. Shannon was fascinated. Shortly afterwards they met up at Shannon's house, the basement of which was packed with mechanical and engineering gadgets, the perfect playground for further roulette experiments.

Ed and Shannon together took the roulette analysis to greater heights, investing $1500 in a full-sized professional wheel. They calibrated a simple mathematical model to the experiments, to try to predict the moment when the spinning ball would fall into the waiting pockets. From their model they were able to predict any single number with a standard deviation of 10 pockets. This converts to a 44 per cent edge on a bet on a single number. Betting on a specific octant gave them a 43 per cent advantage. It is one thing to win on paper, or in the comfort of a basement. It is quite another to win inside a noisy casino.

From November 1960 until June 1961, Ed and Shannon designed and built the world's first wearable computer. The twelve-transistor, cigarette pack-sized computer was fed data by switches operated by their big toes. One switch initialized the computer and the other was for timing the rotation of the ball and rotor. The computer predictions were heard by the computer wearer as one of eight tones via an earpiece. (Ed and Shannon decided that the best bet was on octants rather than single numbers, since the father of Information Theory knew that, faced with n options, individuals take a time $a + \mathrm{b}\ \ln(n)$ to make a decision.)

This computer was tested out in Las Vegas in the summer of 1961. But for problems with broken wires and earpieces falling out, the trip was a success. Similar systems were later built for the Wheel of Fortune which had an even greater edge, an outstanding 200 per cent.

On 30 May 1985, Nevada outlawed the use of any device for predicting outcomes or analysing probabilities or strategies.

Beating the dealer

The first key is in having the optimal strategy. That means knowing whether to hit or stand. You're dealt an eight and a four and the dealer's showing a six, what do you do? The optimal strategy involves knowing when to split pairs, double down (double your bet in return for only taking one extra card), or draw a new card. Thorp used the computer to calculate the best strategies by simulating thousands of blackjack hands. In his best-selling book *Beat the Dealer* (Random House, 1962, revised 1966), Thorp presented tables showing the best strategies.

But the optimal strategy is still not enough without the second key. You've probably heard of the phrase "card counter" and conjured up images of Doc Holliday in a ten-gallon hat. The truth is more mundane. Card counting is not about memorizing entire decks of cards, but about keeping track of the type and percentage of cards remaining in the deck during your time at the blackjack table. Unlike roulette, blackjack has "memory". What happens during one hand depends on the previous hands and the cards that have already been dealt out.

A deck that is rich in low cards, twos to sixes, is good for the House. Recall that the dealer must take a card when he holds sixteen or less; the high frequency of low-count cards increases his chance of getting close to 21 without busting. For example, take out all the fives from a single deck and the player has an advantage of 3.3 per cent! On the other hand, a deck rich in ten-count cards (tens and court cards) and aces is good for the player, increasing the chances of either the dealer busting or the player getting a blackjack (21 with two cards), for which he gets paid at odds of 3 to 2.

In the simplest case, card counting means keeping a rough mental count of the percentage of aces and tens, although more complex systems are possible for the really committed. When the deck favours the player, he should increase his bet; when the deck is against him, he should lower his bet (and this bet variation must be done sufficiently subtly so as not to alert the dealers or pit bosses).

In *Beat the Dealer*, Ed Thorp published his ideas and the results of his "experiments". He combined the card counting idea, money management techniques (such as the Kelly criterion) and the optimal play strategy to devise a system that can be used by anyone to win at this casino game. "The book that made Las Vegas change the rules", as it says on the cover, and probably the most important gambling book ever, was deservedly in the *New York Times* and *Time* bestseller lists, selling more than 700,000 copies.

Passionate about probability and gambling, playing blackjack to relax; however, even Ed himself could not face the requirements of being a professional gambler. "The activities weren't intellectually challenging along that life path. I elected not to do that."

Once, on a film set, Paul Newman asked him how much he could make at blackjack. Ed told him $300,000 a year. "Why aren't you out there doing it?" Ed's response was that he could make a lot more doing something else, with the same effort, and with much nicer working conditions and a much higher class of people. Truer words were never spoken. Ed Thorp took his knowledge of probability, his scientific rigour and his money management skills to the biggest casino of them all, the stock market.

On Thorp

> "Over the years, through Princeton Newport and through his recent ventures, Ed has shown that anomalies can be exploited and successfully traded. In my lectures I use the 1968–1988 Princeton Newport results: 15.9% mean (net) with 4% standard deviation as the standard for superior hedge fund management. Others such as Soros have had higher means but the smoothness of Ed's record rates it right at the top and a challenge for others to duplicate. We all have a lot to learn from Ed and a few of us have had the pleasure to work with him and learn from the master."
>
> Bill Ziemba

> "One time Ed and I attended a fairly large investment conference at La Quinta in the desert near Palm Springs. As an entertainment activity, the conference people were running a 'racetrack' in which they ran films of races and had betting with play money they provided. When it started, Ed looked at the process and said something like, 'I can figure this out'. He stood and thought about it for less than two minutes and then said, 'I've got it'. So we all pooled our money and he placed some bets. An hour or so later

we had cleaned up. As I recall we ended with more 'money' than everyone else put together."

Jerome Baesel, Managing Director, Morgan Stanley Alternative Investment Partners and lead Portfolio Manager on Morgan Stanley's fund of hedge funds. Jerome and Ed worked together at Princeton–Newport Partners for 10 years.

"Despite all his amazing and internationally recognized professional accomplishments, Ed is quite modest and upon a casual meeting with him, a person would not be aware of all his fame. His ethical and moral standards are of the highest quality. He is a very real role model, rare in this day and age. Ed has a great sense of humor and is a wonderful storyteller in person, as you might imagine from his *Beat the Dealer* book. Ed has a large number of personal interests and for each one devours the subject and devises his own quantitative approach. For example, some 20 years ago, Ed and I trained together for some marathons (including Boston and New York). Ed had determined mile markers for a number of routes near his home. I recall Ed then on a training run, looking at his watch and saying that we were running (for example) at a 7 minute 10 second/mile pace. During his competitive running years, Ed kept large quantities of training data, including physiological (pulse rates, etc.) quantities to help him monitor his progress. I'm sure his plots and analyses would be of interest to coaches."

Gordon Shaw, Professor Emeritus of Physics at the University of California, Irvine, and discoverer of the "Mozart Effect".

3
A Perspective on Quantitative Finance: Models for Beating the Market

Ed Thorp

Quantitative Finance Review 2003

This is a perspective on quantitative finance from my point of view, a 45-year effort to build mathematical models for "beating markets", by which I mean achieving risk-adjusted excess returns.

I'd like to illustrate with models I've developed, starting with a relatively simple example, the widely played casino game of blackjack or twenty-one. What does blackjack have to do with finance? A lot more than I first thought, as we'll see.

Blackjack

When I first learned of the game in 1958, I was a new PhD in a part of mathematics known as functional analysis. I had never gambled in a casino. I avoided negative expectation games. I knew of the various proofs that it was not possible to gain an edge in virtually all the standard casino gambling games. But an article by four young mathematicians presented a "basic" strategy for blackjack that allegedly cut the House edge to a mere 0.6%. The authors had developed their strategy for a complete randomly shuffled deck.

But in the game as played, successive rounds are dealt from a more and more depleted pack of cards. To me, this "sampling without replacement", or "dependence of trials", meant that the usual proofs that you couldn't beat the game did not apply. The game might be beatable. I

Contact address: Edward O. Thorp & Associates, 610 Newport Center Drive, Suite 1240, Newport Beach, CA 92660, USA.
E-mail: EOThorp@ix.netcom.com

realized that the player's expectations would fluctuate, under best strategy, depending on which depleted pack was being used. Would the fluctuations be enough to give favorable betting opportunities? The domain of the expectation function for one deck had more than 33 million points in a space with 10 independent variables, corresponding to the 10 different card values in blackjack (for eight decks it goes up to 6×10^{15}).

Voila! There "must be" whole continents of positive expectation. Now to find them. The paper I read had found the strategy and expectation for only one of these 33 million points. That was only an approximation with a smallish but poorly known error term, and it took 12 years on desk calculators. And each such strategy had to address several mathematically distinct decisions at the table, starting with 550 different combinations of the dealer's up card and the player's initial two cards. Nonetheless, I had taken the first step towards building a model: the key idea or "inspiration" – the domain of the visionaries.

The next step is to develop and refine the idea via quantitative and technical work so that it can be used in the real world. The brute force method would be to compute the basic strategy and expectation for each of the 33 million mathematically distinct subsets of cards and assemble a 33 million-page "book". Fortunately, using linear methods well known to me from functional analysis, I was able to build a simplified approximate model for much of the 10-dimensional expectation surface. My methods reduced the problem from 400 million years on a desk calculator to a few hundred years – still too long. However, I was fortunate to have moved to MIT at the beginning of the computer revolution, where as a faculty member I had access to one of the early high-speed mainframe computers, the IBM 704.

In short sessions over several months the computer generated the results I needed in about 2 hours of actual computer time. I was then able to condense all the information into a simplified card counting scheme and corresponding strategy tables that were only moderately more complex than the original "no memory" basic strategy for a complete deck. You can see them in my book *Beat the Dealer*. This work was the second step, the development of the idea into a model that can be tested in the real world – the domain of the quants.

When I announced my results to the mathematical community, they were surprised to see, for the first time, contrary to long-held beliefs, a winning mathematical system for a major casino gambling game. They generally understood and approved. Not so, universally, however. Several casinos said they loved system players and would send a cab to the airport for me. The *Washington Post* greeted me with an editorial in January, 1961, saying:

> "We hear there's a mathematician coming to town who claims he can beat blackjack. It reminds us of an ad we saw: 'Sure fire weed killer – send in $1'. Back came a postcard saying, 'Grab by the roots and pull like hell'."

So I took the third and last step in building a successful model, real world verification – the domain of the entrepreneurs. In 20 hours of full-scale betting on a first casino trip in 1961 I won $11,000, using a $10,000 bankroll. The midpoint of my forecast was a win of $10,000. To convert to today's dollars, multiply by about seven.

To recap, the three steps for a successful market-beating model are (1) idea, (2) development, and (3) successful real world implementation. The relevant skills are (1) visionary, (2) quantitative, and (3) entrepreneurial.

Early on, I assessed the blackjack idea as worth a million dollars to me, meaning that if I chose to focus my time and energy on exploiting it I could personally extract a million after tax dollars, net of costs, within a few years.

The relevance to finance proved to be considerable. First, it showed that the "gambling market" was, in finance theory language, inefficient (i.e. beatable); if that was so, why not the more complex financial markets? Second, it had a significant worldwide impact on the financial results of casinos. Although it did create a plague of hundreds and eventually thousands of new experts who extracted hundreds of millions from the casinos over the years, it also created a windfall of hundreds of thousands of hopefuls who, although improved, still didn't play well enough to have an edge. Blackjack and the revenues from it surged. Third, it popularized a method that could be used to manage risk more generally in the investment world, namely the Kelly criterion, a.k.a. the capital growth criterion, a.k.a. fixed fraction betting, a.k.a. maximizing expected logarithmic utility.

Now a word on what I learned from this about risk control. The field trip to Nevada was bankrolled by two wealthy businessmen/gamblers, whom I chose from the many who sought me out.

(a) They proposed a $100K bankroll.
(b) I reduced it to $10K as a personal safety measure.
(c) I began play by limiting my betting to $1–10 for the first 8 hours – to verify all was as expected (possible cheating?) and that I got used to handling that amount of money.
(d) Once I was used to that, I went to $2–20 for 2 hours.
(e) Next was $5–50 for 2 hours.
(f) This was followed by $25–200 for 3 hours.
(g) Finally I went to $50–500 (full scale) for 20 hours.

The idea, which has been valuable ever after, was to limit my bet size to a level at which I was comfortable and at which I could tolerate a really bad outcome.

Within each range, I used a fractional Kelly system, betting a percentage of my bankroll proportional to my expectation in favorable situations, up to the top of my current betting range, and the minimum of my current range otherwise. At $50–500 (full scale) I bet full Kelly, which generally turned out to be a percentage of my bankroll a little less than the percentage expectation in favorable situations.

Convertible bonds

My next illustration is the evolution over more than two decades of a model for convertible bonds. Simplistically, a convertible bond pays a coupon like a regular bond but also may be exchanged at the option of the holder for a specified number of shares of the "underlying" common stock. It began with joint work during 1965 and 1966 with economist Sheen Kassouf on developing models for common stock purchase warrants. Using these models, we then treated convertible bonds as having two parts, the first being an ordinary bond with all terms identical except for the conversion privilege. We called the implied market value of this ordinary bond the "investment value of the convertible". Then the value of the conversion privilege was represented by the theoretical value of the attached "latent" warrants, whose exercise price was the expected investment value of the bond at the (future) time of conversion.

Just as with warrants, we standardized the convertible bond diagrams so that the prices of different bonds versus their stock could be compared cross-sectionally on a single diagram. This was a crude first pass at finding underpriced issues. We also plotted the price history of

individual bonds versus their stock on the standardized diagrams. This was a crude first pass at finding an issue that was cheap relative to its history. You can see diagrams like these in *Beat the Market*. The scatter diagrams for individual convertibles were also useful in choosing delta-neutral hedge ratios of a stock versus its convertible bond.

I'd guessed the Black–Scholes formula in late 1967 and used it to trade warrants from late 1967 on. My progress in valuing convertibles quickened after I began to value the latent warrants using the model. Volatility, stock price and the appropriate riskless rate of return were now incorporated. But there remained the problem of estimating the future volatility for warrants (or options). There also was the problem of determining investment value. The future expected investment value of the convertible could only be estimated. Worse, it was not constant but also depended on the stock price. As the stock price fell, the credit rating of the bond tended to fall, reducing the bond's investment value, and this fact needed to be incorporated. We started with *ad hoc* corrections and then developed a more complete analytic model. By the time we completed this analytic work in the early 1980s, we had real-time price feeds to our traders' screens, along with real-time updated plots, calculations of alpha, hedging ratios, error bounds, and so forth. When our traders got an offer they could in most cases respond at once, giving us an edge in being shown merchandise.

To exploit these ideas and others, Jay Regan and I started the first market-neutral derivatives hedge fund, Princeton Newport Partners, in 1969. Convertible hedging was a core profit center for us. I had estimated early on that convertible hedging was a $100 million idea for us. In the event, Princeton Newport Partners made some $250 million for its partners, with half or more of this from convertibles. Convertible hedgers collectively have made tens of billions of dollars over the last three and a half decades.

Along the way we met a lot of interesting people: Names from the movie industry like Robert Evans, Paul Newman, and George C. Scott, Wall Streeters like Robert Rubin, Mike and Lowell Milken, and Warren Buffett, and academics like Nobelists Bill Sharpe, Myron Scholes, and Clive Granger.

Here are some of the things we learned about building successful quantitative models in finance. Unlike blackjack and gambling games, you only have one history from which to use data (call this the Heraclitus principle: you can never invest in the same market twice). This leads to estimates rather than precise conclusions. Like gambling games, the magnitude of your bets should increase with expectation and decrease with risk. Further, one needs reserves to protect against extreme moves. For the long-term compounder, the Kelly criterion handles the problem of allocating capital to favorable situations. It shows that consistent overbetting eventually leads to ruin. Such overbetting may have contributed to the misfortunes of Victor Niederhoffer and of LTCM.

Our notions of risk management expanded from individual warrant and convertible hedges to, by 1973, our entire portfolio. There were two principal aspects: local risk versus global risk (or micro versus macro; or diffusion versus jump). Local risk dealt with "normal" fluctuations in prices, whereas global risk meant sudden large or even catastrophic jumps in prices. To manage local risk, in 1973–1974 we studied the terms in the power series expansion of the Black–Scholes option formula, such as delta, gamma (which we called curvature) and others that would be named for us later by the financial community, such as theta, vega and rho. We also incorporated information about the yield "surface", a plot of yield versus maturity and credit rating. We used this to hedge our risk from fluctuations in yield versus duration and credit rating.

Controlling global risk is a quite different problem. We asked how the value of our portfolio would change given changes of specified percentages in variables like the market index, various shifts in the yield surface, and volatility levels. In particular we asked extreme questions: what if a terrorist explodes a nuclear bomb in New York harbor? Our prime broker, Goldman Sachs, assured us that duplicates of our records were safe in Iron Mountain. What if a gigantic earthquake hit California or Japan? What if T-bills went from 7% to 15%? (they hit 14% a couple of years later, in 1981). What if the market dropped 25% in a day, twice the worst day ever? (it dropped 23% in a day 10 years later, in October 1987. We broke even on the day and were up slightly for the month). Our rule was to limit global risk to acceptable levels, while managing local risk so as to remain close to market neutral.

Two fallacies of which we were well aware were that previous historical limits on financial variables should not be expected to necessarily hold in the future, and that the mathematically convenient lognormal model for stock prices substantially underestimates the probabilities of extreme moves (for this last, see my columns in *Wilmott*, March and May 2003). Both fallacies reportedly contributed to the downfall of LTCM.

Two questions about risk which I try to answer when considering or reviewing any investment are: "What are the factor exposures", and "What are the risks from extreme events?".

Statistical arbitrage

Hedging with derivatives involves analytical modeling and, typically, positions whose securities will have known relationships at a future date. A rather different modeling approach is involved with a product called "statistical arbitrage".

The key fact is the discovery of an empirical tendency for common stocks to have short-term price reversal. This was discovered in December 1979 or January 1980 in our shop as part of a newly initiated search for "indicators", technical or fundamental variables which seemed to affect the returns on common stocks. Sorting stocks from "most up" to "most down" by short-term returns into deciles led to 20% annualized returns before commissions and market impact costs on a portfolio that went long the "most down" stocks and short the "most up" stocks. The annualized standard deviation was about 20% as well. We had found another potential investment product. But we postponed implementing it in favor of expanding our derivatives hedging.

For the record, we had also been looking at a related idea, now called "pairs trading". The idea was to find a pair of "related" stocks that show a statistical and perhaps casually induced relationship, typically a strong positive correlation, expecting deviations from the historical relationship to be corrected.

Meanwhile at Morgan Stanley, the brilliant idiosyncratic Gerry Bamberger (the "unknown creator" of MS mythology) discovered an improved version in 1982: hedge within industry groups according to a special algorithm. This ran profitably from 1982 or 1983 on. However, feeling underappreciated, Bamberger left Morgan in 1985, co-ventured with us, and retired rich in 1987. As a result of his retirement, we happened to suspend our statistical arbitrage operation just before the 1987 crash. We restarted a few months later with our newly developed factor-neutral global version of the model.

Too bad: simulations showed that the crash and the few months thereafter were by far the best period ever for statistical arbitrage. It was so good that in future tests and simulations we had to delete this period as an unrealistically favorable outlier.

Statistical arbitrage ran successfully until the termination of Princeton Newport Partners at the end of 1988.

In August 1992 we launched a new, simpler principal components version. This evolved into the "omnivore" program, which incorporated additional predictors as they were discovered. The results: in 10 years, from August 1992 through October 2002, we compounded at 26% per annum net before our performance fee, 20% net to investors, and made a total of about $350 million in profit. Some statistics: 10 day average turnover; typically about 200 long and 200 short positions; 10,000 separate bets per year, 100,000 separate bets in 10 years. The gross expectation per bet at about $(2/3)\% \times 1.5$ leverage $\times$ 2 sides $\times$ 25 turnovers per year is about 50% per year. Commissions and market impact costs reduced this to about 26%.

Concluding remarks

Where do the ideas come from? Mine come from sitting and thinking, academic journals, general and financial reading, networking, and discussions with other people.

In each of our three examples, the market was inefficient, and the inefficiency or mispricing tended to diminish somewhat, but gradually over many years. Competition tends to drive down returns, so continuous research and development is advisable. In the words of Leroy Satchel Paige, "Don't look back. Something might be gaining on you".

4

Psychology in Financial Markets

Henriëtte Prast

Quantitative Finance Review 2003

During the 1980s financial economists, confronted with phenomena in financial markets that were difficult to explain within the rational expectations and expected utility framework, started to consider the possibility that some market participants behave less than rationally, and to study whether this might affect markets as a whole.

Initially, they made no explicit use of insights from psychology. Although the literature by psychologist Daniel Kahneman and his co-author Amos Tversky on prospect theory had already been published in 1979, financial economists were not aware either that this literature existed or that it might be relevant for finance. They introduced information asymmetries and shifts in preferences to explain the apparent anomalies, or simply assumed that people do not always behave rationally.

At a later stage, economists became aware that prospect theory and the psychological literature on heuristics and biases in judging information may provide a sophisticated model of why people make decisions for what seem to be non-rational reasons. Perhaps the 1987 crash provided an additional impulse to question the validity of the rational expectations framework. Anyway, during the 1990s, the finance literature that uses psychological concepts to explain the behaviour of market participants became a separate field of research. And it invented its own label: "behavioural finance". The 2002 economics Nobel prize awarded to Kahneman was a further recognition of the contribution of psychology to the explanation of economic behaviour.

This study surveys the behavioural finance literature. It is set up as follows. In Section 1, some financial puzzles, or anomalies, are briefly sketched. Section 2 introduces prospect theory. This is a theory of decision making under risk which takes actual decision-making processes by people into account, rather than postulating rationality. Prospect theory is to be seen as an alternative to expected utility theory. Section 3 introduces the heuristics and biases used by people when judging information. Again, these heuristics and biases are found in actual behaviour, and are to be seen as an alternative to the rational expectations hypothesis. Section 4

Contact address: De Nederlandsche Bank, Supervision Division, PO Box 98, 1000 AB Amsterdam, The Netherlands. E-mail: h.m.prast@dnb.nl Telephone: (31)20 5243738

The views in this chapter are those of the author and not necessarily those of the Nederlandsche Bank.

describes how behavioural finance may help solving the six puzzles mentioned in Section 1. In Section 5, the implications of behavioural finance for market (in)efficiency are discussed. Section 6 summarizes and concludes.

1 Six puzzles of finance

Puzzle 1: Asset price over- and underreaction

Various empirical studies conclude that asset prices and exchange rates tend to under- and overreact to news. Cutler, Poterba and Simmons (1991) study various financial markets in the period 1960–1988. They find autocorrelation of returns over a horizon varying from 4 months to 1 year. Bernard (1992) studies the returns on individual stocks in the periods following earnings announcements, measuring the surprise element in earnings and its effect on stock prices. His conclusion is, that the more surprising an earnings announcement is, the more a stock price will rise in the periods following the initial news release. Jegadeesh and Titman (1993) and De Bondt and Thaler (1985) find results that point to inefficient pricing in financial markets. Jegadeesh and Titman's research suggests a pattern of underreaction: over a given period (in the study under consideration, 6 months), the return on winning stocks exceeds that on losing stocks. De Bondt and Thaler show that in the longer run, the opposite holds.

Puzzle 2: Excessive trading and the gender puzzle

Barber and Odean (2000) study trading patterns and returns of over 66,000 accounts held by private investors with stockbrokers[1] in the period 1991–1996. The average investor in their sample would have realized a higher return if he had traded less. Moreover, the difference in net return between the 20% investors that traded the least and the 20% that traded the most was about 7% percentage point. The average net return of the group fell short of that of Standard&Poor's 500 by 1.5 percentage points. On the basis of this empirical evidence, Barber and Odean conclude that the average individual (amateur) investor trades excessively. Barber and Odean (2001) study the difference in investment behaviour between men and women by analysing the behaviour of more than 35,000 investors over a 6 year period, distinguishing between investment accounts opened by women and by men. They study the frequency of transactions and the return on the individual accounts. Their study reveals that, on average, men trade 1.5 times more frequently than women, and earn a return that is one percentage point lower. The gender gap is even larger for singles. Single men trade 67% more often than single women, and earn a return that is 1.5 percentage points lower.

Puzzle 3: Hypes and panic

Kaminsky and Schmukler (1999) investigate investors' response to news in 1997–1998, at the time of the Asian crisis. They conclude that the 20 largest daily price changes cannot be fully accounted for by economic and political news. Kaminsky and Schmukler also find that prices overreact more strongly as a crisis worsens, and that in such periods prices respond more strongly to bad news than to good news. In a similar analysis, Keijer and Prast (2001) analyse the response to news of investors in ICT companies quoted on the Amsterdam Stock Exchange in the period 1 October 1999–1 March 2000, in the heydays of the ICT bubble. Classifying

daily telecom news as good or bad, they study the difference in price development between the Amsterdam technology index (MIT index) and the general AEX index. They find that this difference turns out to respond significantly more strongly to good news than to bad news.

Puzzle 4: The equity premium puzzle

Mehra and Prescott (1985) find that between 1926 and 1985, the premium between risky and risk-free assets was on average about 6% per year. In order to be able to explain this equity premium within a rational framework, an unrealistically high degree of risk aversion had to be assumed. Mehra and Prescott show that, in a model where individuals aim at smoothing consumption, the coefficient of relative risk aversion would need to exceed 30 to account for the equity premium. This is a puzzle, since both from a theoretical point of view and on the basis of earlier estimations, this coefficient should be approximately 1.

Puzzle 5: The winner/loser puzzle

Investors sell winners more frequently than losers. Odean (2000) studies 163,000 individual accounts at a brokerage firm. For each trading day during a period of one year, Odean counts the fraction of winning stocks that were sold, and compares it to the fraction of losing stocks that were sold. He finds that from January through November, investors sold their winning stock 1.7 times more frequently than their losing stocks. In other words, winners had a 70 per cent higher chance of being sold. This is an anomaly, especially as for tax reasons it is for most investors more attractive to sell losers.

Puzzle 6: The dividend puzzle

Investors have a preference for cash dividends (Long, 1978; Loomis, 1968; Miller and Scholes, 1982). This is an anomaly as, in the absence of taxes, dividends and capital gains should be perfect substitutes. Moreover, cash dividends often involve a tax disadvantage. Bhattacharya (1979) argues that dividends have a signalling function. However, signalling does not seem capable of explaining all the evidence, hence many consider this to be a puzzle (Brealey and Myers, 1981).

2 Prospect theory

In 1979, Kahneman and Tversky launched their prospect theory in what, in retrospect, proved a seminal paper. On the basis of experiments conducted among colleagues and students, they concluded that the theory of expected utility maximization does not hold in practice. Expected utility theory assumes that the individual maximizes his expected return on the basis of the weighted sum of the various possible outcomes, with each weight being equal to the probability that the corresponding outcome will be realized. Furthermore, the theory assumes that the utility of a final state only depends on the final state; how this final state was reached is irrelevant. Finally, the theory usually assumes that the individual is risk-averse. These assumptions imply that:

$$U(x_1, p_1; \ldots\ldots; x_n, p_n) = p_1 u(x_1) + \cdots + p_n u(x_n) \tag{1}$$

where U is the overall utility of a prospect, $(x_1, p_1; \ldots\ldots; x_n, p_n)$ is a prospect (or gamble), which is defined as a contract that results in outcome x_i with probability p_i and where $p_1 + p_2 + \cdots + p_n = 1$.

$$(x_1, p_1; \ldots\ldots; x_n, p_n) \text{ is acceptable at asset position } w \text{ if } U(w + x_1, p_1; \ldots\ldots; w + x_n, p_n) > u(w) \tag{2}$$

$$u'' < 0 \tag{3}$$

Condition (2) implies that, according to expected utility, a prospect is acceptable to an individual if the utility resulting from integrating the prospect with the individual's assets exceeds the utility of those assets, $u(w)$. Condition (3), the concavity of the utility function, is not necessary for expected utility theory, but it is generally assumed to describe the preferences of a representative individual and implies that the typical individual is risk-averse (Kahneman and Tversky, 1979).

In the experiment set up by Kahneman and Tversky, subjects were asked to solve a range of choice problems. It turned out that in their choices they consistently deviated from expected utility maximization. For example, they evaluate losses and gains in an asymmetric manner. In situations of winning they were risk-averse, while in situations of losing they were risk-seeking. The experiments also showed that people are more sensitive to losses than to gains.[2] In fact, losses have a psychological impact that is about twice as large as the impact of gains. Moreover, further experiments show that people's risk attitude has more dimensions. Thus, a person's risk attitude depends on his recent history. After experiencing a financial loss, people become less willing to take risks. After a series of gains, risk aversion decreases.

A simple value function according to prospect theory can be described by:

$$v(x) = x^a \text{ for } x \geq 0; \; v(x) = -\lambda(-x)^b \text{ for } x < 0 \tag{4}$$

where v is the psychological value that the individual attaches to situation x. From experimental research it appears that the value of λ is approximately 2.25 and that a and b both equal 0.88 (Kahneman and Tversky, 1992).

Figure 1 gives a graphical presentation of a value function according to prospect theory.

Another important piece of prospect theory is the finding that people's decision weights do not correspond to objective probabilities. According to prospect theory, a decision process consists of two stages. The first is the editing stage. In this stage, people frame prospects in terms of losses and gains relative to a benchmark. In doing so, they apply rules of thumb, or heuristics, that facilitate the interpretation of the various possibilities among which they have to choose. The second stage of the decision process is the evaluation stage. After the various prospects have been edited and framed as losses and gains, they are evaluated and the prospect with the highest value is chosen. The rules of thumb used when editing and evaluating are necessarily a simplification. For example, probabilities or outcomes are rounded, and extremely unlikely outcomes tend to be discarded. As a result, decision weights are a non-linear function of probabilities. Thus, for small p, $p(p) > p$, where p is the probability of an outcome and $p(p)$ is the decision weight. Thus, after the individual has passed the two stages of editing and evaluation, he chooses the prospect that maximizes:

$$\sum \pi(p_i) v(x_i). \tag{5}$$

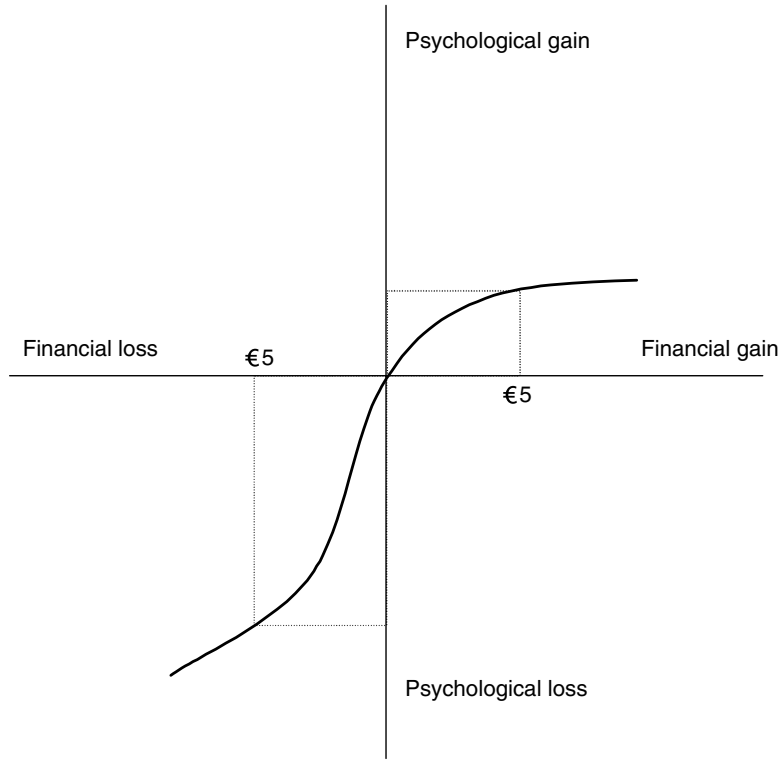


Figure 1: Prospect theory. The psychological value of gains and losses. Based on Kahneman and Tversky (1979)

Prospect theory shows that people use *mental accounting* when making financial decisions. Mental accounting is the tendency to classify different financial decision problems under separate mental accounts, while ignoring that it would be rational to integrate these choices into one portfolio decision. Prospect theory decision rules are then applied to each account separately, ignoring possible interaction. Mental accounting explains why people buy a lottery ticket, while at the same time taking out insurance, or, in other words, why people seek and hedge risk (Friedman and Savage, 1948). Investors mentally keep separate accounts, one for each investment, or one for covering downward risks – for which they use such instruments as bonds – and one for benefiting from the upward potential, for which they use stocks. Although portfolio theory predicts that it would be optimal to integrate these elements mentally, in practice people behave differently. One reason for this behaviour may be that the investor wishes to exert *self-control*. If he keeps separate accounts for different sorts of expenditure, he may be less easily tempted to use his nest egg for an impulse purchase (Thaler and Shefrin, 1981). When a new stock is purchased, a new mental account is opened (Thaler 1980; also, see Shefrin and Statman, 1985).[3]

Mental accounting, combined with loss aversion and a multidimensional risk attitude, results in the *framing effect*. This is the phenomenon that decisions under risk are influenced by the way the decision problem is framed. If a decision is framed in terms of losses, people tend

to choose a risky outcome, whereas they tend to avoid risk when the problem is presented in terms of winning. A frequently cited example to illustrate the framing effect is the following:

> Imagine that you are an army official in a war, commanding 600 soldiers. You have to choose between route A, where 200 soldiers will be saved, or route B, where there is a one-thirds chance that all soldiers will be saved and a two-thirds chance that none will be saved. Which route do you take?

Most people tend to choose route A when the decision problem is framed in this way. However, the decision problem can also be framed as follows:

> You have to choose between route A, where 400 soldiers will die, or route B, where there is a one-third chance that no soldiers will die and a two-thirds chance that all will die.

When the decision problem is framed in this way, most people choose route B, although the objective characteristics are no different from the first problem (Belsky and Gilovich, 1999).

Another result of loss aversion and mental accounting is that in evaluating outcomes people tend to attach value to both changes and final states, rather than to final states only. An example, taken from Antonides (1999), may illustrate this. Students were asked to judge who was happier, Mr A or Mr B. Mr A bought a New York State lottery ticket and won \$100, but he damaged the rug in his apartment and had to pay his landlord \$80. Mr B bought a lottery ticket and won \$20. About 70% of the students believed that Mr B was happier, although their final states – a gain of \$20 – are identical. This evaluation is the result of the fact that the payment, or loss, of \$80 has a stronger psychological impact.

From the value function, the following mental rules can be derived for the combined value of outcomes or events. Examples are based on a situation with two outcomes, x and y (Antonides, 1999).

- *Both outcomes are positive*. In this case (concavity of value function in region of gains), $v(x) + v(y) > v(x + y)$: segregation, that is experiencing these two events separately, is preferred. Moral: do not wrap all Christmas presents together.
- *Both outcomes are negative*. In this case (convexity of value function in region of losses), $v(-x) + v(-y) < v(-x - y)$, so integration of losses is preferred. Example: the psychological cost of suffering two losses on the same day, of say £100 and £50, exceeds the psychological cost of suffering one loss of £150.
- *Mixed outcomes, net result is positive*. This is the outcome $(x, -y)$ with $x > y$, and so $v(x) + v(-y) < v(x - y)$. Hence in this case, integration is preferred. An example: withdrawal of income taxes from payments is less difficult to accept than having to pay taxes separately next year.
- *Mixed outcomes, net result is negative*. In this event $(x, -y)$ with $x < y$, integration is preferred if the positive event x is a little bit smaller than y, whereas segregation is preferred if $x \ll y$. This preference for segregation is called the "silver lining" effect and is deliberately and frequently used by marketeers of financial products.

Figure 2 illustrates the mixed outcomes and silver lining effect. The essential characteristics of prospect theory apply to both decision problems with a financial character and to non-monetary

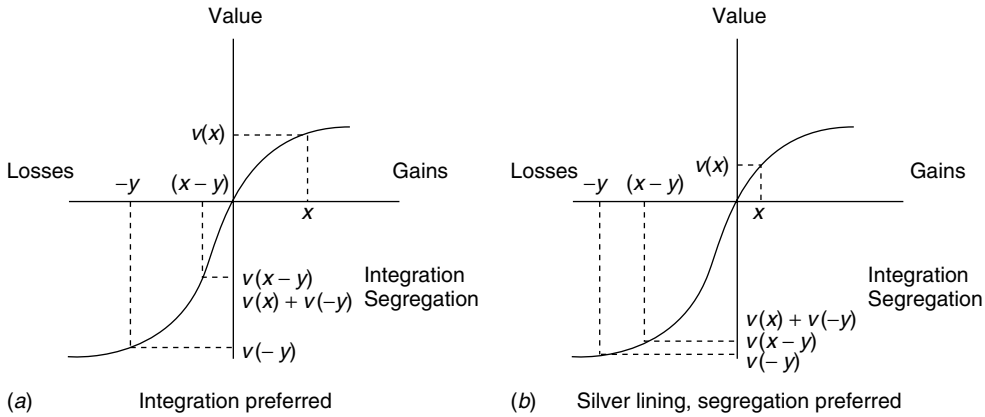


Figure 2: Mental accounting: the silver lining effect. Based on Antonides (1999), p. 257

choices. The Appendix presents some of the decision problems used by Kahneman and Tversky in the experiments used for their 1979 paper on prospect theory.

3 Heuristics and biases in the use of information

Prospect theory deals with the evaluation of financial and non-monetary outcomes, or preferences, and is the first pillar of behavioural finance. The second pillar of behavioural finance concentrates on beliefs, or the way in which people use information. Cognitive psychology has found that people use heuristics and are biased in forming beliefs and in processing information. As a result of these heuristics and biases, information is not used in an objective manner. This section introduces a number of heuristics and biases that behavioural finance uses to account for irrational behaviour in financial markets. They are: cognitive dissonance, conservatism, overconfidence, biased self-attribution, availability heuristic, and representativeness heuristic.

Cognitive dissonance

Cognitive dissonance is the phenomenon of two cognitive elements – an opinion, new information – conflicting with each other (Festinger, 1957). People want to reduce cognitive dissonance in order to avoid the psychological pain of a poor self-image. Therefore, they tend to ignore, reject or minimize information that suggests that they have made a wrong decision or hold on to an incorrect belief. The result is that people filter information in a biased manner. Filtering information is easier when the individual is part of a group whose members hold similar opinions or have taken similar decisions.[4] Therefore, herding may facilitate the reduction of cognitive dissonance and reinforce biased information filtering. The theory of cognitive dissonance may explain not only hypes, but also panic in financial markets, for it predicts that if much dissonant information is released, it becomes more difficult to ignore it. At a certain point the dissonance is equal to the resistance to revise the existing opinion, and the individual will switch to actively seeking information that confirms that his earlier decision was wrong. If he was part of a group, he will now break away from it. The group becomes smaller, and this increases the dissonance of the remaining group members. This may lead to a sudden change of direction of the herd.[5]

Conservatism

Conservatism is defined as the phenomenon that people only gradually adjust their beliefs to new information (Edwards, 1968). It therefore resembles the mechanism that plays a role in the theory of cognitive dissonance. Experimental research indicates that it takes two to five observations to bring about a change of information or opinion, whereas in the case of Bayesian learning one observation would have sufficed. The more useful the new information, the stronger is the conservatism. This is because new information that is at variance with existing knowledge is harder to accept.

Overconfidence

Empirical research in cognitive psychology concludes that the average individual is overconfident. Overconfidence implies that an individual overestimates his ability. The degree of overconfidence varies among professions. It is strongest in professions that can easily shift the blame for mistakes on others or unforeseen circumstances (Odean, 1998b). An economist or financial market professional who in retrospect has failed to predict economic growth incorrectly may put this down to all sorts of unforeseeable political and economic events, or perhaps even to irrational behaviour of investors and consumers. On the other hand, in professions where no-one else is to blame, overconfidence is limited. Thus, a mathematician who cannot prove a theorem has no one to blame but himself. There are also gender differences in overconfidence. Men have been found to be, on average, more overconfident than women (Barber and Odean, 2001).

Self-serving bias and biased self-attribution

The individual is inclined to interpret information in a way that is most favourable to himself, even when he tries to be objective and impartial. People tend to discount the facts that contradict the conclusions they want to reach and embrace the facts that support their own viewpoints (Babcock and Loewenstein, 1997). This mechanism is called the self-serving bias. Also, people tend to blame failures on others and attribute successes to their own ability. This phenomenon is referred to as biased self-attribution (Zuckerman, 1979). The self-serving bias and biased self-attribution contribute to the dynamics of overconfidence. The asymmetry in dealing with successes and failures makes sure that people do not learn enough from their mistakes. In fact, biased self-attribution increases overconfidence.

Availability heuristic

The availability heuristic is the tendency of people to estimate the frequency or probability of an event by the ease with which it can be brought to mind (Herring, 1999). The car driver who witnesses an accident immediately starts to drive more cautiously, even though he knows that the probability of a car accident has not increased. It could be argued that seeing the accident has contributed to his insight into the hazards of driving and that his decision to drive more carefully is due to learning, and therefore consistent with rationality. But in practice, in the course of time the driving style becomes more reckless again. In other words, the cautious driving style is not the result of learning, but of a temporary increase in the subjective probability of car accidents brought about by having recently witnessed one.

Representativeness heuristic

The representativeness heuristic is defined as the phenomenon that people look for a pattern in a series of random events (Tversky and Kahneman, 1974). The representativeness heuristic

leads to stereotyping and serves to make the world look more organized than it really is. It may cause people to draw far-reaching conclusions on the basis of merely a few indications. The representativeness heuristic is often illustrated by the 'Great Bear' effect. People watching a starry sky are usually firmly resolved to detect a familiar pattern. The mechanism is also known as the law of small numbers. People tend to generalize and draw conclusions on the basis of too little statistical information.

4 Application to financial markets

4.1 Introduction

Behavioural finance aims at explaining these puzzles using elements from prospect theory to explain investor preferences, and assuming that investors use heuristics, or rules of thumb, when judging information and forming beliefs. Before turning to a behavioural finance explanation of the six puzzles of finance (Section 5), the next sections will introduce prospect theory (Section 2) and heuristics and biases in the judgement of information (Section 3).

This section shows how behavioural finance may explain the six financial puzzles introduced in Section 1: over- and underreaction, excessive trading and the gender puzzle, the equity premium puzzle, the winner/loser puzzle and the dividend puzzle. Table 1 presents an overview of the puzzles and the behavioural concepts used to explain them. Puzzles 1, 2 and 3 are explained by heuristics and biases in the judgement of information and the formation of beliefs. Puzzles 4, 5 and 6 are explained with the help of prospect theory.

TABLE 1: FINANCE PUZZLES AND THEIR BEHAVIOURAL SOLUTIONS

	Puzzle	Solution
1	Over- and underreaction	Conservatism; representativeness heuristic
2	Excessive trading and the gender puzzle	Overconfidence
3	Hypes and panic	Cognitive dissonance theory
4	Equity premium puzzle	Mental accounting and loss aversion
5	Winner/loser puzzle	Mental accounting and loss aversion
6	Dividend puzzle	Mental accounting, loss aversion and self-control

4.2 Over- and underreaction of stock prices

An underreaction of stock prices occurs if the stock market reacts to news not only in the period immediately after the news is released, but also in subsequent periods. Overreaction occurs in the opposite case: the news is immediately followed by a stock price reaction, which in the subsequent periods is partially compensated by one or more changes in the opposite direction.

Various behavioural finance models seek to explain these patterns of under- and over-reactions. Barberis, Shleifer and Vishny (1998) use the concepts of conservatism and the representativeness heuristic; Daniel, Hirschleifer and Subrahmanyam (1998) concentrate on biased self-attribution and overconfidence.

Barberis, Shleifer and Vishny (1998) define underreaction as a situation in which the return in the period following the publication of good news (and after the very first reaction of stock

prices) is on average higher than it would have been had the news been bad. In an efficient market, the news would be fully processed in the period following immediately upon the news release. Hence, in subsequent periods, the development of stock prices would be independent of the news released in the initial period. If, after a favourable news fact, prices continue to rise, there must have been an underreaction in the period immediately following the news. Indeed, if the reaction had been adequate, the rise would have been realized straight away. An overreaction occurs if the price reacts too strongly. In that case, the stock price increase (decrease) will be followed by decreases (increases).

Barberis, Shleifer and Vishny account for the pattern of under- and overreactions by combining conservatism and the representativeness heuristic. They develop a model involving one investor and one asset. All profit is paid out as dividend. The equilibrium price of the asset equals the net present value of expected returns. Stock prices depend on news, because investors use news to update their expectations about future earnings. However, conservatism causes news to be insufficiently reflected in prices in the short term. The average investor learns more slowly than would be optimal and prices take longer to reach the new equilibrium than would be the case with rational Bayesian learning. This explains the short-term underreaction. In the longer term, the representativeness heuristic induces the investor to attach too much value to a news fact if it is part of a series of a random series of similar messages, in which the investor mistakenly perceives a pattern.[6] The investor believes that one of two regimes applies, i.e. either profits are 'mean-reverting', with a positive shock being followed by a negative one, or they are characterized by a trend. If the investor has observed a series of good earnings shocks, his belief that profits follow a trend grows. On the other hand, if he has observed a series of switches from positive to negative earnings shocks and vice versa, he may switch to the belief that earnings are mean-reverting. These updates of beliefs are meant to represent the mechanisms of the representativeness heuristic and conservatism. Simulating earnings with a random walk model, Barberis, Shleifer and Vishny show that, depending on the values chosen for the parameters, these basic assumptions may produce a pattern of underreactions, a pattern of overreactions or a pattern of underreactions alternated by overreactions.

Daniel, Hirschleifer and Subrahmanyam (1998) develop a model of investor behaviour that takes account of overconfidence and biased self-attribution. They model these psychological mechanisms by assuming that investors tend to overestimate their amount of private information and their ability to interpret this information. Information is private if it has not (yet) been disclosed publicly. Because of his overconfidence, the investor believes that he is one of the few, if not the only one, to recognize the relevance of signals he receives. He believes he has discovered a hot tip which gives him an information advantage over others, who will not come into action until after the relevant information is public knowledge. If the private information is favourable, the investor will buy, convinced as he is that this information has not become incorporated yet into the prices. Daniel, Hirschleifer and Subrahmanyam show that the investor following this line of reasoning tends to purchase more (if the private information is favourable) than is warranted by the fundamental, which leads to an overreaction of stock prices. Besides overconfidence, biased self-attribution also comes into play in this model, for the investor interprets public information asymmetrically. If new public information corroborates what the investor has already assumed on the basis of his private information, this will increase investor confidence. If it does not, the investor blames others. Therefore, overconfidence will not diminish and is likely to increase.

4.3 Excessive trading and gender puzzle

Odean (1998b) develops a theoretical model which takes account of overconfidence. He models overconfidence by assuming that market participants overestimate their ability to interpret information. Every market participant believes that he is better in picking up and interpreting information, and that therefore the accuracy of the information he receives is above average. Thus, the model predicts that investors trade excessively. They assume that two types of asset are traded, one risk-free, with zero interest rate, and one risky asset. There are N price-taking investors ($N = 8$). Their *a priori* information is the same. All investors receive a signal about the probability distribution of the return on the risky asset. Each investor believes his signal to be more precise than the signals of others, but knows that there are some traders receiving the same signal. So each investor believes he belongs to the group of investors that is above average. Within this framework, overconfidence causes trading volume and stock price fluctuations to increase, and stock price efficiency to decrease. However, Odean (1998b) shows that overconfidence does not always stand in the way of market efficiency. In a market of noise traders – traders who follow the market trend, despite being aware that share and bond prices are inconsistent with fundamental factors – including an insider overestimating himself, transaction volume and price fluctuations will increase, but pricing will be more efficient

Opinion polls suggest that the average amateur investor is, in fact, overconfident. Gallup conducted 15 surveys in the period June 1998–January 2000, each among 1000 investors (Barber and Odean, 2001). One of the questions was, what return the respondents expected to realize on their portfolio in the following year. The surveys also asked the investors' expectations of next year's average stock market return. On average, respondents thought they could beat the market, which is by definition impossible.[7]

As mentioned in Section 1, Barber and Odean (2000) indeed found that, first, the average investor trades too much, and second, that the investors in their survey who traded the least earned a return that was far above the return of the investors that traded the most. In order to investigate whether overconfidence might indeed be the explanation of the excessive trading, Barber and Odean (2001) studied differences in investment behaviour between men and women. Psychological research has shown that, on average, men are more overconfident than women. If it could be shown that female investors trade less frequently than men, while realizing a higher return, this would support the assumption that the excessive trading might be due to overconfidence, as predicted by Odean's theoretical model. Barber and Odean studied the investment behaviour of more than 35,000 investors over a 6-year period, distinguishing between accounts opened by women and by men. They analysed the investment pattern, the frequency of transactions, and the resulting returns. Their dual hypothesis was that men trade more frequently than women, and that they realized a lower return. The results prove them right. On average, men trade 1.5 times more frequently than women, and earn a return that is one percentage point lower. The superior performance by women cannot be ascribed to their being more experienced investors. Half of the women in the survey claimed to be experienced investors, against over 60% of men. Having found evidence for a possible relationship between overconfidence and excessive trading, Barber and Odean went further to study the subset of singles in their survey. It cannot be excluded that an investment account opened by a woman is managed by a man, and vice versa, but this is less likely for singles than for married couples. Therefore, one would expect the gender difference in trading frequency and return to be even larger in the

singles subset. This is indeed what Barber and Odean found. The average male bachelor traded 67 per cent more frequently than his female counterpart, and realized a return that was almost 1.5 percentage points lower.

Barber and Odean considered an alternative explanation for excessive trading. It could be that the average investor considers trading to be a hobby. In that case, the lower return might be interpreted as the price the investor is willing to pay for this leisure activity. And the difference between men and women might be explained by assuming that investing is more of a hobby to men than it is to women. However, Barber and Odean rejected this possibility. They calculated that the most active trader loses 3.9 per cent of his annual household income by trading excessively. This exceeds all expenditures on leisure activities of a typical family with an income similar to that of those in the sample.

4.4 Hypes and panic

Both empirical research by Kaminsky and Schmukler into the reaction of investors to news during the Asian crisis, and empirical research by Keijer and Prast into the reaction of ICT stock prices to news, found that investors seem to filter information in a biased manner, as predicted by the theory of cognitive dissonance. Kaminsky and Schmukler found that prices overreact more strongly as a crisis worsens, and that in such periods prices respond more strongly to bad news than to good news. Keijer and Prast (2001) analysed the response of investors in ICT companies quoted on the Amsterdam Stock Exchange in the period 1 October 1999–1 March 2000 to relevant news. Classifying daily telecom news as good or bad, they studied the difference in price development between the Amsterdam technology index (MIT index) and the general AEX index. This difference turned out to respond significantly more strongly to good news than to bad news. Thus, Keijer and Prast found that the reaction coefficient to good news is more than twice as large, in absolute value, than the reaction coefficient to bad news. These reaction patterns fit in with the theory of cognitive dissonance, which predicts that once people hold a fundamental opinion, they tend to ignore or minimize information that suggests they may be wrong, and tend to pay too much attention to information that confirms their opinion. The results by Kaminsky and Schmukler indicate panic, those by Keijer and Prast are suggestive of a hype.[8]

4.5 Equity premium puzzle

Benartzi and Thaler (1995) showed that the equity premium puzzle is solved if it is assumed that individuals behave in accordance with prospect theory. They modelled the behaviour of investors who have a long planning horizon and whose aim of investing is not to realize speculative profits, but rather to have a high return on a long-term investment. In their model, the investor must choose between a portfolio only consisting of stocks and one containing just bonds. These investors evaluate their portfolios on a regular basis, say a year, not with the aim of changing it but, for example, because they need to state their income to tax officials or to a compliance officer. The evaluation of their portfolio does have a psychological effect. Losses, even if they are not realized, have a larger psychological impact than gains. This implies that a portfolio consisting of risky assets should earn an expected return that compensates for the emotional cost of these "paper losses". Assuming a plausible degree of loss aversion, namely a coefficient of 2.25, Benartzi and Thaler showed that an investor who has a 30-year planning horizon and evaluates his portfolio annually, requires an equity premium of about 6.5 percentage points to be indifferent between stocks and bonds. At evaluation frequencies of

two, five and ten times per year, the equity premiums should be 4.65, 3.0 and 2.0 percentage points, respectively.[9] Benartzi and Thaler note that it is conceivable that the psychological involvement of individual investors regarding the value of their portfolios is stronger than that of institutional investors. As the latter are major players in financial markets, they may be expected to be less hampered by loss aversion. Still, the psychological impact of a regular portfolio evaluation by their clients may be relevant to the position of fund managers and other institutional investors. Loss aversion may also explain why pension funds, whose horizon is basically infinitely long, invest relatively little in stocks.[10] Recent research has shown, though, that the equity premium is declining (Jagannathan, McGrattan and Scherbina, 2001). According to Benartzi and Thaler's model, this would signify that the average evaluation period has grown longer. This phenomenon is not accounted for by the theory.

4.6 Winner/loser asymmetry

Investors are predisposed to hold their losing stocks for too long, and sell their winning stocks too early (Shefrin and Statman, 1985). This is an anomaly, especially as in many countries selling losers offers a tax advantage. Shefrin and Statman (1985) use prospect theory to explain the asymmetry in the sale of losers and winners. Take the case where an investor needs cash. He may choose between selling share A, which gained 20 per cent since he bought it, and share B, which fell 20 per cent since he added it to his portfolio. The investor applies prospect theory rules separately to the accounts of A and B. In doing so, he evaluates the selling prices in terms of gains or losses relative to the price he paid for each stock. Thus, the price paid is the reference point for the investor. Selling share B would imply that the investor would have to close his mental account of share B with a loss. When selling share A, the investor can close the mental account of share A with a profit. Thus, mental accounting and loss aversion make the investor prefer selling winners rather than losers.

The prediction of the model by Shefrin and Statman is confirmed by the results of empirical work by Odean (1998a) and Shefrin (2002). Odean studied 163,000 individual accounts at a brokerage firm. For each trading day during a period of one year, Odean counted the fraction of winning stocks that were sold, and compared it to the fraction of losing stocks that were sold. He found that from January through November, investors sold their winning stock 1.7 times more frequently than their losing stocks. In other words, winners had a 70 per cent higher chance of being sold. In December, these investors sold their losers more quickly, though only by 2 per cent.

4.7 Dividend puzzle

The preference for cash dividends can be explained by mental accounting. Two different explanations can be distinguished. The first explanation focuses on the need for self-control. The investor puts capital gains and cash dividends into separate mental accounts. This is one way of keeping control of spending. The investor worries that, once he decides to finance consumption from spending part of his portfolio, he may spend his savings too quickly. As Shefrin puts it: " 'Don't dip into capital' is akin to 'don't kill the goose that lays the golden eggs' " (Shefrin, 2002, p. 30). When stock prices fall, dividends serve as a 'silver lining'. Statman (1999) formulates its as follows: " 'Not one drop' is a good rule for people whose self-control problems center on alcohol. 'Consume from dividends but don't dip into capital' is a good rule for investors whose self-control problems center on spending." The second explanation concentrates on loss aversion, mental accounting and framing. According to this explanation,

when stock prices fall, dividends serve as a "silver lining". This is the mixed outcome example with $x \ll y$ of Section 2 above. On the other hand, when stock prices rise, the investor likes dividends because they are regarded as a separate gain. This is the "don't wrap all Christmas presents together" example of Section 2. Here dividend and capital are two positive outcomes.

5 Evaluation. Behavioural finance and market (in)efficiency

Fama (1998) states that, while the existence of cognitive-psychological mechanisms may explain why the average individual investor does not behave rationally, this need not imply that markets are inefficient. Even if many market participants behave irrationally, arbitrage by a few rational investors, he and others argue, is a sufficient condition for market efficiency.

Barberis and Thaler (2002) challenge this view, using two mottos to this end, i.e. Keynes's well-known statement "Markets can remain irrational longer than you can remain solvent" and "When the rest of the world is mad, we must imitate them in some measure". Barberis and Thaler give a number of reasons for their proposition that it is unlikely that arbitrage always leads to efficient pricing in financial markets. Their main argument is that there are risks and costs involved in arbitrage. Thus, the irrationality of the participants in financial markets may increase. The rational arbitrageur who buys undervalued stocks will incur a loss if market participants grow even more pessimistic, no matter how right he may be about fundamentals.

Therefore, in view of the arbitrage risk, traders may wish to go along with the market, even if they know that asset prices do not reflect economic fundamentals (Black, 1986; De Long, Shleifer, Summers and Waldmann, 1990). This risk is even more important because of the principal-agent problem in financial markets resulting from, as Shleifer and Vishny (1997) formulate it, a "separation of brains and capital". Professionals do not manage their own money, but that of customers who, on average, do suffer from cognitive-psychological mechanisms. If a professional incurs short-term losses by trading against the irrational market, this may harm his reputation and induce customers to withdraw. The professional who anticipates this response will adjust his behaviour accordingly. The professional who does not, and trades on the basis of fundamentals, will lose customers, be increasingly restricted in arbitraging, and eventually be forced to quit the market. From this perspective, it may be rational for a professional to be myopic.

Barberis and Thaler (2002) mention several reasons why there is no full arbitrage. Contrary to what theory suggests, there are costs involved in arbitrage, such as commission fees. Besides, arbitrage often requires going short. This not only carries additional costs, but also meets with regulatory constraints. Some, often major, market participants, e.g. pension funds, are simply prohibited from taking short positions. Moreover, the identification of price efficiencies is costly. Tracking market inefficiencies in order to conduct arbitrage is only rational if the expected benefits exceed the costs, including those of gathering information (Merton, 1987). One final reason, not given by Barberis and Thaler, for assuming that the market as a whole may be inefficient is the fact that, in practice, well-informed individuals, too, appear to be suffering from a subconscious tendency of biased judgement. Experimental studies of the *self-serving bias* reveals that subjects, even after having been informed of the existence of a bias, thought that not they themselves but others were liable to this bias (Babcock and Loewenstein, 1997). For this reason it seems implausible that market participants are free from any biases in searching and interpreting information.

The assumption that, given fundamentals, prices can be inefficient for a long period is empirically supported by Froot and Dabora (1999). On the basis of the price movements of Royal Dutch/Shell stocks, Froot and Dabora show that prices may deviate from their equilibrium for a long time. In 1907, Royal Dutch and Shell decided to merge. This was realized on a 60:40 basis. Henceforth, Royal Dutch stocks would represent 60 per cent of the two companies' cash flows, and Shell stocks 40 per cent. In other words, the price of a share in Royal Dutch should be 1.5 times that of a share in Shell. Froot and Dabora establish that, irrespective of the fundamental value of the Royal Dutch Shell share, there are structural deviations from the equilibrium price ratio, which may amount to as high as 35%.

Fama also criticizes behavioural finance because the models that make use of cognitive-psychological concepts only account for one anomaly at a time. In his opinion, this is an *ad hoc* approach that sometimes leads to inconsistency between behavioural models. Fama does have a point, in that a new paradigm may be expected to provide a consistent framework. Behavioural finance appears to be on its way to doing just that. Its practitioners systematically employ empirically established psychological mechanisms of human behaviour in addition to, or instead of, the conventional assumption of rationality. Actual decision-making under risk appears to be less simple than would be consistent with the assumptions of expected utility theory and efficient markets theory.

6 Summary and conclusions

Behavioural finance has made two valuable innovative contributions to finance theory and to empirical research. In the first place, it shows that market participants evaluate financial outcomes in accordance with prospect theory, rather than expected utility theory. Many anomalies in preferences result from rules of thumb that are applied when editing prospects to facilitate decision making. Moreover, a greater sensitivity to losses than to gains implies that decisions differ according to how a choice problem is framed. In the second place, behavioural finance uses insights from cognitive psychology to take into account that people, when judging information and forming beliefs, use heuristics and biases that are difficult, if not impossible, to overcome.

As Statman (1999) puts it: "Standard finance people are modelled as 'rational', whereas behavioural finance people are modelled as 'normal' ". Behavioural finance explains financial markets anomalies by taking actual behaviour as a starting point, rather than by postulating rationality both as a norm and as a positive description of actual behaviour. One particularly important question to be answered within this context is, of course, whether irrational behaviour of individual market participants may also lead to inefficiency of the market as a whole. Indeed, it is conceivable that, even if the average investor behaves according to the psychological mechanisms mentioned, the market as a whole will generate efficient outcomes anyway. However, this is not the case, behavioural finance argues, for example because the arbitrage required to compensate for price inefficiencies is costly and risky.

What does this imply for the future of finance? Statman (1999) raises the question of market efficiency in a fundamental manner. According to Statman, it is important that a distinction be made between two definitions of efficient markets. One reads that investors cannot beat the market systematically, the other says that stock prices are always and fully determined by economic fundamentals. Statman makes a plea to agree on two things: (1) that investors cannot systematically beat the market; and (2) that prices may reflect both fundamental and emotional

factors. This would pave the way for a further analysis of financial markets, allowing room for both economic fundamentals and systematic psychological factors.

FOOTNOTES & REFERENCES

1. i.e. "Discount brokers", who, unlike "retail brokers", do not advise their customers on purchases and sales.
2. Loss aversion may explain money illusion. A nominal wage decrease at zero inflation is less easily accepted than exactly the same decrease in real wages in situations of inflation. For the macroeconomic effects of money illusion, see Fehr and Tyran (2001).
3. People at auctions tend to bid a lot higher for the same product if they can pay with a credit card than if they were to pay in cash, even if they do not have to do so instantly. Also, foreign currency is more easily spent than home currency.
4. In religious sects, people continue to believe in events that have been shown to be impossible. The most extreme response, of course, is that of killing the messenger of bad news. See Chancellor (1999).
5. Akerlof and Dickens (1982) give several examples of the economic effects of cognitive dissonance, for example in the labour markets. They do not pay attention to financial markets.
6. As an illustration and motivation of the basic assumptions of their approach, Barberis, Shleifer and Vishny show the results of an experiment in which the subjects were requested to toss a coin. The group was informed in advance that the probability of heads or tails was not 50:50, but 70:30, but not whether the 70% probability applied to heads or tails. While learning too slowly in the beginning, it appears that, after a great many tosses, the subjects too hastily arrive at a conclusion about the probability of heads or tails showing up.
7. Irrational as it may seem, overconfidence may contribute to success. The individual who overestimates his abilities will be prepared to take more risks, will be more motivated and even be able to present himself more effectively (Taylor and Brown, 1988). Perhaps private entrepreneurship only exists because of overconfidence, judging by the finding that every starting entrepreneur thinks that his/her enterprise has a greater probability of succeeding than that of the average starter (Cooper, Woo and Dunkelberg, 1988).
8. It should be noted that this type of analysis is not without pitfalls. In the first place, reporters publish news facts that they have selected at their discretion. Secondly, the researcher cannot always objectify why a given news fact is relevant, or why a news fact should be rated as good or bad, as the case may be. Thirdly, news facts are always assigned the same weight. No distinction is made between news that is rather good, quite good, or just bad.
9. According to Jagannathan, McGrattan and Scherbina (2001), in the United States the premium was about 7 percentage points in the period 1926–1970, and less than 1 percentage point since 1970. Based on stocks from the S&P 500, on the one hand, and long-term bonds on the other hand, the premium was 0.7 percentage point in the 1970s, −0.6 percentage point in the 1980s and 1 percentage point in the 1990s.
10. Pension funds are also subject to regulatory constraints.

■ Akerlof, G. (2002) Behavioural macroeconomics and macroeconomic behaviour. *American Economic Review* 92(3), 411–433.

■ Akerlof, G. and Dickens, W. T. (1982) The economic consequences of cognitive dissonance. *American Economic Review* 72(3), 307–314.
■ Antonides, G. (1996) *Psychology in Economics and Business*. Kluwer Academic, Dordrecht.
■ Babcock, L. and Loewenstein, G. (1997) Explaining bargaining impasse: the role of self-serving biases. *Journal of Economic Perspectives* 11(1), 109–126.
■ Barber, B. and Odean, T. (2000) Trading is hazardous to your wealth: the common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.
■ Barber, B. and Odean, T. (2001) Boys will be boys: gender, overconfidence and common stock investment. *Quarterly Journal of Economics* 116, 261–292.
■ Barberis, N., Shleifer, A. and Vishny, R. (1998) A model of investor sentiment. *Journal of Financial Economics* 49, 307–343.
■ Barberis, N. and Thaler, R. (2002) A survey of behavioral finance. *NBER Working Paper* 9222.
■ Bazerman, M. H., Loewenstein, G. and Moore, D. A. (2002) Why good accountants do bad audits. *Harvard Business Review*, 97–102.
■ Belsky, G. and Gilovich, T. (1999) *Why Smart People Make Big Money Mistakes – and How to Correct Them*. Fireside, New York.
■ Benartzi, S. and Thaler, R. (1995) Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics* 110, 73–92.
■ Bernard, V. (1992) Stock price reactions to earnings announcements. In Thaler R. (ed.), *Advances in Behavioral Finance*. Russel Sage Foundation, New York.
■ Brealey, R. and Myers, S. (1981) *Principles of Corporate Finance*. McGraw-Hill, New York.
■ Chancellor, E. (1999) *Devil Take the Hindmost: A History of Financial Speculation*. Farrar Strauss Giroux, New York.
■ Cooper, A. C., Woo, C. Y. and Dunkelberg, W. C. (1988) Entrepreneurs' perceived probabilities for success. *Journal of Business Venturing* 3, 97–108.
■ Cutler, D. M., Poterba, J. M. and Summers, L. H. (1991) Speculative dynamics. *Review of Economic Studies* 58, 529–546.
■ Daniel, K., Hirschleifer, D. and Subrahmanyam, A. (1998) Investor psychology and security market under- and overreactions. *Journal of Finance* LIII(6), 1839–1885.
■ De Bondt, W. and Thaler, R. (1985) Does the stock market overreact? *Journal of Finance* 40, 793–805.
■ De Long, J. B., Shleifer, A., Summers, L. and Waldmann, R. (1990) Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–738.
■ Edwards, W. (1968), Conservatism in human information processing. In B. Kleinmutz (ed.), *Formal Representation of Human Judgment*. Wiley, New York.
■ Erlich, D., Guttman, P., Schonbach, V. and Mills, J. (1957) Postdecision exposure to relevant information. *Journal of Abnormal and Social Psychology* 54, 98–102.
■ Fama, E. F. (1991) Efficient capital markets II. *Journal of Finance* 46(5), 1575–1617.
■ Fama, E. F. (1998) Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49, 283–306.
■ Fehr, E. and Tyran, J.-R. (2001) Does money illusion matter? *American Economic Review* 91(5), 1239–1262.
■ Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford University Press.
■ Friedman, M. and Savage, L. J. (1948) The utility analysis of choice involving risk. *Journal of Political Economy* 56(4), 279–304.

■ Froot, K. and Dabora, E. (1999) How are stock prices affected by the location of trade? *Journal of Financial Economics* 53, 189–216.
■ Froot, K, and Obstfeld, M. (1991) Intrinsic bubbles: the case of stock prices. *American Economic Review* 81(5), 1189–1214.
■ Hawawimi, G. and Keim, D. B. (1998) The cross section of common stock returns: a review of the evidence and some new findings. Working Paper University of Pennsylvania.
■ Hirschleifer, D., Subrahmanyam, A. and Titman, S. (1994) Security analysis and trading patterns when some investors receive information before others. *Journal of Finance* 49, 1665–1698.
■ Jagannathan, R., McGrattan, E. R. and Scherbina, A. (2001) The declining U.S. equity premium. *National Bureau of Economic Research Working Paper* 8172.
■ Jegadeesh, N. and Titman, S. (1993) Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48, 65–91.
■ Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision making under risk. *Econometrica* 47, 263–291.
■ Kahneman, D. and Tversky, A. (1984) Choices, values and frames. *American Psychologist* 39, 341–350.
■ Kahneman, D., Knetsch, J. and Thaler, R. (1991) Anomalies: the endowment effect, loss aversion and status quo bias. *Journal of Economic Perspectives* 1, 193–206.
■ Kaminsky, G. L. and Schmukler, S. L. (1999) What triggers market jitters? A chronicle of the Asian crisis. *Journal of International Money and Finance* 18(4), 537–560.
■ Keijer, M. and Prast, H. M. (2001) De telecomhype: hij was er echt [The telecomhype: it really existed], *Economisch Statistische Berichten* 86(4302), 288–292.
■ Keynes, J. M. (1936) *The General Theory of Employment, Interest and Money*. Macmillan, London.
■ Loewenstein, G. (2000) Emotions in economic theory and economic behavior. *American Economic Review* 90(2), 426–432.
■ Loewenstein, G. and Prelec, D. (1991) Negative time preference. *American Economic Review* 81, 347–352.
■ Long, J. (1978) The market valuation of cash dividends. *Journal of Financial Economics* 6, 235–264.
■ Loomis, C. (1968) A case for dropping dividends. *Fortune Magazine*, June 15.
■ Mehra, R. and Prescott, E. (1985) The equity premium: a puzzle. *Journal Of Monetary Economics*, 145–161.
■ Merton, R. (1987) A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42, 483–510.
■ Miller, M. H. (1986) Behavioral rationality in finance: the case of dividends. *Journal of Business* 59(4), 451–468.
■ Miller, M. and Scholes, M. (1982) Dividends and taxes: some empirical evidence. *Journal of Political Economy* 90, 1118–1141.
■ Odean, T. (1998a) Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–1798.
■ Odean, T. (1998b) Volume, volatility, price and profit when all traders are above average. *Journal of Finance* 53(6), 1887–1934.
■ Odean, T. (1999) Do investors trade too much? *American Economic Review* 89, 1279–1298.
■ Shefrin, H. (2002) *Beyond Greed and Fear*. Oxford University Press, Oxford.

■ Shefrin, H. and Statman, M. (1984) Explaining investor preference for cash dividends. *Journal of Financial Economics* 13(2), 253–282.
■ Shefrin, H. and Statman, M. (1985) The disposition to sell winners too early and ride losers too long: theory and evidence. *Journal of Finance* XL(3), 777–792.
■ Shiller, R. (1981) Do stock prices move too much to be justified by subsequent changes in dividend? *American Economic Review* 71(3), 421–436.
■ Shleifer, A. and Vishny, R. (1997) The limits of arbitrage. *Journal of Finance* 52, 35–55.
■ Statman, M. (1999) Behavioral finance: past battles and future engagements. *Financial Analysts Journal*, 18–27.
■ Taylor, S. and Brown, J. D. (1988) Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin* 103, 193–210.
■ Thaler, R. (1980) Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* I, 39–60.
■ Thaler, R. and Shefrin, H. (1981) An economic theory of self-control. *Journal of Political Economy* 89(2), 392–410.
■ Tversky, A. and Kahneman, D. (1974) Judgement under uncertainty: heuristics and biases. *Science* 185, 1124–1131.
■ Tversky, A. and Kahneman, D. (1982) Availability: a heuristic for judging frequency and probability. In Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Appendix (based on Kahneman and Tversky, 1979)

Choice problem 1:
You may choose between:

(A) 50% probability of a three-week holiday to England, France and Italy, or
(B) a guaranteed one-week trip to England.

Choice problem 2:
You may choose between:

(C) a 5% probability of a three-week holiday to England, France and Italy, or
(D) a 10% probability of a one-week trip to England.

Whatever an individual's preference, from the viewpoint of maximization of expected profit, choice problems 1 and 2 are equivalent. The individual choosing B in Problem 1 is bound to choose option D in Problem 2. Yet practice proved otherwise. In the first choice problem, 78 per cent of the respondents chose option B, while in the second problem 67 per cent of the same group of respondents chose option C, instead of option D. Apparently, an increase in the probability to win a holiday from 50 to 100 per cent (Problem 1) has a different effect than an increase in the probability of winning a holiday from 5 to 10 per cent (Problem 2). The same

phenomenon was perceived in experiments involving a choice between various probabilities of winning a sum of money. Conclusion: as long as the odds of winning are high, people tend to choose the option that offers the *highest probability of winning*. In situations where winning is possible but not very likely, people tend to take the option offering the *highest profit*. However, as a result of loss aversion, the opposite phenomenon occurs when a choice must be made between two negative prospects. People appear to be eager to avoid situations in which they are bound to lose, as the following example shows.

Choice problem 3:
You may choose between:

(E) an 80% probability of losing €4000
(F) the certainty that you will lose €3000

Choice problem 4:
You may choose between:

(G) a 20% probability of losing €4000 or
(H) a 25% probability of losing €3000

92 per cent of the respondents chose option E of Problem 3, while 58 per cent went for option H in Problem 4. Apparently, the respondents considered the certainty of losing 3000 euro unacceptable, although the expected loss entailed by this option is lower (and less variable, hence involving a lower risk!) than the alternative option, E. As soon as an uncertainty factor is introduced in both loss prospects, such as in Problem 4, the majority (but no more than 58 per cent) opt for minimization of the expected loss.

5

Credit Risk Appraisal: From the Firm Structural Approach to Modern Probabilistic Methodologies

Hugues E. Pirotte Spéder

Quantitative Finance Review 2003

The purpose of the present chapter is to review the evolution of credit risk modelling developments in the recent past and to present some main research lines proposed in 2003. Particular attention is devoted to the increasing need to study the implications of the credit risk modelling framework on the general cash-flow cycle of the firm, represented by the ability of the firm to create value and growth.

A little bit of history . . .

Before the 1990s, a few academicians being left aside, credit risk was a modelling problem that was dropped, thanks to either some assumption or by assuming the perfect availability of guarantees and protections. In its search for other domains of application, the arbitrage pricing way of thinking then started to generate interest in the field . . . but credit risk really emerged as a fashionable theme of research when practitioners and regulators began to put the old Basel Accord under the microscope. Suddenly, mathematicians, statisticians, finance researchers, auditors, consultants, all the financial community, became aware of this new venture. From the few earlier volunteers, we have to deal today with an exponentially growing literature. But what are we really left with?

Contact address: Solvay Business School, University of Brussels, 21 av. F.D. Roosevelt, B-1050 Brussels, Belgium. E-mail: hugues.pirotte@ulb.ac.be

While insurers have always relied on mitigating credit risk through guaranties, financial markets have always been looking for a pricing formula to be able to exchange new products and thus complete the market.

Based on the ability of option pricing models to describe future financial decisions, a big branch emerged, led by Merton in 1973, on the roots of the seminal Black–Scholes model, in which Merton has indeed participated. This stream of research includes Shimko, Tejima, Van Deventer, Longstaff, Schwartz, Anderson, Sundaresan, Leland, Toft, Briys and de Varenne, among others. Some of them, like Leland and Toft, were indeed showing the importance of considering some parameters as being endogenously determined.

Here are some figures extracted from Pirotte (1999), showing the results of Merton (1974), Briys and de Varenne (1997) and Pirotte (1999) using equivalent parameter values. Figure 1 presents the shape of the expected cost of default, depending on the level chosen to stop the going concern and the volatility of the assets of the firm, in a model where default may happen earlier than at maturity and with a two-factor term structure of interest rates.[1] Figure 2 compares the three cases mentioned. The hump-shaped pattern of credit spreads is one of the most interesting, as far as we can show that credit spreads should not necessarily display a monotone pattern.

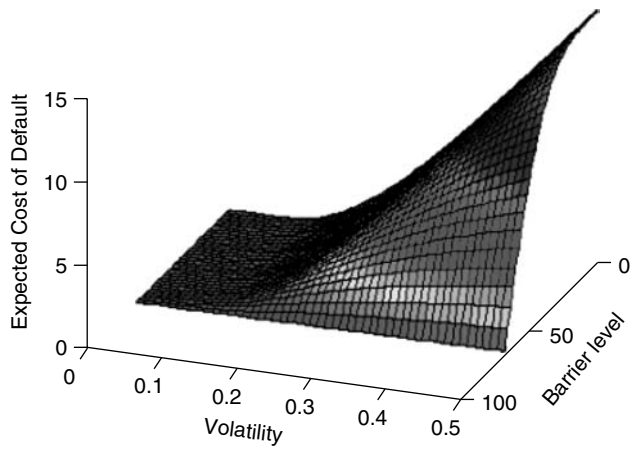


Figure 1: Expected cost of default for various default triggering levels (barrier level) and assets volatility. Reproduced from Pirotte (1999) with permission

But, while this structural approach of credit risk helps in understanding the value to shareholders and debtholders in the firm's capital structure context, it soon became clear that this bottom-up approach was giving poor results in terms of efficient market pricing and was uncalibrable. This opened the path to Jarrow and Turnbull, Duffie and Singleton and Madan and Unal to present probabilistic approaches known as "reduced-form models". Made for calibration, they were typically useful for the pricing of credit derivatives after calibrating on standard debt issues.

But soon, as with market risks, understanding how credit risk was behaving in a portfolio became the target. A big contributor to that enthusiasm was the desire to modernize the old

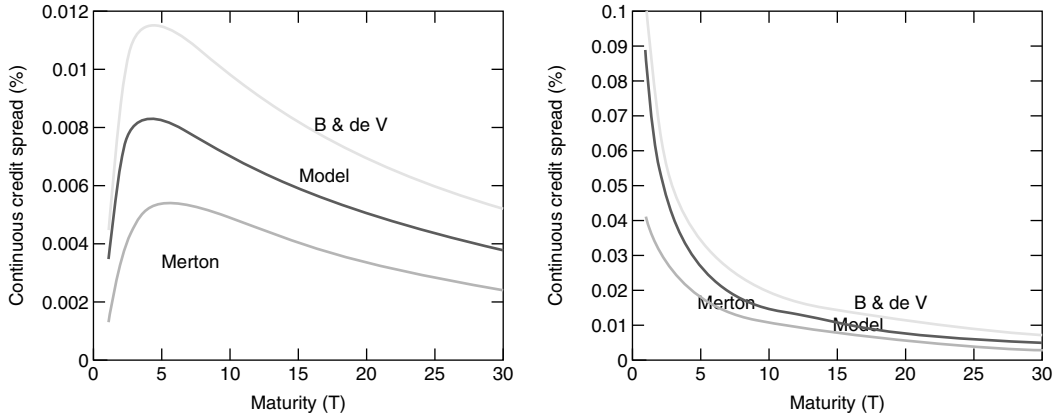


Figure 2: Credit spreads as a function of the maturity for various structural models. "B & deV" stands for the model of Briys & de Varenne (1997). The left-hand side of the figure presents the obtained shapes for a low-leverage firm, while the right-hand figure shows the case for a highly leveraged firm. Given the differences in the modelling of the term-structure of interest rates, parameters are chosen so that we obtain equivalent interest-rate frameworks (constant *r* for Merton) to analyse the models together. Different results will obtain if a much higher volatility of long-term rates is permitted as we currently observe it. Reproduced from Pirotte (1999)

Basel accord. While the Cooke ratio of 8% had been a static notion offering no surprises for almost a decade, the desire to make this accord more sensitive to institutional credit risk management lit the fire again.

Research has literally exploded. Practitioners developed portfolio models, such as CreditMetrics™, CreditRisk + ™, Credit Portfolio View™. Regulators proposed frameworks for the appraisal of the probability of default and the recovery rate. They also proposed formulations for the correlation of assets returns in a credit portfolio along with the capital requirement. KMV restated the honour of structural models, proposing a marked-to-market rating framework that showed real practical results. It was finally bought by Moody's as a signal of the acceptance of a new era in rating methodologies, more desirable than analysis based on simple financial ratios or on "logits on any seemingly important variable". A simple rule obtained, providing consensus: structural models are for explaining, reduced-form models are for pricing. There is still one problem, however: if credit risk models are also to be used for regulatory and control purposes, how can we match the power of calibration to real credit spreads structures and the power of rating determination?.[2] Choices are obviously being made in the recommendations but the authorities do not provide any argument on the cross-effects of these and their impact on the requirements for different activities. They have indeed asked financial institutions to run analyses on their own under the name of "quantitative impact studies".

Is credit risk a new issue? Not really, it was already analysed before in contracts such as convertible bonds. But today, financial markets now have credit derivatives to blow away the credit risk exposure underlying the contract. Thinking in the "option" terminology, a convertible bond is equivalent to a straight bond plus a stock conversion option minus an option to default held by the issuers. Indeed, credit derivatives have boosted the motivation for the need of steady marked-to-market credit risk pricing.

On the other hand, we are still lacking a realistic and steady portfolio model. Even though the Basel II discussions have raised concerns about the way to compute correlations between assets, the subject does not seem to be really uncovered. Basel formulas present a deterministic way to infer asset correlations and nobody knows really what they do mean *in fine*. Optimally, correlations in a default model are linked to the propensity of two assets in reaching default over a certain horizon. Are these correlations stable? Much less than sure. Uncovering credit risk also means relaxing the perfect liquidity hypothesis. Unless we examine credit risk only in a perfectly liquid market, a true issue is to know if "default passages" are concentrated (in time) in our portfolio and how this concentration may bring "negative spirals", where cash ends up to be the only desired collateral (in opposition to what a good collateral may be in normal market conditions).

As a consequence, several issues were raised. First, empirical studies tried to shed some light on the relative importance of specific and systematic factors in the evolution of the counterparties' credit-worthiness. Second, other techniques were proposed, such as the "copula" method. But this seems to make the default occurrence of a second counterpart to be perfectly revealed in time, once first one's default is known. Finally, making statements on the effective diversification in a portfolio is still dangerous, since that effect may vanish if correlations revert to strongly positive in stressed market conditions. There is just one step then to the extreme value theory. In particular, if default is a rare event and if "collapsing markets" is even a rarer event, how can you sustainably switch to abnormal market conditions for your experiment? Since every rare event is driven by particular conditions, how can we account for on an average behaviour of the creditworthiness of market participants? This has all to deal with "survivorship". The Markov property of market-sensitive prices is no longer sustainable when the existence of the position may be questionable for the foreseeable future.

Another field where little research was conducted previously was "sovereign credit risk".

Table 1 lists some main contributions of the recent past, i.e., at the passage to the new century, and the achievements of 2003, respectively. They are far from exhaustive and they just pretend to be an essence of various branches of research. Please forgive me for the many missing articles that would also deserve to be referenced here.

TABLE 1: CREDIT RISK APPRAISAL MODELS AND EVIDENCE

Theme	Article
Recent past	
Earlier modelling	
Corporate finance	Acharya, Huang, Sundaram, Subramanyam (2000) Costly financing, optimal payout policies and the valuation of corporate debt.
Asset-Pricing	Mella-Barral & Tychon (1999) Default risk in asset pricing.
Structural models	Collin-Dufresne and Goldstein (2001) Do credit spreads reflect stationary leverage ratios?
	Sousa (2002) Corporate credit risk valuation using option pricing theory methodology.
	KMV, Crosbie (2002) Modeling default risk.
	Hsu, Saa-Requejo and Santa-Clara (2002) Bond pricing with default risk.
	Neftci (2001) Correlation of default events: some new tools.
	Duffie and Singleton (1999) Modeling term structures of defaultable bonds.

TABLE 1 (*continued*)

Theme	Article
Reduced-form models	Duffie and Lando (2000) Term structures of credit spreads with incomplete accounting information.
Portfolio Models	Schonbucher (2000) Factor models for portfolio credit risk.
	Schonbucher and Schubert (2001) Copula-dependent default risk in intensity models.
Others	Moody's (2002) LossCalc model for predicting LGD.
Derivatives pricing	Huge and Lando (1999) Swap pricing with two-sided default risk in a rating-based model.
Earlier evidence	
Market evidence	Eberhart, Altman and Aggarwal (1998) The equity performance of firms emerging from bankruptcy.
	Chen and Huang (2001) Credit spread bounds and their implications for credit risk modeling.
	Dai and Singleton (2002) Term structure dynamics in theory and reality.
	Newman and Rierson (2002) How downward-sloping are demand curves for credit risk?
	Dietsch and Petey (2002) The credit risk in SME loans portfolio.
	Barnhill, Joutz and Maxwell (2000) Factors affecting the yields on noninvestment-grade bond indices: a cointegration analysis.
	Liu, Longstaff and Mandell (2000) The market price of credit risk – IRS spreads.
Default risk vs. other sources	Bakshi, Madan and Zhang (2001) Investigating the sources of default risk lessons from empirically evaluating credit risk models.
	Collin-Dufresne, Goldstein and Spencer (2001) The determinants of credit spread changes.
	Delianedis and Geske (2001) The components of corporate credit spreads.
	Elton, Gruber, Aggrawal and Mann (2001) Explaining the rate spread on corporate bonds.
	Joutz, Mansi and Maxwell (2001) The dynamics of corporate credit spreads.
	Campbell and Taksler (2002) Equity volatility and corporate bond yields.
	Driessen (2002) Is default event risk priced in corporate bonds?
	Mhuang (2002) How much of corporate-treasury yield spread is due to credit risk?
	Gabbi and Sironi (2002) Which factors affect corporate bonds pricing?
Structural models are not so bad?	Bohn (1999) Characterizing credit spreads.
	Gemmill (2002) Testing Merton's model for credit spreads on zero-coupon bonds.
	Ericsson and Reneby (2002) Estimating structural bond pricing models.
	Eom, Helwege and Huang (2002) Structural models of corporate bond pricing: an empirical analysis.
Model testing in general	Houweling, Hoek and Kleibergen (2001) The joint estimation of term structures and credit spreads.

(*continued overleaf*)

TABLE 1 (*continued*)

Theme	Article
Achievements in 2003	
Evidence	
Market evidence	Hull, Predescu and White (2003) The relationship between credit default swap spreads, bond yields, and credit rating announcements.
	Huang and Kong (2003) Explaining credit spread changes: new evidence from option-adjusted bond indexes.
	Lando and Mortensen (2003) Mispricing of StepUp bonds in the European telecom sector.
	Purda (2003) Controlling for anticipation in stock price reactions to credit downgrades.
	Schuermann (2003) What do we know about loss-given-default?
	Varotto (2003) Credit risk diversification evidence from the Eurobond market.
	Zhang (2003) What did the credit market expect of Argentina default evidence from default swap data?
	Odders-White and Ready (2003) Credit ratings and stock liquidity.
Sources of credit risk	Houweling, Mening and Vorst (2003) How to measure corporate bond liquidity?
	Moody (2003) Systematic and idiosyncratic risk in middle-market default prediction.
	Pesaran, Schuermann, Treutler and Weiner (2003) Macroeconomic dynamics and credit risk.
	Huang and Huang (2003) How much of the corporate-treasury yield spread is due to credit risk?
	Altman, Brady, Resti and Sironi (2003) The link between default and recovery rates.
Structural?	Hull, Nelken and White (2003) Merton's model, credit risk, and volatility skews.
Modelling	
Asset pricing	Giesecke and Goldberg (2003) The market price of credit risk.
	Giesecke (2003) Default and information.
	Yu (2003) Accounting transparency and the term structure of credit spreads.
	Binnenhei (2003) An analytic approach to rating transitions.
	Chen, Filipovic and Poor (2003) A firm-specific model for bond and stock valuation.
	Chen and Filipovic (2003) A simple model for credit migration and spread curves.
Credit risk modelling	Gourrieroux, Monfort and Polimenis (2003) Affine models for credit risk analysis.
	Nikolova (2003) The informational content and accuracy of implied asset volatility as a measure of total firm risk.
Estimation – modelling on parameters	Guha and Sbuelz (2003) Structural RFV recovery form and defaultable debt analysis.
	Albanese and Chen (2003) Implied migration rates from credit barrier models.
	Hahnenstein (2003) Calibrating CreditMetrics correlation concept for non-publicly-traded corporations.

TABLE 1 (*continued*)

Theme	Article
	Hillegeist, Cram, Keating and Lundstedt (2003) Assessing the probability of bankruptcy.
	Schuermann and Jafry (2003) Measurement and estimation of credit migration matrices.
	Yu (2003) Default correlation in reduced-form models.
	Houweling, Mentink and Vorst (2003) Valuing Euro rating-triggered StepUp telecom bonds.
	Bierens, Huang and Kong (2003) An econometric model of credit spreads with rebalancing, ARCH and jump effects.
	Chen (2003) The extended Geske–Johnson model and its consistency with reduced form models.
	Johannes and Sundaresan (2003) Pricing collateralized swaps.
Derivatives pricing	Ayache, Forsyth and Vetzal (2003) Convertible bonds with credit risk.
	Schönbucher (2003) A note on survival measures and the pricing of options on credit default swaps.
	Das, Sundaram and Sundaresan (2003) A simple unified model for pricing derivative securities with equity, interest-rate and default risk.
Monitoring and control	Krishnan, Ritchken and Thomson (2003) Monitoring and controlling bank risk: does risky debt serve any purpose?
	Leippold, Ebnoether and Vanini (2003) Optimal credit limit management.
Capital structure	Huang, Ju and Yang (2003) A model of optimal capital structure with stochastic interest rates.

Expectations for the future . . . some thoughts

Figure 3 presents a simple but already complex picture. Credit risk appraisal should be consistent and compatible across uses (and therefore among agents in the economy): capital adequacy (with the Basel II accord), credit risk sharing (with the existence of credit derivatives), shareholder value creation, firm sustainable growth.

A great step will be made when we will be able to use such structural models to explain further the capital equilibrium of the firm in a marked-to-market perspective. Let's take a concrete example.

We may want first to determine what is the objective: maximizing shareholder value vs. maximizing firm value. Traditional finance based on financial statements give us some hints. On the side of the shareholder value, we can think of the ROE and the value of equity represented by a call on the assets of the firm. ROE depends on the return on invested capital (ROIC, independent of the leverage), impacted negatively by the interest payout after tax and positively by the leverage effect.

On the side of the value of the firm, we can mix ideas from the agency costs theory together with pecking-order considerations, the avoidance of debt overhang problems and the minimization of the WACC of the firm.

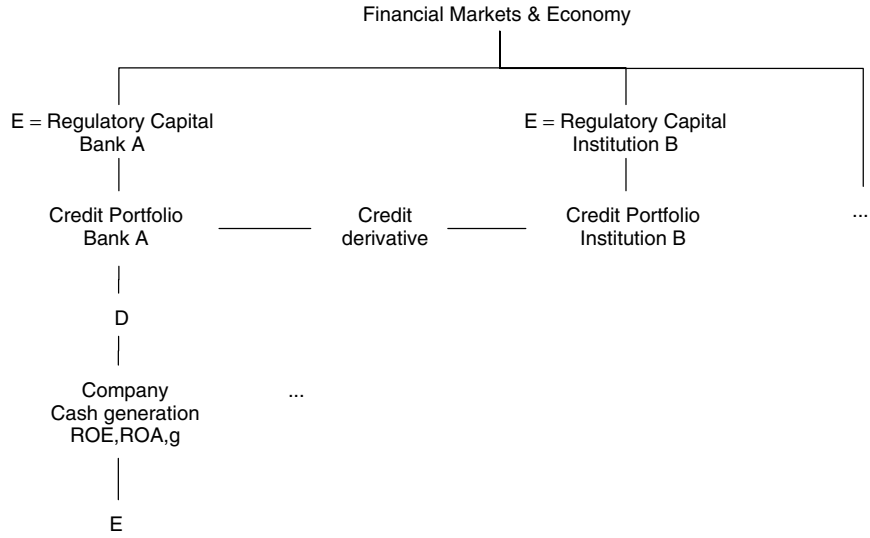


Figure 3: Systemic credit risk and creating shareholder value. D stands for "debt" and E for "equity". Each financial intermediary is required to put in an amount of capital (E), given the riskiness of their operations

Both perspectives will lead to the determination of a first estimation of the optimal capital structure for the firm. This optimal capital structure, together with the payout policy and the profit margin generated by the activity, will allow us to infer the sustainable growth rate for the company.

Up to this point, each step makes the assumption that the firm and its shareholders may choose the structure on their own without the feedback of debtholders (naïve approach). But structural models allow us to easily compute the fair value of debt (and equity) for the various inputs.[3] In order to avoid instantaneous value transfers between shareholders and debtholders at the inception of the debt contract, contractual parameters should make the theoretical value of debt equal to the original investment by debtholders. Either the investment or the leverage may be calibrated, resulting in a given credit spread out of the model.

After iterating and converging towards a set of inferred inputs, we can find out the optimal leverage policy of the firm and the consequent sustainable growth that we can expect. It has to be noted that previous research, such as Leland (1994) and Leland and Toft (1996), has already dealt analytically with the endogenous determination of some parameters, obtaining optimal capital structures and cost of capital for different debt and default designs. But no-one really deals with the impact of credit risk requirements on the evolution of the economics of the firm (value creation and growth). Structural models can go further, thanks to their modelling of the firm, and "old" financial diagnosis can be recovered and combined with it to give some sort of balance scorecard of financing and show the impacts on the firm's activity as a going concern.

Moreover, we stand today on the modernization of the old Basel Accord for a determination of capital requirements more sensitive to changes in creditworthiness. But we are completely unaware of the impact that such modification will have on the supply of debt, and therefore on the equilibrium between small, medium and big companies. Also, in this case, we deal not only with credit spreads at equilibrium but also with the access to debt. What would happen if small and medium enterprises (SMEs) would be forced to be systematically below their optimal

capital structure? The next story would then undoubtedly be, "How to artificially generate value from mergers". On the other hand, all our "optimization" research ideas stand on the fact that debtholders are valuing precisely the credit risk exposure in a game between shareholders and debtholders. The new accord could also force some financial institutions (from the roots of their cost of capital) to offer more sensitive credit conditions to the market.

Finally, future outcomes in the study of specific and systematic components of credit risk are not obvious to forecast. In the latter paragraph, we typically have a situation where individual reactions of SMEs to a common legal framework could lead to a bigger concentration, while at the origin, the aim of the new accord is to be more sensitive to each issuer of the marketplace.

Here is a little summary to conclude on the ideas presented above:

- We need a steady consensus on what drives credit risk, having a better understanding of the role of business cycles and the firm sensitivity to macro factors. What is then left as specific risk?
- We need a methodology for estimating probabilities of default and recoveries, consistent with firm-specific conditions of value creation and sustainable growth.
  - Managers need understanding.
  - Credit risk issues should be integrated to the balance scorecard.
  - There is currently no relationship with the economics of the firm's activity.
  - Sustainable growth of startup companies could be approached.
- We need a consensus on an acceptable portfolio approach to analyse credit risk diversification.
- We need to check:
  - Moral hazard and strategic consistency in pricing models.
  - Absence of bias in regulation leading to artificial competing advantages. This is essential for credit derivatives pricing.

FOOTNOTES & REFERENCES

1. It is not an extension at all of Merton. Contrary to what we could think, this can indeed be seen as an unfortunate reduction of Merton's case on the credit risk point of view. By assuming that a barrier can trigger default earlier, we need now to define a parameter for the loss function, while it is completely implicit in Merton's model.

2. Some have presented ways to analyse and reconcile both approaches, such as Gordy (2000) with a paper entitled "A comparative anatomy of credit risk models".

3. Be careful about the coherence of assumptions: if the firm model assumes payouts to shareholders and debtholders before the average maturity of the debt burden, then the debt pricing model must include such cash outflows.

Keeping aside the numerous references mentioned in the previous tables, the text presented above was highly inspired by the following authors:

■ Huge, Brian & David Lando (1999), 'Swap Pricing with Two-sided Default Risk in a Ratings-Based Model', *European Finance Review* **3**:239–268, 1999.

■ Leland, Hayne E., 1994, "Corporate Debt Value, Bond Covenants and Optimal Capital Structure", Journal of Finance, 49(4), September 1994, pp. 1213–1252.

■ Leland (1994), Leland and Toft (1996) for the determination of the optimal capital structure (and cost of capital).
■ Leland, H.E. and Toft, K.B., 1996, "Optimal Capital Structure, Endogenous Bankruptcy and the Term Structure of Credit Spreads", Journal of Finance, 51(3), July 1996, pp. 987–1019.
■ Lintner, J. (1965), "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets", Review of Economics and Statistics, 47, (February): 13–37.
■ Merton (1973, 1974) for the seminal structural model.
■ Merton, R.C. (1974), "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates", Journal of Finance, 29 (May): 499–70.
■ Merton, R.C. (1977), "On the Pricing of Contingent Claims and the Modigliani-Miller Theorem", Journal of Financial Economics, 5 (November): 241–9.
■ Mossin, J. (1967), "Equillibrium in a Capital Asset Market", Econometrica, 35 (October): 768–83.
■ Pirotte (1999) for the barrier extension in a two-factor setting for the interest-rate risk.
■ Pirotte, H., 1999a, "implementing a Structural Valuation Model of Swap Credit-Sensitive Rates", Working paper, Institute of Banking and Finance, Ecole des HEC, University of Lausanne, December 1999, 32 pp.
■ Pirotte, H., 1999b, "A Structural Model of the Term Structure of Credit Spreads with Stochastic Recovery and Contractual Design", Working paper, Institute of Banking and Finance, Ecole des HEC, University of Lausanne, December 1999, 85 pp.
■ Sharpe, W.F. (1964), "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk", Journal of Finance, 19 (September): 425–42.
■ Sharpe, Lintner & Mossin (1964, 1965, 1966) for the CAPM results and the use of the beta.

6
Modelling and Measuring Sovereign Credit Risk

Ephraim Clark

Quantitative Finance Review 2003

Contemporary credit risk modelling is dominated by two types of models, the structural models and the reduced-form models. The structural models are based on Merton (1974, 1977) and view bonds as contingent claims on the borrowers' assets. The credit event is modelled as timing risk when the assets of the borrower reach a threshold. In Merton (1974, 1977), Black and Cox (1976), Ho and Singer (1982), Chance (1990) and Kim, Ramaswamy and Sundaresan (1993), default is modelled as occurring at debt maturity if the assets of the borrower are less than the amount of the debt due. More recent models, starting with Longstaff and Schwartz (1995), have randomized the timing of the default event determined by when the value of the assets hits a predetermined barrier. The reduced-form models, such as Jarrow and Turnbull (1995), Madan and Unal (1998) and Duffie and Singleton (1999), model the timing of the default event as a Poisson process or a doubly stochastic Poisson process (Lando, 1994).

These models are often applied to sovereign credit risk, but, because of the nature of sovereign debt, they have some fundamental shortcomings. The reason is that both types of models were conceived in the context of corporate bankruptcy, where there is a legal framework for settling claims when the borrower is unable to service its debt. The inability to pay can be *de jure*, such as when certain conditions on asset values are not respected, or *de facto*, such as when the borrower runs out of cash. The assumption is that when default occurs, creditors can seize the firm's assets to "recover" as much as possible of what is owed them.

Because of the principle called "national sovereignty", which lies at the heart of the political world order, the foregoing context is clearly different from the context surrounding sovereign credit risk. First of all, there is no recognized legal framework for sorting out sovereign defaults.[1]

Contact address: Middlesex University, The Burroughs, London NW4 4BT, UK.
Email: e.clark@countrymetrics.com Telephone: 44-(0)208-411-5130. www.countrymetrics.com

When a sovereign defaults, creditors have very little scope for seizing assets, as is the case in corporate defaults. Consequently, instead of seizing assets that can be sold to pay off some or all of the delinquent debt, creditors and sovereigns negotiate. The result usually includes some combination of forgiveness, new money and, more importantly, the rescheduling of outstanding contractual maturities, i.e., the transformation of contractual maturities into longer-term liabilities.[2] This introduces an element of uncertainty regarding the actual maturities of the contractual cash flows, something that the structural and reduced-form models are not equipped to handle.

Besides making asset recovery in the case of default practically impossible, "national sovereignty" also endows countries with the *de facto* power to unilaterally abrogate or suspend contractual obligations when deemed in the national interest. This is another major characteristic that distinguishes sovereign debt from corporate debt. In the structural and reduced-form models, the creditworthiness of a corporate borrower depends, for all practical purposes, on its ability to pay. Where a sovereign borrower is concerned, however, besides the ability to pay, creditworthiness depends on the government's willingness or unwillingness to pay, even if it has the ability.

The literature on sovereign credit risk has not overlooked the importance of the willingness factor. Eaton, Gersovitz and Stiglitz (1986), for example, argued that because a country's wealth is always greater than its foreign debts, the real key to default is the government's willingness to pay. Borensztein and Pennacchi (1990) suggest that, besides other observable variables that are tested, the price of sovereign debt should be related to an unobservable variable that expresses the debtor country's willingness to pay. Clark (1991) suggests that the price of sovereign debt is related to a country's willingness to pay, which is motivated by a desire to avoid the penalties of default. Although analytically seductive, the problem with the concept of the willingness (unwillingness) to pay is that it is not readily observable. Consequently, empirical testing of the price of sovereign debt on the secondary market has tended to exclude this variable and focus on financial variables (Feder and Just, 1977; Cline, 1984; McFadden *et al*., 1985; Callier, 1985), structural variables (Berg and Sachs, 1988) or other phenomena, such as prices (Huizinga, 1989) or debt overhang (Hajivassiliou, 1989).

Some recent developments of sovereign debt modelling deal explicitly with the foregoing problems. Clark (2003) shows how the willingness to pay can be modelled as an American-style call option, where the decision to default depends on the government optimizing the trade-off between the gains to be reaped through non-payment and the costs associated with not paying. Clark and Zenaidi (forthcoming) show how the reality of rescheduling and the uncertainty of contractual maturities can be captured by modelling losses due to rescheduling and default as Poisson processes covered by a hypothetical insurance policy.

The rest of this chapter is organized as follows. In Section 1, I review the latest applications of the structural and reduced form models to sovereign debt. In Section 2, I outline the elements of the models that deal with the willingness to pay and rescheduling, while Section 3 and Section 4 conclude.

1 Structural and reduced form modelling of sovereign debt

There has been no reduced-form modelling specific to sovereign debt. Work on sovereign debt with respect to reduced-form models has concentrated on testing existing models and especially

the Duffie and Singleton (1999) model. For example, this is the case for Andritzsky (2003) and Zhang (2003) for Argentine bonds, Merrick (2001) for Argentine and Russian bonds and Keswani (1999) and Pagès (2000) on Latin-American Brady bonds. Dullman and Windfuhr (2000) test two affine diffusion models, the Vasicek (1977) and Cox–Ingersoll–Ross (1985) models, on European government credit spreads.

The key to structural modelling of sovereign debt is constructing an appropriate underlying security for a country that can drive the model. In the structural models for corporate debt, the underlying security is the value of the firm's assets. Clark (1991) has shown that the corresponding underlying security for sovereign credit risk is the country's international market value (net export value), measured as the present value of a country's expected net exports.[3] Thus, in order to apply a structural default model, it is necessary to estimate this variable and its parameters. Two methods have been proposed. Clark and Kassimatis (2003) use historical investment data, interest rate parity and forward rate parity, while Clark (2003) suggests an econometric estimation using historical data on net exports.

One of the most interesting developments is that this underlying international market value can be combined with sovereign debt market data to create derivative products capable of managing overall country risk. For example, one way to exploit the market information on traded sovereign debt is to borrow from standard option pricing techniques and estimate the implied volatility of the country's international market value (see Clark, 2002b). This implied volatility is a market-based measure of the country's underlying riskiness. It can be obtained by running the option pricing model with volatility as the unknown. The asset does not actually have to be traded. It can be an indicator similar to those used for weather derivatives, like the number of sunny days in a given month.

Once the country's implied volatility has been estimated it can be used directly as a risk management tool or serve as the underlying variable in other types of derivative products. Clearly, if we have a reliable underlying variable, a whole range of derivative products for country risk management, including forwards, futures, swaps and options, would be feasible. Within this range, volatility swaps are particularly pertinent to the management of country risk. Volatility swaps involve one side paying a constant, pre-determined volatility, S, on a notional amount of L, while the other side pays ω, which would be the implied volatility of the country's international market value calculated on the payment dates.

2 Modelling the unwillingness to pay

The argument for modelling the unwillingness to pay as an American-style call option goes as follows. Based on the generally accepted concept of national sovereignty, a government has an ongoing *de facto* right to repudiate or default on its foreign debt, if this is deemed in the national interest. There is, however, no obligation on the part of the government to default. Thus, it is an option. It is an American-style option because the government can default at any time it chooses. As I mentioned above, this it should do when deemed in the national interest, and the national interest is when the benefits from defaulting are large enough to offset the costs of doing so. Hence, if we measure the relative value of default with respect to the costs of defaulting, we are in effect measuring the degree of the government's unwillingness to honour its contractual debt obligations. The higher the value of the option to default, the less willing is the government to pay.

Thus, the value of the option, noted as Y, depends on the nominal amount of foreign debt outstanding, noted as x, and the indemnities and costs associated with default, noted as C:[4]

$$Y = Y[x(t), C(t)] \tag{1}$$

Both x and C are assumed to follow geometric Brownian motion:

$$dx(t) = \alpha x(t)\, dt + \sigma x(t)\, dz(t) \tag{2}$$

where α is the growth rate of the foreign debt, which depends on the economy's requirements for external financing, $dz(t)$ is a Wiener process with zero mean and variance equal to dt, and σ^2 is the variance parameter of $dx(t)/x(t)$:

$$dC(t) = \pi C(t)\, dt + \omega C(t)\, dw(t) \tag{3}$$

where π is the trend parameter, ω^2 is the variance parameter of the percentage change in $C(t)$ and $dw(t)$ is a Wiener process with zero mean and variance equal to dt, with $dz(t)\, dw(t) = \rho\, dt$, where ρ is the instantaneous correlation coefficient between x and C.

The argument for modelling debt as geometric Brownian motion is based on the nature of external transactions and on the fact that debt cannot be negative. More precisely, the continuous, random element stems from the balance of payments identity and the random, continuous nature of autonomous commercial and capital transactions.[5]

The argument for modelling indemnities and costs associated with default as geometric Brownian motion relates to the two forms that these indemnities and costs can take. The first revolves around the costs associated with the loss of access to capital markets (Eaton and Gersovitz, 1981). The second concerns the costs due to direct sanctions, such as the elimination of trade credits or the seizure of assets (Bulow and Rogoff, 1989).

The costs related to capital market access will be influenced by the economy's overall performance, which varies stochastically over time, and the extent to which foreign resources, both imported and borrowed, which also vary stochastically over time, play a role in the economy. The costs related to direct sanctions will be influenced by the reactions to the default of a wide range of players, including politicians, businessmen, bankers, international civil servants and consumers. Typically, these reactions vary according to circumstances and current perceptions surrounding them. Finally, perceptions themselves are likely to vary according to the evolution of a complex set of economic, political, social, environmental, etc. variables at the local, regional and international levels. In short, the sources of variation are numerous and unpredictable enough that there should be a considerable random element in variations of C.

Since the option to default can be exercised at any time, its value depends on when the option is exercised. The government will want to exercise at the point that maximizes its value. This problem can be solved with standard techniques. First, generate a new variable, $g = x/C$, the value of the investment per dollar of default cost, where the time arguments have been dropped for simplicity of notation.[6] Next, make the change of variables $y(g, 1) = Y(x, C)/C$ and assume time independence.[7] Use the capital asset pricing model to find R_g, the required rate of return on g, so that the instantaneous payout rate κ is equal to $R_g - \mu = \kappa$. Then, going through the well-known steps of setting up a riskless hedge consisting of one unit of

the option and $-y'(g)$ units of the investment, and applying Ito's Lemma, gives the following differential equation:

$$\tfrac{1}{2}\delta^2 g^2 y'' + (r-\kappa)gy' - ry = 0 \tag{4}$$

using the boundary conditions:

$$y(0) = 0 \tag{5}$$

$$y(g^*) = g^* - 1 \tag{6}$$

$$y'(g^*) = 1 \tag{7}$$

makes it possible to solve for y where g^* represents the value of g, where it will be optimal for the government to act. Reversing the change of variables gives:

$$Y = CK_1 g^{\eta_1} \tag{8}$$

where

$$\eta_1, \eta_2 = \frac{-(r-\kappa-\delta^2/2) \pm \sqrt{(r-\kappa-\delta^2/2)^2 + 2\delta^2 r}}{\delta^2}$$

$$K_1 = \frac{1}{\eta_1 - 1} g^{*-\eta_1}$$

and

$$g^* = \frac{\eta_1}{\eta_1 - 1}$$

Equation 8 expresses the government's unwillingness to honour its international debt obligations. The unwillingness grows as the gains from default grow with respect to the costs of defaulting. In this context, default itself is the result of a rational welfare optimizing decision based on relative costs and benefits.

Table 1 shows how changes in the parameters that drive the exercise price affect the value of the option to default. An increase in the trend parameter of x increases the value of the option, an increase in the trend parameter of C, decreases the value of the option, as does an increase in the correlation of the changes in the exercise price with changes in the amount of debt outstanding. The intuition behind these results is straightforward. Increases in x increase the value of the call, increases in the exercise price reduce the value of the call, and the more they move together, the more the increases in the value of the debt outstanding are offset by increases in the cost of default.

TABLE 1: COMPARATIVE STATICS ON THE OPTION TO DEFAULT

$\partial Y/\partial\alpha > 0$	$\partial Y/\partial\pi < 0$	$\partial Y/\partial\rho < 0$	$\partial Y/\partial\sigma \leq 0, \geq 0$	$\partial Y/\partial\omega \leq 0, \geq 0$

Interestingly, the effect of changes in σ and ω, the volatilities of x and C, is ambiguous. Depending on the levels of the other parameters, as well as the levels of σ and ω themselves,

the effect can be positive, negative or zero. In fact, because of the interaction of σ, ω and ρ in determining the dynamics of g (δ and μ in footnote 6), the effect of an increase in either σ or ω, up to a certain critical level, has the effect of reducing the value of x. Above this critical level, further increases have the effect of increasing the value of x.

The foregoing model can be used for measuring and monitoring sovereign risk. The distance to default can be measured as $g^* - g$, the difference between the optimal ratio of x to C and the current value of this ratio. Default occurs when $g = g^*$. The probability of default, then, is the probability that g will remain below g^*. To calculate the probability of default, the fact that g is lognormally distributed gives:

$$P[g < g^*] = 1 - P[g \geq g^*] \tag{9}$$

where

$$P[g \geq g^*] = \frac{1}{\sqrt{2\Pi\delta^2 T}} \int_{-\infty}^{c_1} \frac{1}{g} e^{-(\ln g - m)^2/2\delta^2 T} dg \tag{10}$$

and

$$c_1 = \frac{\ln(g/g^*) + (\mu - \delta^2/2)T}{\delta\sqrt{T}} \tag{11}$$

To implement this procedure, first determine the time horizon or the family of time horizons, $T = T_1, T_2 \ldots T_n$, deemed to be relevant and then apply equations 9–11.

3 The total expected loss from default as the value of a hypothetical insurance policy

Modelling the total expected loss from default as the value of a hypothetical insurance policy that pays off any and all losses due to *de facto* or *de jure* default addresses the reality that there is no recognized legal framework for sorting out sovereign defaults and very little scope for creditors to seize assets as a means of recovering their loans. Thus, default is usually not a one-off affair, as in corporate default. It is more likely to be a series of events where losses result from a negotiated settlement involving a combination of forgiveness, new money, concessional terms and the rescheduling of outstanding contractual maturities, i.e., the transformation of contractual maturities into longer-term liabilities.

In this model there are two types of default events that occur at random times, according to Poisson arrival processes. Type 1 refers to defaults accompanied by forgiveness, rescheduling and/or new lending that cause losses but keep the loans alive. Type 2 refers to the more extreme class of defaults, such as repudiations, that effectively kill the loans. Let q_1 for type 1 defaults and q_2 for type 2 defaults represent random variables that increase by steps of 1 every time a Poisson event occurs, with λ and ϕ their constant intensity parameters, such that:[8]

$$dq_1(t) = \begin{cases} 1 & \text{with probability } \lambda\, dt \\ 0 & \text{with probability } 1 - \lambda\, dt. \end{cases} \tag{12}$$

$$dq_2(t) = \begin{cases} 1 & \text{with probability } \phi\, dt \\ 0 & \text{with probability } 1 - \phi\, dt. \end{cases} \tag{13}$$

The amount of losses due to type 1 events will be influenced by the reactions to the default of a wide range of players, including politicians, businessmen, bankers, international civil servants and consumers. Typically these reactions vary according to circumstances and current perceptions surrounding them. Finally, perceptions themselves are likely to vary according to the evolution of a complex set of economic, political, social, environmental, etc. variables at the local, regional and international levels. In short, the sources of variation are numerous and unpredictable enough that there should be a considerable random element in variations of the loss in the case of default. Let x^* represent the loss given default as a percentage of outstanding nominal debt that follows geometric Brownian motion:

$$dx^*(t) = \alpha x^*(t)\,dt + \sigma^* x(t)\,dz(t) \tag{14}$$

Thus, in the case of type 1 events, the expected loss per interval dt is equal to $\lambda x^*(t)\,dt$. In the case of type 2 events, the loan is effectively ended and the expected loss per interval dt is equal to $\phi x(t)\,dt$.

To measure the total expected loss, V represents the value of a hypothetical, open-ended insurance policy that covers creditors against losses arising from the country risk, so that when losses occur, they are reimbursed by the insurance. "Open-ended" refers to the fact that the amount of debt covered by the policy can vary over the life of the policy as new loans are contracted and old loans are rolled over or paid off.[9] In this way, the policy measures the present value of total expected losses due to default over its specified life. The insurance policy, like the debt, is time-independent.[10] Thus, $V = V[x^*(t)]$. The expected total return on the insurance policy is equal to $E(dV)$ plus the expected cash flows generated by the explicit events. For type 1 events, the expected cash flow is equal to the expected loss, $\lambda x^*(t)\,dt$. For type 2 events, the insurance policy is cashed in for x; thus, the expected cash flow is $\phi(x - V)\,dt$.

Applying Ito's Lemma and taking expectations gives:

$$rV\,dt = V_x \alpha x^*\,dt + \frac{1}{2} V_{xx} \sigma^2 x^{*2}\,dt + \lambda x^*\,dt + \phi(x - V)\,dt \tag{15}$$

where the subscripts denote first and second partial derivatives. Using the boundary conditions of no speculative bubbles and an absorbing barrier when x^* goes to zero gives the solution:

$$V = \frac{\lambda x^*}{r + \phi - \alpha} + \frac{\phi x}{r + \phi} \tag{16}$$

Equation (16) quantifies country default risk. For it to make economic sense, $r + \phi - A$ must be greater than zero. It says that the value of the insurance policy is equal to the present value of the expected losses due to default discounted at the riskless rate plus a premium for the probability of a policy ending event.

4 Conclusion

Sovereign debt differs from corporate debt in several significant ways. When a sovereign defaults, creditors have very little scope for seizing assets, as is the case in corporate defaults. Furthermore, countries have the *de facto* power to unilaterally abrogate or suspend contractual obligations when deemed in the national interest. Consequently, sovereign defaults often depend more on the

government's willingness to pay than on its objective ability to do so. When a default does occur, instead of seizing assets that can be sold to pay off some or all of the delinquent debt, its resolution usually includes some combination of forgiveness, new money and the rescheduling of outstanding contractual maturities. In this context, the total expected loss due to sovereign default includes the reality of a series of loss-causing events as well as one, single, claim-ending repudiation.

In this chapter I have outlined how these characteristics can be modelled. The unwillingness to pay can be modelled as a call option on the nominal amount of outstanding debt with a stochastic exercise price. The resulting distance to default and default probabilities thus reflect the reality that the government's willingness to pay is a major determinant of sovereign default risk. The total expected loss due to default that includes a series of loss causing events can be modelled as the value of a hypothetical insurance policy that pays off any and all losses resulting from default, where loss causing events are modelled as Poisson processes. It reflects the fact that the total expected loss includes the reality of a series of smaller loss-causing defaults, as well as the possibility of a total, definitive repudiation.

FOOTNOTES & REFERENCES

1. There are semi-official organizations, such as the Paris Club for sovereign creditors and the London Club for private creditors.
2. The negotiations are usually carried out in the framework of the London and Paris Clubs.
3. Gray *et al.* (2003) propose a sectoral approach to modelling the macro financial risks of an economy. Their macro-economic flow of funds in their Appendix 7 corresponds to net exports.
4. As in corporate defaults, sovereign default on a debt service payment puts the total debt outstanding in default through *pari passu* and cross-default clauses that are routinely written into the debt contracts. In practice, once default has occurred and the government has demonstrated its willingness to suffer the costs this entails, a bargaining process begins, usually within the Paris and London Clubs, whereby the government enters negotiations with its creditors to trade the value of the exercised default option by recommencing payments in exchange for concessions such as forgiveness, reschedulings, etc. Our analysis is limited to the initial decision to default.
5. We consider jumps to new levels of nominal debt outstanding through forgiveness, rescheduling, Brady deals or the like as part of the negotiation process that occurs subsequent to the act of *de facto* or *de jure* default. This will be the subject of the next section.
6. The dynamics of this variable are:

$$dg = \mu g\, dt + \delta g\, ds,\ \mu = \alpha - \pi - \sigma\omega\rho + \omega^2,\ \delta^2 = \sigma^2 - 2\sigma\omega\rho + \omega^2,\ ds = \frac{\sigma\, dz - \omega\, dw}{\delta}$$

7. This is a common assumption, adopted, for example, by Modigliani and Miller (1958), Merton (1974), Black and Cox (1976) and Leland (1994). Leland (1994) justifies this assumption based on the conclusions of Brennan and Schwartz (1978), whereby the optimal strategy is continuously rolled over short term debt under the constraint of positive net asset value. Thus, as long as the firm is able to repay its debt, the debt is automatically rolled over.
8. In fact, the intensity parameters might be stochastic and/or changing over time. This can be incorporated by modelling the jumps as a doubly stochastic Poisson process (Cox process) with Bayesian updating (see Clark and Tunaru, 2003).

9. The hypothetical insurance policy in question should measure the cost of overall country default risk. Thus it is the amount of debt that is important and not the identity of the individual creditors, who can change over the life of the policy.
10. The special case of when the debt has a definite maturity is treated in Clark and Zenaidi (forthcoming).

■ Andritzsky, J. (2003) Implied default probabilities and default recovery ratios: an analysis of the Argentine crisis 2001/2002. University of St. Gallen.
■ Berg, A. and Sachs, J. (1988) The debt crisis: structural explanations of country performance. *Journal of Development Economics* 29(3), 271–306.
■ Black, F. and Cox, J. (1976) Valuing corporate securities: some effects of bond indenture provisions. *Journal of Finance* 31(May), 351–367.
■ Borensztein, E. and Pennacchi, G. (1990) Valuing interest payment guarantees on developing country debt. *IMF Staff Papers*.
■ Brennan, M. and Schwartz, E. (1978) Corporate income taxes, valuation and the problem of optimal capital structure. *Journal of Business* 51, 103–114.
■ Bulow, J. and Rogoff, K. (1989) Sovereign debt: is not to forgive to forget? *American Economic Review* 79(1), 43–51.
■ Callier, P. (1985) Further results on countries' debt servicing performance: the relevance of structural factors. *Weltwirtschaftliches Archiv* 121, 105–115.
■ Chance, D. (1990) Default risk and the duration of zero coupon bonds. *Journal of Finance*, 45, 265–274.
■ Clark, E. (1991) *Cross Border Investment Risk*. Euromoney Publications, London.
■ Clark, E. (2002a) *International Finance*. International Thomson Business Press, London.
■ Clark, E. (2002b) Measuring country risk as implied volatility. *Wilmott*.
■ Clark, E. (2003) Sovereign debt default risk: quantifying the (un)willingness to pay. *Wilmott*.
■ Clark, E. and Zenaidi, A. (in press) The quantification of country risk: determinants of secondary market sovereign debt discounts. *Sovereign Debt and Country Risk*, Karmann, A. and Scholtens, B. (eds). Springer-Verlag, Heidelberg.
■ Clark, E. and Kassimatis, K. (2003) Country financial risk and stock market performance: the case of Latin America. *Journal of Economics and Business* 55(6).
■ Clark, E. and Tunaru, R. (2003) Quantification of political risk with multiple dependent sources. *Journal of Economics and Finance* 27(2), 125–135.
■ Cline, W. R. (1984) *International Debt: Systematic Risk and Policy Response*. Institute for International Economics, Washington, DC.
■ Cox, J., Ingersoll, J. and Ross, S. (1985) A theory of the term structure of interest rates. *Econometrica* 53, 385–408.
■ Duffie, D. and Singleton, K. (1999) Modeling term structures of defaultable bonds. *Review of Financial Studies*, 12, 687–720.
■ Dullman, K. and Windfuhr, M. (2000) Credit spreads between German and Italian sovereign bonds: do affine models work? Working Paper, University of Mannheim.
■ Eaton, J. and Gersovitz, M. (1981) Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies* 48(152), 289–309.
■ Eaton, J., Gersovitz, M. and Stiglitz, J. (1986) A pure theory of country risk. *European Economic Journal*.
■ Feder, G. and Just, R. (1977) A study of debt servicing capacity applying logit analysis. *Journal of Development Economics* 4(March), 25–38.

■ Gray, D. F., Merton, R. and Bodie, Z. (2003) A new framework for analyzing and managing macrofinancial risks of an economy. MF Risk Working Paper, 1-03, August 1.
■ Hajivassiliou, U. A. (1989) Do the secondary markets believe in life after debt? Working Paper 252, International Economics Department, The World Bank, pp. 1–42.
■ Ho, T. and Singer, R. (1982) Bond indenture provisions and the risk of corporate debt. *Journal of Financial Economics* 10, 375–406.
■ Huizinga, H. (1989) How has the debt crisis affected commercial banks? Working Paper 195, International Economics Department, The World Bank, pp. 1–32.
■ *International Financial Statistics*, The International Monetary Fund, Washington, DC: several issues.
■ Jarrow, R. A. and Turnbull, S. M. (1995) Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50(1), 53–85.
■ Judge, G. G. *et al*. (1985) *The Theory and Practice of Econometrics*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.
■ Keswani, A. (1999) *Estimating a Risky Term Structure of Brady Bonds*. London Business School.
■ Kim, J., Ramaswamy, K. and Sundaresan, S. (1993) Does default risk in coupons affect the valuation of corporate bonds? A contingent claims model. *Financial Management* 117–131.
■ Lando, D. (1994) On Cox processes and credit risky bonds. Working Paper, Institute of Mathematical Statistics, University of Copenhagen.
■ Leland, H. (1994) Corporate debt value, bond covenants, and optimal capital structure. *Journal of Finance* 49(September), 1213–1252.
■ Longstaff, F. and Schwartz, E. (1995) A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50(3), 789–819.
■ Madan, D. and Unal, H. (1998) Pricing the risks of default. *Review of Derivatives Research* 2, 79–105.
■ Merrick, J. (2001) Crisis dynamics of implied default recovery ratios: evidence from Russia and Argentina. *Journal of Banking and Finance* 25, 1921–1939.
■ Merton, R. (1973) Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–183.
■ Merton, R. (1974) On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* 29(May), 449–470.
■ Merton, R. (1977) On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* 5, 241–249.
■ Modigliani, F. and Miller, M. (1958) The cost of capital, corporation finance and the theory of investment. *American Economic Review* 38(June), 261–297.
■ Pagès, H. (2000) Estimating Brazilian sovereign risk from Brady bond prices. Working Paper, Banque de France.
■ Palac-McMiken, E. D. (1995) *Rescheduling, creditworthiness and market prices*. Avebury, London.
■ Saunders, A. (1986) The determinants of country risk. *Studies in Banking and Finance* 3, 2–38.
■ Vasicek, O. A. (1977) An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–188.
■ Zhang, F. X. (2003) What did credit market expect of Argentina default? Evidence from default swap data. Federal Reserve Board, Division of Research and Statistics, April 16.

7

The Equity-to-credit Problem (or the Story of Calibration, Co-calibration and Re-calibration)

Elie Ayache*

Quantitative Finance Review 2003

This essay was initially intended as the elaboration of the presentation I gave at the Quantitative Finance Review in London in November 2003. The original title of the talk was: "The equity-to-credit problem, or how to optimally hedge your credit risk exposure with equity, equity options and credit default swaps".

1

As will be soon apparent from my line of arguing, I will indeed tackle, in this essay, what has become an urgent issue in those fields where credit volatility and equity volatility intermingle, typically the pricing and hedging of convertible bonds. Originally perceived as equity derivatives, the hybrid securities known as convertible bonds are steadily drifting towards the class of credit derivatives. Indeed, the last stories of default have demonstrated the frailty of what used to be the rock-bottom value of the convertible bond and a bedrock notion in its literature

Contact address: ITO 33, 36 Rue Lacépède, 75005 Paris, France.
E-mail: numbersix@ito33.com

* I am greatly indebted to Philippe Henrotte, whose thought and brilliant mind lie at the heart of the argument developed in this essay. Philippe's continuing advice and tireless conversation have been my most reliable guides, not only through my present writing process but also in my most general thinking.

and analysis, the "bond floor". It is probably more appropriate today to say that convertible bonds are derivative both on the equity state and the default/no default state of the issuer as it is no longer sufficient, for the purposes of quantitative analysis, to only specify the payoff of the convertible bond in case of conversion into the underlying share. Indeed, a state-of-the-art pricing model almost certainly requires, in addition, the specification of the payoff of the convertible bond in case of default. I refer the reader to Ayache *et al* (2002, 2003) where all these insights are quantitatively fleshed out.

Convertible bond pricing is perhaps the first derivative pricing problem to have raised the question of the explicit relation between the credit spread and the share price. The hybrid nature of the instrument is not the only reason, for the delta-neutral convertible bond volatility arbitrageurs have long been worrying about the adjustment of the equity delta implied by such a relation. I shall concern myself with this question, which epitomizes the equity-to-credit problem. Let it be noted, in passing, that traders and arbitrageurs of non-convertible corporate debt and even, in some cases, of pure credit derivative instruments such as credit default swaps, are awakening to the notion of equity delta too. Hedging the credit exposure with the traded equity of the issuer is another name for the equity-to-credit problem, and it reaches far beyond the confines of convertible bonds.

2

Although a proper exposition of the equity-to-credit problem is supposed to follow the particular order of moving from the credit problem as such to the bearing of the equity process on it, I will follow the reverse order in this essay. Instead of asking what the equity can bring to the credit problem, I will ask what the credit risk, or in other words the likelihood of default of the issuer, can bring to an outstanding problem in the equity derivatives field, namely the equity volatility smile. Convertible bonds are a bad case for distinguishing between subtleties of orders of exposition like the ones I am pointing out. Obviously the convertible bond quantitative analyst must concern himself *both* with credit spread term-structures and implied volatility smiles, as witnessed in Andersen and Buffum (2002). Let it be noted, however, that convertible bond pricing has complexities of its own that have recently engaged, and are still absorbing, the specialists. Beside the complexity inherent in the structure itself (the non-linear interplay of the multiple embedded options, the option to convert the bond, to redeem it earlier than maturity, to put it back at a fixed price, etc.) and the growing popularity of exotic features (make-whole premiums, contingent conversions, variable conversion ratios, etc.), it is the proper treatment of default risk that has been the greatest subject of concern recently. Again, I refer the reader to Ayache *et al* (2002, 2003). One will hardly want to worry about stochastic volatility and volatility smiles on top of all these problems! The analyst is usually content to specify a certain constant volatility parameter in his convertible bond pricing engine, inferring it roughly from the implied volatility of equity options of similar strike and maturity, and turns to what is the most pressing problem to his mind, that of calibrating a suitable default intensity term-structure and, in the most demanding cases, of worrying whether default intensity should not be made a function of the underlying equity as well.

It has been the rule, for all those reasons, that the equity volatility smile problem did not pose itself *per se* in convertible bond pricing. This is despite the fact that convertible bonds are highly exotic derivative instruments! With the issuer's option of early redemption acting as a knock-out and contingent conversion acting as a knock-in and their triggers being located far away from

the conversion price, can we feel comfortable using a single implied volatility number in our pricing engine? Not to mention that early puts, and likelihood of default, can also considerably shorten the effective life of the bond. Have we not learnt from the independent smile literature that exotic option pricing is irreducibly entangled with smiles and smile dynamics (Ayache *et al*, 2004)? Yet convertible bond pricing models continue to lead their lives separately from smile models. A very tempting proposition could be to make use of local volatility in convertible bond models. State-of-the-art convertible bond pricing engines, such as those produced by my own company, are equipped to deal with local default intensity, or hazard rate, anyway; they rely on unconditionally stable finite difference schemes and adaptive stepping as a prerequisite for solving what may become a very hard numerical problem when local hazard surfaces need to be calibrated to credit default swap market data, and convertible bond deltas and gammas computed off them; so why not simply overlay the numerical solver with a local volatility surface?

Tempting as the proposition may be – as a matter of fact it is starting to receive some attention (Andersen and Buffum, 2002) – I am resentful of the very way it suggests itself to us. I will not dwell on the well-rehearsed and very deep arguments that go against local volatility, precisely when exotic option pricing is of concern. The general philosophy of my company, in this regard, has been given expression in Ayache *et al* (2004). What is noticeable, and I think most dangerous, here, is the way the complexity – or shall I say the perversity? – compounds itself. Since the model already accommodates local hazard rate surfaces, so the argument goes, let us add local volatility surfaces. For how can we sustain the pressure exerted on the single volatility number in our pricing engine any longer? Can we reasonably become the specialists of local hazard rate surfaces and their calibration routines, can we reasonably propose a tool which has exhausted one side of the problem and allows the calibration of full non-parametric hazard rate surfaces to full surfaces of credit spread data (as against maturity and equity level), and not set volatility free on the other side? What meaning can a single volatility number retain, and how can it sit tight as the only remaining hinge, in the midst of such a complex system? Shouldn't the two surfaces be considered simultaneously and the two problems solved hand-in-hand? Can we solve for local hazard rate and not jointly solve for local volatility? Can we calibrate to surfaces of credit default swap spread data and not jointly calibrate to implied volatility surfaces?

3

This is our problem, precisely. This is precisely the equity-to-credit problem. Since default (or its probability) and the subsequent drop of the underlying share are among the greatest creators of implied volatility skew and term structure, there is no way we could calibrate the default process without worrying about the option data. The equity-to-credit problem is essentially a smile problem. As a matter of fact, it is even worse than a smile problem. It speaks of a more complex underlying process (jump-diffusion) and compounds two ill-posed inverse problems: calibration of volatility instruments and calibration of credit instruments. (There already exists a thick literature on the subject of smoothing, interpolating and extrapolating an arbitrage-free local volatility surface, so try to imagine the result of adding the worry about a hazard rate surface!)

Again, notice how the historical build-up of the problem, and the historical succession of its proposed solutions (among which those proposed by my own company), have got us into trouble. Because the convertible bond pricing problem is such a complex problem to begin

with (hybrid nature, embedded options, endless exotic features), trying to formulate it outside a strict diffusion framework is the last thing we want to do! It is not as though the pricing of the convertible bond had become difficult, or perhaps even impossible, on account of the underlying diffusion process, and this gave us sufficient reasons to reconsider that process and contemplate jump-diffusion or stochastic volatility instead! The pricing of convertible bonds has become difficult – and in some cases of outdated software, even impossible – for reasons having strictly to do with the convertible, not the process. Only in simpler cases, where simpler exotic structures are liquidly traded, for instance FX barrier options, has it prominently appeared that the smile problem should be treated as such and solved as such. We have so many other things to worry about in convertible bond pricing, before we get down to the smile problem! Enough for now to have taken that first step outside diffusion, which consisted in adding a hazard rate and a jump to zero of the underlying share! (How many people, by the way, recognize the fact that this "simple" jump-diffusion process is already posing a full smile problem, and that the implied volatility number they are importing into their convertible bond pricing system from the prices of equivalent equity options, no longer means the same?)

Against this conservative-sounding kind of argument, I say this is precisely the time when, on the contrary, a break with history should occur. Who said we should ever consider local hazard rates in the first place? Because of the wrong sense of complexity that the convertible bond may convey, the convertible bond problem is probably a wrong place to pose the equity-to-credit problem afresh, and start looking for radical alternatives for solving it. Like we said, convertible bond pricing specialists have been distracted by the wrong kind of difficulty – and I sometimes fear we might be among them – and for this reason perhaps, lack the freshness of the eye. This is why I have set out to pose the equity-to-credit problem as an equity smile problem rather than a hybrid problem, and consider my task today to be the continuation of the work on smiles and smile dynamics pioneered in Ayache *et al* (2004) as well as to find out whether default risk, and its traded instruments such as credit default swaps, cannot help us frame the smile problem better, by any chance. Only when the equity-to-credit problem is addressed and solved as a smile problem proper, with the same critical spirit and independence of vision as we have exercised in our previous work, will we go back to convertible bonds and realize what we should have known from the start, namely, that their complexity ought to be the driving motive for wanting to base their valuation on a sound theoretical ground – be it at a big intellectual and cultural cost – not an excuse for evading it.

4

Anyone familiar with our general philosophy of derivative pricing must be guessing my point at this juncture. If one thing is really dear to my heart when it comes to the smile problem, it is the defence of homogeneous models. The case for homogeneity has been extensively argued in Ayache *et al* (2004). There we showed, among other things, that the pricing of exotics and the hedging of both vanillas and exotics are crucially dependent on the smile dynamics, and that the need to discriminate between the empirical smile dynamics (with sticky-strike and sticky-delta sitting at the two extremes) need not be answered by inhomogeneous models. We argued, on the contrary, that a sufficiently comprehensive homogeneous model – a model we christened "Nobody" – can perfectly address the exotic pricing and hedging issues under all kinds of smile dynamics, provided that (a) the hedging problem is aptly formulated in the incomplete markets framework, and (b) the model is calibrated to those *quoted* market prices which are

the empirical reflection of the projected smile dynamics, the prices of the one-touches and the forward starting options.

This is how we dispensed with the local volatility model and all its cross-breeds (universal volatility, etc.). In yet another paper (Henrotte, 2004), we argued that even in fields where inhomogeneous models are so deeply entrenched as to go unnoticed, for instance the modeling of yield curves and credit spread curves, homogeneous models can have right of way. Term structure is the direct consequence of stochastic character, after all, not of a dissymmetry inherent in time, and a parsimonious time-homogeneous stochastic interest rate model, or stochastic hazard rate model, or stochastic volatility model, can demonstrably reproduce any shape of interest rate, or credit spread, or volatility term-structure.

Insisting that the equity-to-credit problem shall be posed and solved as a smile problem will therefore strike our reader as an overt directive against inhomogeneous models. Unpacking the claim, this would mean that volatility and hazard rate have to be modelled as two independent, time and space homogeneous, stochastic processes, instead of being deterministic functions of time and the underlying. Smile dynamics and credit dynamics would then be accounted for by a suitable correlation with the equity process.

This sounds like a daring claim indeed, all the more so when structural models of the firm (Merton, KMV, CreditGrades™) seem to have entrenched the view that the triggering of default is a deterministic function of the underlying equity price. The myth of the bankruptcy threshold has transformed the very liberal notion of probability of default into the very concrete vision of a "distance to default", and it is commonplace nowadays to speak of credit spreads that explode to infinity with a falling share, and of bond floors that strictly collapse to zero. As if somebody had ever managed to measure such infinite spreads, or truly held bonds whose value vanished strictly prior to default and not directly after!

Rather, this picture strikes me as the result of the confusion of possibility and actuality, and it is unfortunately imposed on us by the particular mathematical representation. Take, for instance, the so-called "reduced-form" models which are supposed to disconnect the structural link between default and the equity price. The underlying equity, however, remains the state variable governing the probability of default and this particular choice naturally leads us to sampling equity values which are very close to zero (if only because the PDE numerical schemes require suitable boundary conditions). And now the additional twist brought by default is that a vanishing equity price – which may be vanishing only formally, just for the sake of writing the boundary condition – cannot not *get mixed up* with the question of default. This causal confusion stems from the fact that it all looks as if the equity price were the *sole* state variable, therefore the cause of everything. As a matter of fact, since default is supposed to be triggered independently by a Poisson process, a second state variable should be recognized here: the Boolean admitting of "default" or "no default" as values.

Now of course a natural and most comprehensible recommendation in such a setting is to suggest that the intensity of that process (or in another words the probability of its triggering) may indeed increase with a falling equity price (see Appendix). But this need not imply in any way that the probability of default should become equal to one exactly, or the intensity reach infinity, at the limit where the share price is equal to zero. There is no reason why a phenomenon (default), which normally originates from an independent cause (or a complex set of causes summarized by the Poisson process), should know of no other cause, at the limit, but the share price. The problem with equity-to-credit, when forced to fit inside the inhomogeneous, deterministic representation of the relation, is that it forces all kinds of boundary conditions

on us, not just the mathematical, formal one but also a material, causal one. Just because the intensity of the default process is driven by the equity, we are forced to assign a value to the probability of default for all equity prices, particularly when the equity is worth zero, and our state of confusion about the true causal origin of the phenomenon (whether it is default which causes the zero share price or the zero share price which causes default) leaves us no choice but to assign infinity at the boundary and hope to have settled the issue with that. Any number other than infinity would indeed seem unsatisfactory, or call for an even bigger problem. Why should the credit spread be worth 20% rather than 30% at the limit of a worthless share? We certainly do not wish to get involved into this extreme sort of corporate finance and, like a game theorist friend of mine once said, "Infinity, in some cases, is the best approximation of an otherwise arbitrary finite number!"

5

Think, by contrast, of a situation where the default intensity is an independent stochastic process. Think, for instance, of a simple two-factor model such as Nobody, where the second factor is taken care of by a discrete number of "regimes" (depending on the particular problem, this may be volatility, or the hazard rate; see Henrotte, 2004). Each regime is characterized by a given, constant, hazard rate, while the share price can formally vary from zero to infinity in each. Regimes can be interpreted as different ratings of the issuing firm and transitions between regimes are governed by a probability transition matrix. We must consider correlation between the share price process and the regime transitions as it is indeed very unlikely that the share should start falling dramatically and the firm not switch over to a regime of higher default intensity, or indeed to the default regime! (I forgot to say that default receives an interpretation, in such a framework, which is of a piece with the rest. In our discrete regime approach, default is a regime like any other, only a very special one where the hazard rate, or the *probability* of default, is no longer defined. See Appendix.)

Whatever the realistic interpretation of the model may be, the point is that the model does not impose on us restrictions, or philosophical boundary conditions, such as imposed by the inhomogeneous model above. When the firm is alive, it subsists in one of our discrete regimes, with a given finite hazard rate. As the regime representation is only a discretization of the credit state space, the general objection that could be levelled against us, according to which the firm may, as a matter fact, not fall exactly in any of our discrete regimes but somewhere in between, is answered by the fact that the regimes are probabilistically connected and it is only the *probabilistic averages* (i.e. prices) that matter. So long as default is recognized as the extreme regime in our discrete collection, what specific hazard rate numbers, or credit spreads, the other regimes get, no longer matters. And that each individual regime may formally allow, as its boundary condition, that the share may approach zero and the hazard rate remain finite, even constant, will not matter either.[1]

The material process and the formal process are separated in our model. Although there formally exist states of the world where the share is close to zero and the credit spread no larger than a given finite number, these scenarios will materially get a very low probability because of the correlation between the share and the regime shifts. Moreover, correlation can itself depend on the regime and increase as the firm shifts to regimes of higher default intensity. In other words, to answer the question: "What happens when the share collapses?" all you have to do is follow its *material* price process as it really unfolds in real time, and observe that

the firm will jump between regimes, and eventually default. Yet you ask: "Will the hazard rate ever reach infinity?" The answer is: "Not in any of the live regimes, no, and not in the default regime either, for then it is too late." Yet again: "Can we not imagine that the firm is actually climbing to regimes of explosively high default intensity as the share approaches zero, only those regimes are invisible to our discrete representation?"

Why not indeed, but then it is precisely the advantage of indeterminism over determinism (of correlation over deterministic functions) not to force on us a determinate answer to that question. It is precisely the advantage of having separated the hazard rate and the equity price into two distinct state variables, and of having distinguished between formal cause and material cause, to allow us to solve (or shall we say, dissolve) both problems simultaneously. "Why have 20% credit spread instead of 30% when the share price is equal to zero?" and the answer is: "Let us have both. As a matter of fact, let us have other values as well, and the question of the boundary condition will not matter anymore because it is only a formal device, required in each regime." "What is the real boundary condition then?" and the answer is: "It disappeared in the probabilistic 'interspace'. While the inhomogeneous model cannot avoid mixing the bottom of the share price spectrum with the certainty of default, all that our homogeneous model offers, by contrast, is a collection of regimes where the bottom of the share price is in each case immaterial to default, and a *separate* default regime."

6

Taking the suggestion that default is a regime like any other more seriously, we now realize that it has been with us all along and implicit in everything we have ever said about default, even in the traditional inhomogeneous framework. Surely enough, the hazard rate was a function of the share price in that framework and default became certain as soon as the share hit its lowest boundary, but the jump into default itself and its consequence for both the derivative instrument and the underlying share were no different from what we are contemplating today for the homogeneous case. As a matter of fact, we have already suggested in the previous convertible bond article (Ayache *et al*, 2002) that "a softer appellation of the state of default could be 'distress regime' " and that "it would certainly make sense to imagine a continuation of life after default." What makes life end at default, after all, is just the assumption that the share drops to zero and the derivative instrument to its recovery value and that the game is over. But what if the share did not drop to zero but to some recovery value as well, then resumed its trading? What if the event of default triggered a restructuring of the issuing firm and of its outstanding debt, and the holders of the bonds were offered the opportunity to hold on their assets and stand by their positions until further notice? What if the holder of the convertible bond found it more optimal to postpone until later his right to convert into the underlying share – provided he still owned a convertible bond after the restructuring – rather than opt out of the game at the moment of default and take away the recovery value of the bond?

Theoretical and impractical as these questions may sound, they have at least the merit of making us think how to properly extend our framework if need be. The suggestion in Ayache *et al* (2002) was that we would end up solving a pair of coupled PDEs, one describing the pre-default regime and the other the default, or distress, regime. The regimes are coupled through the Poisson jump to default, and the transition is supposed to be irreversible in the sense that the firm cannot recover from the distress regime, back to the normal regime. As the share starts its journey, in the distress regime, with the recovery value it has hit after default, it was even

suggested that its subsequent volatility might be different from the volatility in the normal, pre-default regime.

To summarize: If the assumption should ever be made that the share might not drop to zero upon default but to some recovery value and then resume its trading life, and if we should ever consider holding on our instrument for reasons such as restructuring or rescheduling, then the proper way to value the instrument *as of today*, would be to solve a system of two coupled PDEs, possibly with different diffusion coefficients, perhaps even with different payout conditions written on the instrument (for instance, the conversion ratio may change after restructuring, or coupons may no longer be paid, etc.).

7

As we stand now, we are just one step away from the full, homogeneous, equity-to-credit model that we've been hinting at. If the thread of default is capable alone of leading us to a regime-switching model, even in the classical inhomogeneous case, what is to keep us from spreading the idea over to the no-default side of the picture? If volatility is allowed to be different in the default regime, why wouldn't it be different in each of the pre-default regimes which corresponded, in our homogeneous model, to different ratings of the issuing firm? In other words, the suggestion here is this: Let *both* a stochastic volatility process and a stochastic hazard rate process be taken care of by the regime representation. The regime is not a traditional state variable, after all, in the sense that volatility is one such, and the hazard rate is another, in the traditional stochastic volatility or stochastic hazard rate models that people usually have in mind. A regime can be identified, not just by one hazard rate number λ as proposed in Henrotte (2004) or one volatility number σ as proposed in Ayache *et al* (2004), but by the mathematical pair (σ, λ). (See Appendix.) As a matter of fact it can generally be identified by the n-tuple $(\sigma, \lambda, r, s, \ldots)$ where r, s, ... might be other parameters of the pricing equation that we want to turn stochastic, such as the short-term interest rate, etc.

We are touching here on an interesting, and I think, very deep, idea. This is the idea that a pricing problem might be multi-factor yet we are able to handle it with the same unchanged regime representation. I am being very cautious here in picking the right words. Notice that I did not speak of a "multi-dimensional" pricing problem. People are used to thinking that anything stochastic has somehow to be diffusing, therefore to be mathematically represented by a full, continuous, spatial dimension. A three-factor pricing problem would mean solving a three-space-dimensional PDE, etc. Whereas I contend that a regime-switching model with a few regimes, say three or four, where each regime is characterized by a different triplet (σ, λ, r) and the underlying share diffuses in each regime, can in theory handle the pricing of, say, convertible bonds under stochastic volatility, hazard rate and interest rate. And this is achieved at no extra computational cost other than solving a system of three or four coupled one-dimensional PDEs. You can see now why I am at a loss for words, trying to frame the *nature* of the regimes. Surely enough, the regime is our extra state variable, but I must refrain from calling it a "dimension" as it is able to embed *multiple* dimensions, or rather, multiple factors, and I hesitate to call the individual factors, σ, λ, r "dimensions", as they do not represent separate state variables that live independently in their own individual, continuous spaces. A multidimensional solid need not be the tensor product of the individual dimensions, after all. I am not even sure we can put a name on the regime. A regime is a state variable, surely enough. But what is it a state of? Volatility, hazard rate, interest rate? Or is it an abstract entity with no name, a sort

of container which may contain the name of volatility, of hazard rate and interest rate – or any other collections of names depending on the particular pricing problem – and will assign to these names the particular numbers that they get depending on the particular calibration to market? It seems we have found one more reason why our model should be called "Nobody", the model with no name.

It remains to deal with the general objection that the market may, as a matter of fact, not fall in any of the n-tuple regimes we are considering, all the more certainly so that our distinguished regimes now require that volatility should be the particular number that the particular regime says it is and, *simultaneously*, that the hazard rate should be the particular number, *and* the short rate the particular number, that the particular regime says they are. Even worse, the real pairwise correlation between these three factors may be such that no transition between any pair of the n-tuple regimes can reflect it. As before, we reply that the real volatility, the real hazard rate, the real interest rate, and the real correlation are not observable. All that really matters are the observable prices of traded instruments. And these prices are probabilistic averages over the regimes. It is up to us to calibrate the regime-switching model to the market prices of the volatility, credit risk, interest rate, and correlation instruments that we think are most representative. Maybe we should increase the number of regimes – always an open question that the relative ease of calibration can answer alone.

As the values that the parameters get inside the regimes and the intensities of transitions between the regimes are determined by calibration only, the hope is that the calibration procedure will settle on a certain solution – consequently our model on a certain specification – which will serve no other purpose, in the end, but to price other instruments relative to the initial ones. The general philosophical idea being here that, just as we could not put a *name* on the regime, even less so will we be able to *read* in the regime some value of volatility, or hazard rate, or interest rate, that we believe shall obtain in reality. This is another way of saying that the philosophical doctrine known as instrumentalism is perfectly acceptable as an alternative to metaphysical, or even semantic, realism.

8

Let me turn now to what is probably the deepest, and by far the most original, insight about the regime-switching representation. (Notice that I am no longer calling it "regime-switching model" as it is, I think, more general than a model.) The argument is rather subtle so please bear with me. Earlier I said that a regime could be characterized by an n-tuple rather than the single value of a single parameter, and I gave the combination of the hazard rate, volatility, and the interest rate, as an example. The underlying share followed a diffusion process in each one of the regimes anyway, so let us not worry about the underlying share for the moment; as a matter of fact, I can refer to Ayache *et al* (2004), where it is suggested that the process of the share *inside* a regime might as well be a full jump-diffusion, or that the share might undergo one, or indeed several, jumps in value, as it switched over between regimes. But let us leave it at that. Now suppose, for a moment, that the stochastic processes that we wish to worry about are not exactly the volatility process and the process of some other financial variable, such as the hazard rate. Suppose volatility is stochastic alright and the coefficients of its *own* stochastic process are stochastic too. Since writing is such an endless process, all I am proposing here is to take it one step further. Just as a diffusive, mean-reverting process was once written for the volatility of the underlying Brownian diffusion, and gave us the Heston model or the Hull

and White model of stochestic volatility, I propose today to write a further process for the *volatility of volatility* – why not another diffusion? – or indeed for the mean-reversion coefficient of volatility, or for the correlation coefficient between the underlying share and volatility, etc. More specifically, let our tentative model be:

$$\begin{aligned} dS &= rS\,dt + \sqrt{v}S\,dZ_1 \\ dv &= \kappa(\theta - v)\,dt + \varepsilon\sqrt{v}\,dZ_2 \\ d\varepsilon^2 &= \xi\,dt + \varphi\,dZ_3 \end{aligned} \tag{1}$$

In theory, the writing process should never stop, for this is the essence of trading and the result of submitting the theoretical model to the market (Ayache and Tudball, 2004). Just as the trading of the vanilla options turned the coefficient of the Black–Scholes formula, implied volatility, into a stochastic variable and created the need for stochastic volatility models such as Heston (or volatility smile models more generally), all we are noting here is that daily calibration of the given smile model, as well as trading the derivative instruments which are higher up in the hierarchy and specifically sensitive to the smile (for instance, the barrier options), will in turn create the need for a higher-level model such as the one we are proposing. Just as implied volatility once became a widely talked-about and a liquidly traded commodity, we are now talking of the next evolutionary stage where smiles become a commodity and are traded in turn. (Necessity of recalibration of the smile models and its counterpart, this open-endedness of the trading/writing process, are perhaps the greatest challenges facing any theory of smiles today. In my sense, they properly belong to the *metatheory* of smiles, or in other words, the philosophy of derivative pricing.)

Taking my cue from what I said earlier, namely that a three-factor pricing model need not be confused with a three-dimensional pricing problem, what I propose next is to apply the regime-switching idea to the third process. Instead of assuming a diffusion process for the volatility of volatility ε, why not consider a three-regime representation, $[\varepsilon_1, \varepsilon_2, \varepsilon_3]$, and appropriate transitions between the regimes? $[\varepsilon_1, \varepsilon_2, \varepsilon_3]$ will be our regime-switching model of stochastic volatility of volatility, and each individual regime ε_i will act as a super-container containing a full Heston model with a different volatility of volatility parameter. And now the regime-switching idea can be further invoked and the question asked how the Heston model inside the super-container can be itself replaced with a regime-switching model of stochastic volatility. Most probably the answer will be: put containers inside the super-container, with sub-regimes of volatility, $[\sigma_{i1}, \sigma_{i2}, \sigma_{i3}]$, now occurring inside each super-regime ε_i.

The problem is, ε was meant initially as the diffusion coefficient of the *Heston volatility process*. As soon as the Heston model is replaced by a *regime-switching model* inside the super-containers, ε becomes meaningless. The stuff inside the container blows up the super-container. To keep out of trouble, we should turn the problem on its head. We should really start with the Heston model, replace it with a regime-switching model of stochastic volatility, then find an appropriate meaning for the super-container supposed to make the latter model stochastic. As apparent from Ayache *et al* (2004), the regime-switching model of stochastic volatility $[\sigma_{i1}, \sigma_{i2}, \sigma_{i3}]$ is characterized by a bunch of parameters, (the individual values σ_{ij}, the intensities of transitions between sub-regimes j and the sizes of the simultaneous jumps of the underlying), and not just three like Heston. There is no diffusion coefficient, or mean-reversion coefficient, or long volatility coefficient in the regime-switching model of stochastic volatility,

at least not explicitly, but a conjunction of regime and regime-switching parameters which can only act *together* to reproduce any of these features. So the way to make our regime-switching model of stochastic volatility become stochastic in its turn is to assign different conjunctions of parameters to each super-container and not just a different single-valued parameter ε_i. Each one of the super-regimes of the model of stochastic volatility of volatility becomes, so to speak, a full regime-switching model of stochastic volatility.

Since transitions between super-regimes are modelled the same way as the transitions between sub-regimes, i.e. through Poisson processes of given intensity (possibly with simultaneous jump in the underlying), and given the associative character of the operation of grouping the sub-regimes into super-regimes, *our regime-switching super-model of stochastic volatility of volatility is in the end indistinguishable from a regime-switching model of stochastic volatility with a large number of regimes*. To fix the ideas, if we are talking about three super-regimes supposed to represent the stochastic character of the volatility of volatility and three sub-regimes, occurring inside each super-regime, supposed to represent the stochastic character of volatility, then the resulting construction will be indistinguishable from a regime-switching model of stochastic volatility, with nine regimes.[2]

9

And now we are ready for the last step of the argument. Remember that calibration to the market prices of derivative instruments is all that matters in our derivative pricing philosophy. It is the only key to unlocking the relative value of other instruments (and locking their hedging strategies; Ayache *et al*, 2004). There is nothing to tell us whether our model should have three, four or nine different volatility regimes other than the number and the variety of the instruments we are calibrating against, and our satisfaction with the calibration results. Given the large number of parameters implied by a nine-regime-switching model, chances are that a more parsimonious regime-switching model will fit the market prices just as well, and perhaps even exhibit more robust and more stable calibration behaviour. Perhaps a three-regime-switching model will do the job!

The most extreme form of the thought here is this: our volatility regime-switching representation, Nobody, is not just a model of stochastic volatility; *it is also a model of stochastic volatility of volatility, and a model of stochastic volatility of volatility of volatility, etc.* It contains at once the whole endless model-writing chain. Or rather, it is open like the writing process is open. So the question becomes: What could ever determine the particular hierarchical level at which the particular instance of Nobody shall land? Like I said, the answer lies in the particular nature and the particular prices of the derivative instruments we are calibrating against in the particular instance. Imagine that the option prices are not given by the market but artificially produced by a "first-level" stochastic volatility model such as Heston. Then our regime-switching model will match the corresponding vanilla smile and, for all practical purposes, mimic the behaviour of a first-level stochastic volatility model. Now suppose the instruments of concern are not plain vanilla but exotic structures, whose value, we know, is dependent on a level of complexity higher up than the given static smile, e.g. smile dynamics. For instance, we know from Ayache *et al* (2004) that first-level stochastic volatility models such as Heston may not simultaneously match the prices of the vanillas and the cliquets, for they imply a certain smile dynamics which may not accord with the cliquet prices. Higher-level models are called for in this case, with volatility dynamics more complex than Heston, for instance "universal volatility

models" or indeed models even more general, such as Nobody. Once Nobody is calibrated to the vanillas and the cliquets, it will behave like a higher-level stochastic volatility model. Finally imagine a situation where the instruments are very complex structures with no definite sense of the particular hierarchical level at which their sensitivity stops, and that their prices are just given by the market. Then hopefully Nobody will match those prices, and only the market "will know", in that case, at which level we landed.

Yet you complain:

> "Surely there must be something out there to help us distinguish between models of stochastic volatility and models of stochastic volatility of volatility (or at least, between models of significantly different hierarchical levels). The probability distributions must come out different, and there surely must come a stage where Nobody is ruled out *a priori*. There must be some probability distribution that gets generated at some level yet cannot be reached by the regime-switching representation, no matter the number of regimes or the value of the parameters. For how could a two-factor model, such as the particular instance of Nobody that we are considering (where volatility is the only identifier marking the regimes), ever reproduce the richness of a multi-factor model such as stochastic volatility of volatility of volatility ...?"

My reply to you is that you first try to realize what you are saying. Your perplexity relates back to the confusion we have already mentioned, between number of factors and number of dimensions. Recall the nine-regime-switching stochastic volatility model $[\sigma_{ij}]$ that we had obtained after unpacking the sub-regimes j of stochastic volatility occurring inside the super-regimes i of stochastic volatility of volatility. The double-index notation reflected the three-factor nature of the model, so everything looked OK. Are you now saying that as soon as associativity is invoked and we realize that all we have on our hands is a nine-regime-switching model $[\sigma_k]$, we lose the third factor? How could a change of notation, or in other words, a mere *change of name*, have such deep consequences?

Underlying your worry about Nobody failing to account for a multi-factor situation is in fact just a worry about different *names* that a thing can be called. And we had warned you that Nobody was precisely the model with no name, and the regime precisely the state variable with no particular label and no particular dimension attached to it! To put it differently, a regime-switching model which has the name of a *single* factor – for instance volatility – written on each regime, is not necessarily a one-factor, or a two-factor, or a three-factor model of stochastic volatility. It can be anything.[3] What it is really will largely depend on the "depth" and the variety of the derivative instruments we are calibrating against. Let me put it this way. If Nobody calibrates successfully to prices of vanillas, cliquets, as well as higher-order structures which may be sensitive to the volatility of volatility of volatility, and none of the other models, commonly known as one-factor or two-factor, does, then Nobody may be said to be of a "level higher than two-factor". Pushing the argument a little further, we may even wonder: Why should what we have to say about Nobody depend on what other models can do? For all we know, the other models may have never existed. Is it not our purpose to break with tradition anyway? All that we can hope to say, then, is that Nobody *was* able to calibrate to the prices of certain derivatives instruments traded in the market, period.

10

And by the way, isn't the whole language of "number of factors" just a heritage of the tradition? At least this much is certain: the regime-switching representation is not reducible to a classification in terms of number of factors. This, we have shown with two complementary arguments: (a) the argument that the regime can be identified by an n-tuple of names relating to different financial categories (volatility, hazard rate, interest rate); and (b) the argument that it can alternatively be identified by a single name (for instance, volatility) yet the picture be richer than the traditional two-factor framework. There is no question that talking of multiple factors is legitimate in the first case. When we contemplate stochastic volatility, stochastic hazard rate and stochastic interest rate, the situation is different from the one where the second process concerns the volatility of the first, and the third the volatility of the second. Typically, derivative instruments can have radically different underlyings in the first case. The underlying that a credit default swap is written on (the state of default or no default of a certain issuer) is very different from the underlying of an equity option, which is in turn very different from the underlying of an interest rate option. Derivative instruments such as the vanilla equity option and the cap and floored cliquet, by contrast, are written on the *same* underlying, although we did refer to them, a paragraph back, as instruments of different "depth". Even a variance swap is written on the same underlying as the vanilla option. The only difference is that its payoff explicitly depends on a variable, realized variance, which can only be measured over the whole path of the underlying.

The point of this distinction is to suggest that talking of "multiple factors" in the second case might be misleading and might have imposed itself on us for no other reason than the written tradition and the tradition of *writing* – for no other reason, indeed, than that a model, say, of stochastic volatility of volatility, has traditionally been *written* in three lines, as shown above. The volatility process has a diffusion coefficient ε and this coefficient diffuses in turn, etc. A new line is written every time and this suggests that a new kind of derivative instrument, specific to the new kind of factor, can be written every time. Vanilla option writing is specific to stochastic underlying; variance swap writing is specific to stochastic volatility, etc.

When you think about it, however, all that is really meant by the three lines written above – or any additional number of lines for that matter – is jut a complex stochastic process the underlying is supposed to be following, and the corresponding complex probability distribution. Again, compare the situation where some truly different processes are involved: a hazard rate process, an interest rate process.

The other real things are the derivative instruments. True, they may be differently styled, and may admit of different levels of complexity, but they all fall back, in the end, on the underlying they are written on. Our writer has created himself a fiction (that the Brownian diffusion coefficient might be diffusing) and he is now growing a new fiction inside the fiction (that the diffusion coefficient of the coefficient of the Brownian diffusion might itself be diffusing, etc.). Think, by contrast, how the un-writable, un-nameable, dimension-less, story of regimes manages to describe the world just as well, or even describe it better (calibrate to the traded prices of derivatives, propose optimal hedging strategies), yet offers no predetermined format, no predetermined number of lines, to fit the story in.

Another way of looking at things, and seeing how our writer entraps himself in his own fiction, is to compare the *a priori* attitudes of the traditional representation and the regime-switching representation, when both are brought face-to-face with the market. By committing

himself to two or three lines of writing, the traditional writer faces a market which can be completed *a priori* with the help of two or three traded instruments. Markets can be completed *a priori*, under Heston or any other "first-level" stochastic volatility model, by trading an option together with the underlying as a dynamic multi-hedging strategy, and they can be completed *a priori*, under "second-level" models such as the one written above, by adding one further instrument, etc. Our writer faces a market that he knows he can complete *a priori* with the help of a given number of instruments, regardless of the variety, the depth, and the price structure of the market he will face in effect. The regime-switching representation, by contrast, does not impose such strictures on the future story. It cannot tell, in advance, the degree of incompleteness of the market. It cannot tell *a priori* at which particular "writing level" it will fall. Only the market and the result of calibration can. Philippe Henrotte, the head of theory in my company and the father of Nobody, summarized the point beautifully: "The reason why the regime-switching representation cannot be completed *a priori* is precisely that the regimes bear no particular name!" As Jacques Derrida, the leading figure of French theory, would put it, we've been held captive by a long tradition of logocentrism and the presence of the name. Indeed the noticeable consequence of naming is to make present and actual for us what could very well be different and have to be *deferred*, what could only *be* later.

11

The point is of importance because the real test of a smile model is to see how robust it is to a reality that may contradict its assumptions. (This is the whole point of the *metatheory* of smiles.) Complete markets are the least robust notion. You perturb them a little bit and they become incomplete. Surely enough, Nobody will fall on what we have called a "particular writing level" once it is calibrated; the market it describes will assume a certain degree of incompleteness, and it will be susceptible of completion by using the appropriate number of hedging instruments in the dynamic hedging strategies. The point is that such degree and such level are dictated by the reality of the instruments we calibrated against. They are determined *a posteriori*, not *a priori*. Should a new instrument become traded the next day, and its price fall outside the range of prices that were attainable by the completed market of the day before, then new calibration to that new instrument will open new levels to Nobody, and new degrees of incompleteness. By contrast, you cannot but throw away your Heston, when such a situation occurs. This also tells us that we should always leave the door open, in Nobody, for such new possibilities and such self-upgradings. We should not strive to complete the market, at any level of writing that we may be standing. To keep our hedging robust, we should always keep it optimal, and never try to make it perfect. HERO is the measure of residual risk borne by the optimal hedging strategy.[4] What is interesting is how HERO decreases when additional instruments are used in the hedging portfolio, or in other words, the hedging opportunity that the additional instruments may offer. But HERO should never be driven down to zero. What is interesting is the way we approach completeness, not completeness as such. For surely we lose the sense and the measure of all that when HERO is equal to zero!

What we are really saying is that we might be offered a chance, with the regime-switching representation, which was not available in any smile model before: the invaluable possibility that Nobody might just be equipped to deal with *model risk* on top of the risk which is the normal, *contained* subject matter of the models of risk. Somehow, Nobody might be "aware" of its own metatheory. Recall that Nobody is not just a first order stochastic volatility model, but

can potentially instantiate any of the higher order models corresponding to the higher levels of writing. This level-invariance is due to the fact that a regime-switching model *made stochastic*, and consequently calling for "regimes of regimes," is in the end just another regime-switching model. The regime-switching representation does not iterate or pile up in the same fashion as the traditional lines of writing. Only the variety of the derivative instruments we are calibrating against and the richness of their price structure can set the level of writing for us. For instance, a newly traded instrument and a new range of prices can precipitate an "internal" transition from a certain level to a higher level. Our stochastic volatility model suddenly becomes a stochastic volatility of volatility model, yet with no visible change. We will still be looking at the same old regime-switching representation, only we will be calling it different names. As a matter of fact, this can work both ways. Why should the change be invisible in the one way, and not in the other? What is to stop us indeed from thinking that the stochastic volatility of volatility model was already available to us the day before? True, we may have not calibrated to the *relevant* instruments the day before, or the relevant instruments may have not been available the day before, but is this reason not to think that the *thought* was available to us the day before?

Put differently, Nobody may have already opened itself to the self-upgrading the day before, only the visible derivative instrument, specifically crystallizing the upgrading, was simply not available the day before! Maybe the upgrading was *in part* contained in the set of derivative instruments of the day before, and was just missing one last instrument to express itself in full. Things may have already been "on the cards", as the saying goes, or in other words, the market may have already anticipated its own upgrading the day before. Or we may argue the other way round. Although the new instrument takes us up one level, there might be no reason to believe that its introduction will automatically bring a mutation in the market. At least not the first day. Chances are, on the contrary, that the newly created instrument – I am thinking, for instance, of a new complex structure, something like a complex cap and floored cliquet, which, although written on the same underlying, is sensitive to higher order distortions, such as volatility of volatility of volatility – chances are that it will begin its journey right on the tracks from the day before, as the market participants will no doubt start pricing it with state-of-the-art models not yet aware of the next level. It is only when the instrument comes alive and starts leading a trading life of its own that the true change in the market will begin.

I guess my whole point is that changes of "writing levels", or degrees of incompleteness, can take place smoothly within the regime-switching representation, and can only cause breaks and fractures within the *writing tradition*. Nobody can thus enjoy continuity of life and allow us, for the first time, to really address the question of *history*. Indeed a big question, perhaps the biggest, is whether the given smile model should be calibrated to the instant prices of derivative instruments or to their history. This is the story of re-calibration looming again. People are sooner or later led to back-test their model and they become very excited when they notice that the parameters of the model are stable over successive re-calibrations. They think they have hit upon some deeply significant invariant. Whether they admit it or not, everybody is striving towards this end, and calibration to the history of prices is one way of making it explicit. The main objection, however, is that any time series of any given length will end up revealing some invariant or other when submitted to a model of sufficient complexity. This is another way of objecting that the "true" data generating process may in fact admit of no finite moments and may require time series of infinite length before any parameter is stably estimated. As a matter of fact, being able to *name* an invariant may be the worst thing that could ever

happen to the searchers engaged in that kind of quest. For what is then to stop us from writing an extra line, and turning the name of the invariant into the name of a new stochastic process?

12

"Model risk", "necessity of re-calibration", "endlessness of the writing process", are all different names for the same big problem. Perhaps *the* big foundational problem which puts into question the very possibility of quantitative finance as a science (Ayache in press; Ayache and Tudball, 2004). We all know that the option pricing tools that we are using are forward-looking. Precisely for that reason, we all know that they are subject to the necessity of recalibration and to the threat of model change, or in other words, model risk. Yet we lack the means to address that problem almost by definition. A model can do anything except look into its own assumptions. ("What by definition can hurt you is what you expect the least"; Taleb (2001).) This is why we almost inevitably turn to history as the only way out of our predicament. That we may not have the faintest idea how to address the problem of re-calibration makes historical calibration look useful, to say the least, in comparison. Although a backward-looking procedure can be the last thing we want to consider when addressing a forward-looking question, it is unfortunately the only thing available. We have no choice but to entertain the momentary hope that history might repeat itself.

Now think again of the essentially nameless character of our regime-switching representation. Since our regimes bear no particular name (the name of volatility, or volatility of volatility, etc.), the temptation is simply *not there* to look back at the recent history of calibration of Nobody, and try to identify a stable parameter. First of all, it is not even clear what level of writing is getting instantiated everyday! Alternatively, we can look back at the series and read into it any story we want. We can argue, for instance, that the full richness and the full incompleteness of the market were right there with us from the start – just as Nobody was with us from the start! – only the hidden variables became manifest, and Nobody was able to calibrate to them, from that day when the relevant complex instrument became alive and started leading a trading life of its own. So I guess the second part of the recommendation never to complete the market and always to leave the door open for future upgrading is the recommendation never to trust that Nobody *has been perfectly calibrated to the market*. "Calibration is just another word for completion," as Philippe Henrotte says, for an instrument whose price process we could not attain with a suitable self-financing strategy involving the existing calibration instruments would mean that a new level is up and that the new instrument adds richness to our existing information set. Therefore it should independently be included in the calibration procedure, and our past calibration was not right.

Whatever interpretation it may be that we care to put on Nobody's recent history of calibration, the fact remains ... that only the facts remain. Whether the richness was all there from the start and Nobody was simply not "perfectly calibrated to the market" or whether the richness all emerged when the new instrument first diverged from the known tracks and opened the door for an actual upgrading of Nobody, is in the end a purely nominalistic issue and just a question of how we want to *call* the story. The absence of writing inherent in Nobody relieves us completely of the necessity either to read the past into the future or to read the future into the past. Only the fact of recalibration remains and the hope, like I said, is that the HERO shall pick up a component of model risk (or meta-risk) on top of the standard risk that it is picking up at the object level, thus allowing smoothness of upgrading. Although theoretically

unjustified – for the computation of HERO presupposes that the level of writing is fixed and the model is final – this is just the hope that successive recalibrations and successive rehedging operations shall prove robust *as a matter of fact*. Since a "stochasticized" Nobody is just an instance of Nobody, the hope is that the two instances shall bear a few similarities other than just by "name"!

In the end, the reason why I think that Nobody may just be offering us a chance to *finally* address the question of history, is that this is the question whose answer we should expect to be the least naïve of all, or in other words, the least *expected* of all (for any answer falling within our range of expectations will become history, therefore will be overtaken by history), and that Nobody, whose inherent non-writing is essentially a non-answer, seems to offer just the right kind of divergence and the right kind of ... digression. By remaining open to the future and by coming out virtually unchanged through the future upgradings, Nobody in fact *postpones* any temptation we may have to look back at the past in order to figure out the future. (As I said, the temptation may be due to a lack of choice.) Nobody takes over the task that we thought was reserved for history, the task of softening up the future for us, and teaching us patience in matters strictly relating to the future (as if history was the only thing we could read in patience while waiting for the future). As it is *both* forward-looking and capable of self-upgrading, Nobody allows us at last to deal with the future *seriously* (when we thought history alone could achieve that purpose). In fact, Nobody reinstates the balance of power in favour of the *present* (rather than the past or the future). Nobody is the perfect tool in the hands of the trader (as I shall argue in a philosophical column in *Wilmott* magazine) and the pair that the trader and Nobody will thus constitute is, in the last instance, essentially *present* (like the living trader is present) and utterly self-upgrading (or geared to the future).

13

So far, my strategy has consisted in arguing that default lent itself naturally to the discrete regime-switching representation on account of the two-valued nature of the state variable (default / no default) and that the regimes could be further extended to the no default side of the picture – naturally so as concerns the hazard rate (as different regimes got interpreted as different ratings), and not so naturally as concerns volatility (as people traditionally had in mind continuous diffusion processes of volatility). As a matter of fact, the second part of my argument very quickly developed into a formidable digression on the regime-switching representation, whose main purpose was to establish the originality, and I dare say, the *uniqueness* of the proposed solution. Beside the advantages specific to the regime-switching representation (its economy, its computational efficiency), and beyond the equity-to-credit problem, I have attempted to show that the regimes might in fact provide just about everything everybody ever wanted in a smile model and could not get before: robustness of calibration and its correlate, the absence of presupposition with regard to the degree of incompleteness of the market (and adaptation, instead, to the effective degree of incompleteness through effective calibration), as well as the beginning of an answer to the abysmal question of recalibration. Incidentally, the argument earned us a criticism of the traditional representation, the tradition of *writing*, and debunked its self-created myths and mythicized names. Philosophically, this meant that our regime-switching representation, Nobody, had seized mastery of the subject and taken precedence over the other models, as it was able both to propose a working solution and to take an unprecedented metatheoretical stand.

It remains, however, to clear one last obstacle before moving on to the problem at hand, the equity-to-credit problem. I would like to call this obstacle the "continuous–discrete fallacy". In a word, this is the general worry that the regime-switching representation might just not be suitable for derivative pricing *because* of its discrete character. All is well when the variables assume two, or perhaps only a few, discrete values (such as credit ratings or the default state), but is a regime discretization of the credit spread able, for instance, to handle the pricing of options *on the credit spread*? Is a three-regime-switching model of stochastic volatility able to price volatility swaps? Also, how could such a formidable campaign ever be launched in the field of smiles, and such a systematic attack ever be mounted against the traditional "continuous" smile models, from such a frail and discrete base? Isn't the regime-switching model "negligible", and almost degenerate, in the space of the smile models?

The proximate answer is that the convertible bond is an equity derivative, not a credit spread derivative, and the equity is still getting modelled as a continuous diffusion process, spanning all values from zero to infinity, inside each one of our regimes. As shown by our numerical examples (see Appendix), recognizing two or three hazard rate regimes (actually four, if we count the default regime) is amply sufficient to capture the impact of default on the pricing and hedging of the convertible, or indeed to fully explain the term-structure of credit spreads. Two or three volatility regimes are also sufficient to capture volatility risk and the relevance of vega hedging. Actually, a three-regime stochastic volatility model or a three-regime stochastic hazard rate model are much richer than you think. The intensities of transitions between regimes participate fully in the specification of the model. Also, let us not forget the whole orientation of my essay. My main topic is the equity smile, not credit risk, and my task is the continuation of the work achieved in Ayache *et al* (2004). As I said, I am willing to consider default only to the extent that it is a component of the equity smile and that the credit default swap prices can help us calibrate and hedge our equity smiles better.

The ultimate answer, on the other hand, is that the regime-switching model is not negligible after all. Nothing stops us in theory (or in practice) from multiplying the number of hazard rate or volatility regimes up to the point where they become indistinguishable from the discretization of a "real", continuous space. Nothing stops us from assuming as many regimes (σ_i, λ_j) as there are pairs in the tensor product of the full (discretized) continuous volatility space and the full (discretized) continuous hazard rate space. Solving a diffusion PDE for the underlying equity in each one of those regimes, and coupling this incredibly large number of one-dimensional PDEs through the usual transitions between regimes, will then turn out to be numerically indistinguishable from discretizing a full three-dimensional PDE in the (S, σ, λ) variables, and solving it with the usual techniques. As a matter of fact, the argument can work both ways. For the equity diffusion process that we had assumed was taking place, so far, in each one of our parsimonious regimes, and its numerical treatment by the usual PDE discretization techniques, can now in turn be interpreted within the regime-switching representation. Discretized Brownian motion is just a regime-switching model with a large number of tinily spaced regimes. Numerically speaking, it is all then but a massive regime-switching operation. All I am saying here is that, ultimately, everything becomes discrete (not to mention that everything, actually, is *initially* discrete, as stocks trade by the tick and hedging takes place in discrete time).

In conclusion, the regime-switching representation is the one and all-pervasive representation. Continuous path processes and continuous time finance are but useful fictions that were invented to *summarize* the incredibly disarticulate picture left over by the regimes. The name

of "volatility" (Black–Scholes), and following it, the name of "stochastic volatility" (Heston), etc., are *given names* that merely indicate that the wiring between the regimes has been laid out in a certain way and not another. Two or three names, two or three coefficients (volatility, mean-reversion, correlation, volatility of volatility, etc.), throw order and rigidity on an incredibly rich and multifarious picture.

This immediately poses the following question: Among the two extreme situations we have described, which one is in fact the poorest? Is it the situation where we have only a few volatility and hazard rate regimes and have no particular names attached to the model, or the situation where we have infinitely many regimes but only a couple of names? Might not the order in question be a restraining order? The main objection against Nobody was that the discrete regime representation might be missing something important that the continuous representation could provide. Shouldn't we be worrying about the opposite by now? Shouldn't we be worrying that the continuous representation might be giving us much more than we actually need, and charging us a very high price for it – imposing names on us that we might not even need? Let us not forget indeed that the tradition of writing and naming creates strata that are impermeable to each other (volatility, volatility of volatility, etc.), when the nameless regimes manage at once to *see through* the whole thing. As the numerical fate common to both representations demonstrates their equivalence in the limit, and shows that it is all but a question of representation, perhaps we could now step back from the limit and try to get the best of both worlds. Not wanting an infinite number of regimes and not wanting the names either, perhaps we could find, in the midfield, the compromise which is best adapted to each particular situation. At least we want the freedom to do so.

Again, recall the freedom to be in any of the possible "writing levels", that Nobody is allowed by the absence of names. And imagine, for instance, a situation where calibration to a newly introduced derivative instrument has exhausted all the possibilities offered by three regimes of volatility and still could not be achieved to our satisfaction. This is typically the situation where we contemplate adding a regime. Now adding a fourth regime achieves much more in effect than adding a fourth state of volatility. It opens for us whole new levels of writing and new degrees of incompleteness not previously available. On the other hand, it can have richer consequences than adding a line of writing as in the classical writing tradition. The writing levels of the classical tradition are stratified. Below the name of "volatility of volatility of volatility" there just lies the immutable name of "volatility of volatility". Whereas the three regimes that lie "below" the fourth can now react to calibration in ways we could not even dream of when all we had was three regimes. Actually, the three regimes are not *below* the fourth, but at the same level. There is a sense of wholeness, of richness, and an overall feeling of economy in the regime-switching representation that is unavailable to the stratified writing tradition.

In the end, I do not exactly place the interesting debate between the *discrete* and the *continuous*. This is a wrong divide, and the two are in the end equivalent. Rather, our "digression" into the regime-switching representation will have served the following philosophical purpose (beside providing us with an extraordinary calibration, pricing and hedging tool). It has shown us that the real difference is between worshipping the name and breaking the name, between iconolatry and iconoclasm, if you will.

I have nothing against the Black–Scholes option pricing model, or the Heston option pricing model. Numerically, that is to say, ultimately, they are but instances of Nobody. I guess what I have against you is that you may be tempted to build up your world like a venerable writer,

not like an engineer. You may be tempted to recognize in Black–Scholes or in Heston nothing but the names and the strata that will allow you to append your own model and write your own name.

Analytical pricing formulae are very hard to come by nowadays because of the complexity of the derivative instruments and the complexity of the underlying processes. Computational power, on the other hand, is allowing numerical speed-ups which compare with the analytical formulae of yesterday. So what could be the point of insisting that the model should be cast in terms of elegant continuous processes – so remote indeed from our nameless, shameless, mess of a regime-switching representation! – other than the facility of writing the model *on paper*, and producing a nice mathematical paper? True, continuous path Brownian motion is what afforded the perfect replication in Black–Scholes and created the myth of the complete markets. But the complete markets are precisely the thin leaf that we should try to escape at all costs, or only admire from a distance as an elegant argument written on a nice piece of paper! Complete markets are enmeshed with the guilty "tradition of the name" anyway. Options are redundant in the Black–Scholes world; they do not truly exist and are only the diminutive *name* of a particular dynamic trading strategy involving the underlying alone.

Derivative pricing science has been taken hostage by mathematicians when it should be handed back to the financial theorists – how many working solutions mention incomplete markets and real hedging? – and to the engineers. But how come, you may wonder, a philosophical argument as general as this – the argument against the naming and the writing – happens to occur precisely in the field of finance, and more particularly so, the field of derivative pricing? Doesn't the same massive attack equally apply in other engineering fields? What is so specific about derivative pricing that allows me today to draw such deep conclusions about the fallacy of naming and the lure of writing?

The first answer is that derivative pricing is only *starting* to become a proper engineering field. Inelegant and massively computational solutions have been occurring for some time now in fluid dynamics, or structural mechanics, or thermodynamics. The second and most important answer is that we enjoy a freedom in our specific engineering field which is hard to find in other fields. Laws of gravity, laws of mechanics, and more generally laws of nature, compel an Einstein, a Schrödinger or a Navier and Stokes to *write* the models they have written. (Yet I am not so sure that an antirealist will not argue that laws of nature *are not* written in nature after all, and that the physical theories that we have, and their quantitative models, are mere computational tools.) We, by contrast, can "write" or "name" or "program" the model that we want, so long as the model is robustly calibrated and some derivative instruments are robustly priced and hedged relative to some others. (The searcher of the ultimate data generating process can wait all the time he wants, even wait infinitely – but then he *is not* in the business of derivative pricing.) We have all this freedom, yet some writers insist on following the inherited path of research, and the inherited lines of writing! Path-dependency is an even worse case than inhomogeneity ...

Above all, I think the real strong argument for wanting to erase the previous lines of writing and making the fresh start that I suggest, is that the *open* regime-switching representation can address the problem of *co-calibration* just as easily as it addressed the problem of calibration or re-calibration. Take the equity-to-credit problem, for instance (this way, I can circle back to my original topic). A lot of effort has been spent recently in developing credit risk models quite separately from the smile models. While jump-diffusion, local volatility, stochastic volatility or universal volatility have been suggested on the one side, stochastic interest rates,

stochastic hazard rates and stochastic recovery rates have been suggested on the other. My whole argument about the level-invariance of Nobody with respect to the volatility factor can of course be reiterated with respect to the hazard rate factor. Nobody can scale up to any degree of incompleteness that the market of credit derivatives may impose in effect (stochastic hazard rate, stochastic volatility of hazard rate, stochastic volatility of volatility of hazard rate, etc.), just as it scaled up in the volatility case. More interesting, perhaps, is the fact that Nobody can co-calibrate to credit instruments and volatility instruments *with no visible change*. Wouldn't we have to wait, otherwise, for two *distinct* traditions of writers to deliver to us their successive lines of writing?

This isn't just going over the fact that the regimes of our regime-switching representation can be identified by a pair (σ, λ), for I am now implying a deeper phenomenon, namely, that the volatility skew implied in the market prices of out-of-the-money puts can positively give us information on the default process and, reciprocally, that the CDS term-structure of spreads can positively give us information on the value of equity options. The equity-to-credit problem is precisely a smile problem. As a matter of fact, it may even be better-posed than the pure smile problem, as the information from the CDS will have a tendency to help the calibration, and help determine the solution (see Appendix). Provided, of course, our smile model and calibration tool can *handle* co-calibration, and make it a friend, not an enemy. Nobody *loves* the idea that CDS prices can be included in the calibration set, when other (stratified) models most certainly resent it! What is a complication in the traditional representation is a simplification in ours. This is exactly the correlate, in co-calibration, of the idea we have already explored in calibration and re-calibration, according to which Nobody loves the prospect of adding barrier options, or cliquets, or more complex structures still, in the calibration phase.

Appendix

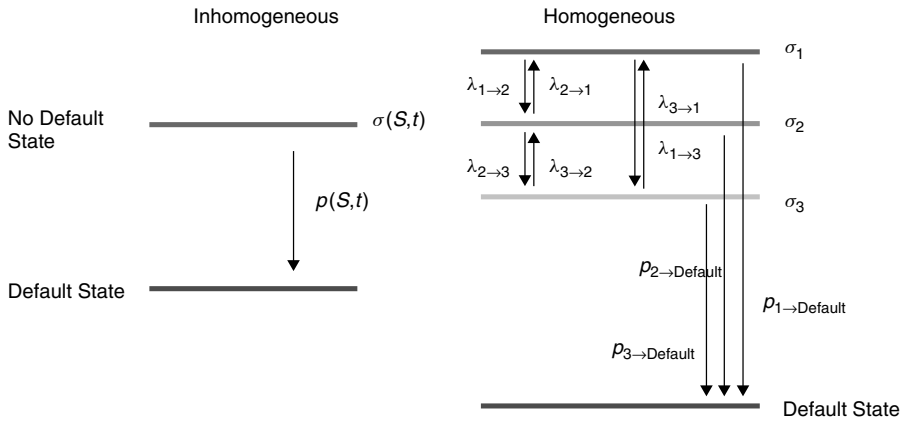


Figure 1: Comparative logic of the inhomogeneous and homogeneous equity-to-credit models. The inhomogeneous model consists of one default regime and one non-default regime. The implied volatility surface and the credit smile surface are explained by a local volatility surface and a local hazard rate surface. In the homogeneous model, volatility and hazard rate are stochastic and switch between the three non-default regimes

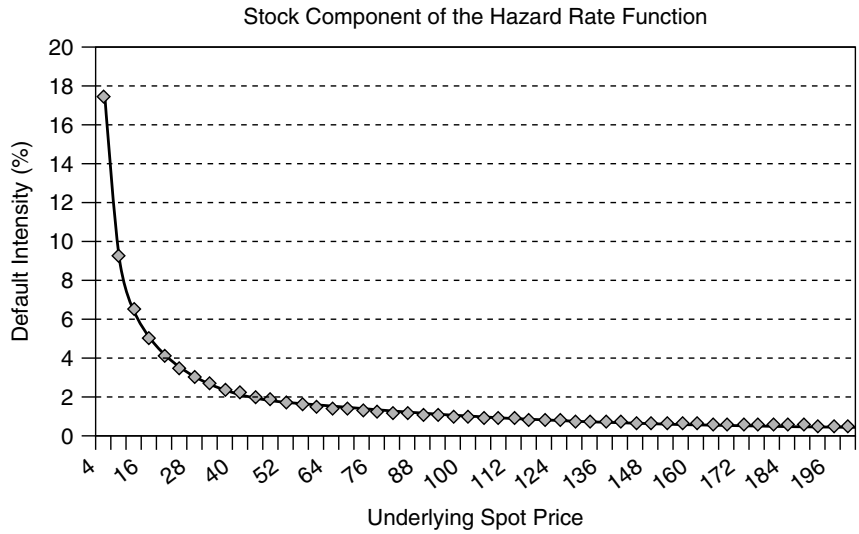


Figure 2: In our inhomogeneous model, the hazard rate function is the sum of a time component $g(t)$ and a space component $f(S)$. The space component accounts for the dynamics of the credit spread curve against the moving underlying stock. The time component ensures that a given credit spread term structure is matched for a given stock level. Above is the plot of the space component, $f(S) = p_0 \left(\frac{S_0}{S}\right)^{\beta}$, $S_0 = \$16$, $p_0 = 5\%$, $\beta = 0.9$

TABLE 1: TERMS OF A CONVERTIBLE BOND (THE TYCO 2.75% 2018 TO THE FIRST PUT DATE)

Maturity date	15/01/2008
Semi-annual coupon rate	2.75%
Nominal	100
Conversion ratio	4.38
Recovery rate	0

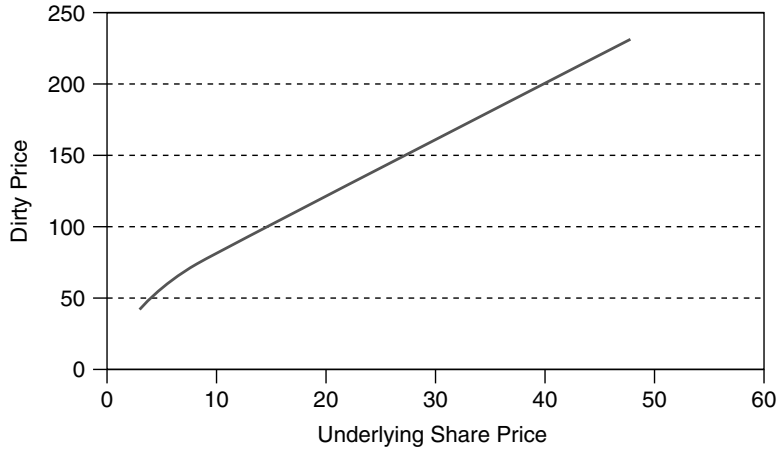


Figure 3: Theoretical value of the convertible bond described in Table 1 against the underlying equity in a inhomogeneous equity-to-credit model. The pricing date is 3 February 2003. The interest rate is flat 3%. The credit spread term structure is given in Table 5 when the underlying stock is $S = \$16$. The Brownian volatility is 45.3%. The CB is worth \$105.50 against $S = \$16$, its Delta is 3.80 shares, its fixed income component is worth \$81.75

TABLE 2: CONVERTIBLE BOND IN THE INHOMOGENEOUS MODEL. WE REPORT ITS DELTA AGAINST STOCK LEVELS UNDER THE EQUITY-TO-CREDIT MODEL WHERE THE HAZARD RATE FUNCTION IS GIVEN IN FIGURE 2. THE EQUITY-TO-CREDIT DELTA INCREASES ON THE WAY DOWN AS THE BOND FLOOR COLLAPSES. WE ALSO REPORT THE DELTA UNDER STATIC SPREAD FOR COMPARISON

Stock price	Equity-to-credit delta	Delta under static spread
16	3.81	2.91
14	3.86	2.66
12	3.96	2.37
10	4.16	2.01
8	4.58	1.59
6	5.38	1.10
4	7.33	0.57
2	10.80	0.13

TABLE 3: CALIBRATED PARAMETERS OF THE HOMOGENEOUS REGIME-SWITCHING STOCHASTIC VOLATILITY AND STOCHASTIC HAZARD RATE MODEL, NOBODY (TWO NON-DEFAULT REGIMES AND ONE DEFAULT REGIME). NOBODY IS CALIBRATED TO THE FULL IMPLIED VOLATILITY SURFACE GIVEN IN TABLE 4 AND THE FULL CREDIT DEFAULT SWAP SPREAD TERM STRUCTURE GIVEN IN TABLE 5. THE SOURCE OF MARKET DATA IS TYCO ON 3 FEBRUARY 2003 AND THE UNDERLYING STOCK IS $S = \$16$

	Brownian diffusion (%)	Total volatility (%)
Regime 1	49.86	61.18
Regime 2	27.54	40.83
	Jump size (%)	Jump intensity
Regime 1 → Regime 2	4.48	3.3429
Regime 2 → Regime 1	−58.68	0.1697
Regime 1 → Default Regime	−100.00	0.1190
Regime 2 → Default Regime	−100.00	0.0324

TABLE 4: QUALITY OF FIT OF A FULL IMPLIED VOLATILITY SURFACE WITH THE EQUITY-TO-CREDIT HOMOGENEOUS REGIME-SWITCHING MODEL. SOURCE: TYCO ON 3 FEBRUARY 2003. THE UNDERLYING STOCK IS $S = \$16$

Maturity		Strike (%)											
(years)		5	7.50	10	12.50	15	17.50	20	22.50	25	30	35	45
21/02/2003	Market			158.10	112.70	79.80	58.10	49.40	56.30	72.40			
	model			175.99	122.30	76.07	55.80	48.39	48.09				
21/03/2003	Market			122.20	92.90	71.20	56.00	48.40	45.40	53.10			
	model			129.28	93.11	66.24	54.82	49.39	49.13	47.17			
17/04/2003	Market	138.20	108	82.80	66.30	54.60	47.20	45.40	45.70	53.20	65.30	78	
	model	150	112.07	83.38	63.62	54.07	49.06	48.28	46.91	45.78	44.90	43.50	
18/07/2003	Market	99.10	87.30	72.60	60.50	52.10	47.00	44.70	43.60	44.30			
	model	113.80	88.77	71.43	59.45	52.22	47.78	46.14	44.72	43.87			
16/01/2004	Market	92.40	75.40	64.40	56.50	51.40	47.10	45.00		42.80	43.20	45.20	
	model	89.87	73.99	63.47	55.90	50.68	46.98	44.94		42.07	40.91	40.49	
21/01/2005	Market	73.70		58.60		49.40		46		42.70	41.10	41.10	
	model	74.50		58.34		50.40		45.88		43.02	41.31	40.19	

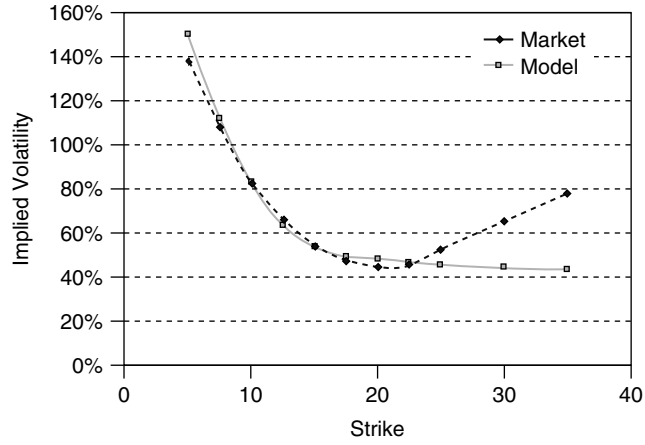


Figure 4: Zooming on the quality of fit of the implied volatility smile of options maturing on 17 April 2003.

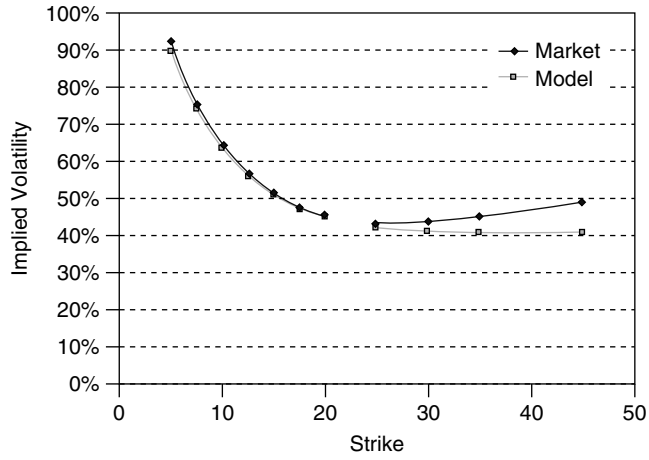


Figure 5: Zooming on the quality of fit of the implied volatility smile of options maturing on 16 April 2004. The model does not do such a good job on the out-of-the-money calls, in both Figure 4 and Figure 5, because of the crazy prices reported in the market. Probably those calls are not so liquid

TABLE 5: QUALITY OF FIT OF THE CREDIT DEFAULT SWAP TERM STRUCTURE WITH THE EQUITY-TO-CREDIT HOMOGENEOUS REGIME-SWITCHING MODEL. SOURCE: TYCO ON 3 FEBRUARY 2003. THE UNDERLYING STOCK IS $S = \$16$

Maturity (years)	Market premium (quarterly paid coupon) (%)	Model premium (%)
1	1.25	1.50
2	1.17	1.22
3	1.14	1.13
4	1.11	1.08
5	1.09	1.05
6	1.05	1.03
7	1.03	1.02
8	1.01	1.01
9	0.99	1.00
10	0.98	0.99

TABLE 6: THE TYCO CONVERTIBLE BOND IS DYNAMICALLY OPTIMALLY HEDGED, IN THE HOMOGENEOUS MODEL, *WITH THE UNDERLYING STOCK ALONE*. THE SIMULATION TAKES PLACE ON 3 FEBRUARY 2003. IT USES NOBODY WITH THE PARAMETERS INFERRED FROM CALIBRATION GIVEN IN TABLE 3. THE CB IS WORTH \$105.50 AGAINST $S = \$16$ (SAME REFERENCE POINT AS IN THE INHOMOGENEOUS MODEL). WE REPORT THE HERO, EXPRESSED IN DOLLARS, AND THE OPTIMAL DYNAMIC HEDGING RATIO. THE HEDGING RATIO IS EXPRESSED AS THE EQUIVALENT STOCK POSITION IN NUMBER OF SHARES (YOU SHOULD SELL THE SHARES IN ORDER TO HEDGE). NOTICE THAT THE HEDGING RATIO FOR $S = \$16$ IS 3.79 SHARES, VERY CLOSE TO THE DELTA IN THE INHOMOGENEOUS MODEL AGAINST THE SAME STOCK LEVEL (SEE TABLE 2). OPTIMAL HEDGING UNDER DEFAULT RISK IS TO OUR MIND THE REAL REASON WHY YOU SHOULD GO HEAVY ON THE DELTA, NOT SOME DETERMINISTIC FUNCTION LINKING THE HAZARD RATE AND THE STOCK

Stock Price	Optimal stock hedge ratio	HERO
16	3.79	9.62
14	3.70	10.25
12	3.60	10.93
10	3.52	11.68
8	3.50	12.48
6	3.71	13.29
4	4.64	14.04
2	8.64	14.51

TABLE 7: THE TYCO CONVERTIBLE BOND IS DYNAMICALLY OPTIMALLY HEDGED IN THE HOMOGENEOUS MODEL WITH A *COMBINATION OF THE CREDIT DEFAULT SWAP AND THE UNDERLYING STOCK*. WE USE THE 5-YEAR MATURITY CDS WHOSE PREMIUM IS REPORTED IN TABLE 5. WE REPORT THE RESIDUAL HERO AND THE DYNAMIC HEDGING RATIOS. THE CDS HEDGING RATIO IS EXPRESSED AS EQUIVALENT CDS POSITION (YOU SHOULD BUY THE CDS AND SHORT THE STOCK TO ACHIEVE THE OPTIMAL HEDGE). THE CDS HEDGE IS IN PERCENTAGE OF NOMINAL. AS THE CDS TAKES CARE OF THE MAJOR JUMP DUE TO DEFAULT RISK, THE STOCK HEDGES AGAINST THE DIFFUSION AND OTHER SMALL JUMPS. ITS CONTRIBUTION IN THE HEDGE IS VERY SIMILAR TO THE DELTA UNDER STATIC SPREAD (SEE TABLE 2)

Stock price	CDS hedge ratio	HERO	Stock hedge ratio
16	−55.9	5.00	3.05
14	−59.7	5.10	2.80
12	−63.8	5.14	2.48
10	−68.2	5.06	2.08
8	−72.7	4.76	1.58
6	−76.9	4.11	1.00
4	−79.9	2.89	0.41
2	−81.2	1.06	0.04

TABLE 8: THE CONVERTIBLE BOND IS DYNAMICALLY OPTIMALLY HEDGED IN THE HOMOGENEOUS MODEL WITH A *COMBINATION OF THE UNDERLYING, THE CDS, AND A CALL OPTION* OF SAME MATURITY AS THE CB AND STRIKE PRICE $K = \$22.50$. WE REPORT THE DYNAMIC OPTIMAL HEDGING RATIOS ON BOTH THE CALL OPTION AND THE CDS. THE HERO IS NOW ALMOST IDENTICALLY EQUAL TO ZERO AS THE CDS AND THE CALL OPTION CANCEL THE DEFAULT RISK AND THE VOLATILITY RISK. NOTICE THE STABILITY OF THE HEDGING STRATEGY. THE CB IS ALMOST PERFECTLY DECOMPOSED INTO A VOLATILITY INSTRUMENT AND A CREDIT INSTRUMENT

Stock price	Hedging ratio (Call option)	Hedging ratio (CDS)
16	4.18	−80.23
14	4.16	−80.36
12	4.12	−80.46
10	4.05	−80.53
8	3.92	−80.59
6	3.63	−80.67
4	2.62	−80.78

TABLE 9: WE NOW TRY TO TAKE FULL ADVANTAGE OF OUR EQUITY-TO-CREDIT HOMOGENEOUS MODEL. WE ANALYSE A NON-CONVERTIBLE BOND, OTHERWISE IDENTICAL TO THE TYCO CB. THIS IS A PURE CREDIT PLAY. WE USE AN OUT-OF-THE-MONEY PUT OPTION OF SAME MATURITY AS THE BOND AND STRIKE $K = \$3$, TO HEDGE DEFAULT RISK. WE REPORT THE DYNAMIC OPTIMAL HEDGING RATIO IN EQUIVALENT PUT POSITION (YOU SHOULD BUY THE PUTS TO ACHIEVE HEDGING) AND THE HERO. THE HERO INCREASES AS THE STOCK GOES DOWN BECAUSE OF THE INCREASING VOLATILITY RISK BORNE BY THE PUTS. WE COULD HAVE HEDGED DEFAULT RISK WITH THE UNDERLYING ALONE, BUT HERO WOULD HAVE BEEN MUCH LARGER ($14.60 AGAINST $S = \$16$)

Stock price	Hedging ratio (put option)	HERO
16	−41.59	1.45
14	−42.29	1.69
12	−43.28	1.99
10	−44.79	2.41
8	−47.27	3.01
6	−51.87	3.91
4	−62.87	5.45

TABLE 10: TO HEDGE AGAINST THE RESIDUAL VOLATILITY RISK MANIFESTED IN TABLE 9 WE NOW ADD AN AT-THE-MONEY CALL IN THE HEDGED PORTFOLIO INVOLVING THE CORPORATE BOND AND THE OUT-OF-THE-MONEY PUT. WE REPORT THE DYNAMIC OPTIMAL HEDGING RATIOS ON BOTH OPTIONS AND THE HERO. OBVIOUSLY, WE SHOULD SELL THE ATM CALL

Stock price	Hedging ratio (put option)	Hedging ratio (ATM call option)	HERO
16	−44.54	1.02	1.22
14	−46.25	1.54	1.36
12	−48.82	2.45	1.54
10	−52.98	4.21	1.74
8	−61.27	8.52	1.95
6	−80.97	21.45	2.21
4	−184.26	110.61	2.53

FOOTNOTES & REFERENCES

1. It won't matter for the prices (which are probabilistic averages) whether the model is inhomogeneous or homogenous. But wouldn't it matter for the hedging? More specifically, would we be hedging with a heavier delta in the homogeneous model, as the share goes to zero, the same way as in the inhomogeneous model (see Table 2)? The answer is yes, and it revolves entirely around the question of incomplete markets and optimal hedging, as in Ayache *et al* (2004). See Appendix.

2. I must qualify the latter statement a little bit. If the individual values of volatility characterizing the different sub-regimes inside a given super-regime i, $[\sigma_{i1}, \sigma_{i2}, \sigma_{i3}]$, are equal respectively to those occurring in a different super-regime k, $[\sigma_{k1}, \sigma_{k2}, \sigma_{k3}]$, then the result will be indistinguishable from a three-regime-switching model of stochastic volatility, only the transitions between the regimes will occur in many more different ways.
3. Of course the picture would have been different if, instead of considering a *nesting* of stochastic processes written one off another, such as stochastic volatility, stochastic volatility of volatility, etc., we had considered a radically different second factor, such as the hazard rate. For in that case, the regimes could not be labelled by anything short of two factors.
4. Our hedging strategies are optimal in the sense that you break even on average and the standard deviation of the P&L of the hedged portfolio is minimal. HERO is this minimal standard deviation (cf. Ayache *et al*, 2004).

■ Andersen, L. and Buffum, D. (2002) Calibration and implementation of convertible bond models. *Journal of Computational Finance*, 7(2).
■ Ayache, E. The philosophy of quantitative finance. *Wilmott* (forthcoming).
■ Ayache, E., Forsyth, P. A. and Vetzal, K. R. (2002) Next generation models for convertible bonds with credit risk. *Wilmott*, December.
■ Ayache, E., Forsyth, P. A. and Vetzal, K. R. (2003) Valuation of convertible bonds with credit risk. *The Journal of Derivatives*, Fall, 9–29.
■ Ayache, E., Henrotte, P., Nassar, S. and Wang, X. (2004) Can anyone solve the smile problem? *Wilmott*, January.
■ Ayache, E. and Tudball, D. (2004) One step beyond. *Wilmott*, January, 6.
■ Henrotte, P. (2004) The case for time homogeneity. *Wilmott*, January.
■ Taleb, N. (2001) 2001 Hall of fame. *Derivatives Strategy*, March.

8

Measuring Country Risk as Implied Volatility

Ephraim Clark

Wilmott magazine, September 2002

Investment in emerging markets has become a prominent feature of the financial globalization sweeping the world. Besides market risk, however, investments in emerging markets are also exposed to political phenomena that are not generally present in the more developed economies. The Mexican peso crisis and the Asian economic meltdown are two of the more spectacular examples. This problem is well known to banks and multinational companies by the name of country or political risk and, although assessment techniques in these domains are relatively well developed, they are not really adapted to economic and financial risk management. Traditional methods for assessing political risk range from the comparative techniques of rating and mapping systems to the analytical techniques of special reports, dynamic segmentation, expert systems, and probability determination to the econometric techniques of model building and discriminant and logit analysis. These techniques are very useful for identifying and analysing individual sources of political risk but are weak when it comes to translating them into standard risk measures such as variance and standard variation. Rating systems, although they are extremely popular, are ad hoc at best and have a dubious track record. Based on subjective analysis, analytical techniques are essentially informed opinions whose value depends on the quality of the analyst(s). The econometric techniques are more objective but are backward looking and their value depends on the quality of the model and the accuracy of the estimates of the exogenous variables. None of the techniques generate the market based information that drives modern financial theory and practice.

Some important market based information is available in the prices of traded sovereign securities. One popular technique is to use market prices to compute interest rate spreads. The problem with spreads on individual instruments is that they reflect the special features of the security being analysed. This can be overcome to a large extent by considering the ensemble of a sovereign's foreign obligations. However, spreads reflect general market risk as well as the specific country's political risk and are difficult to interpret in terms of variance and standard deviation. For example, a three percent spread over a treasury yielding 2% reflects a

Contact address: Middlesex University, The Burroughs, London NW4 4BT, UK
E-mail: e.clark@countrymetrics.com www.countrymetrics.com

lower level of volatility than a 3% spread over a treasury yielding 4%. This having been said, market prices contain a wealth of forward looking information that can be used to measure a country's riskiness.

The implied volatility of a country

One way to exploit the market information on traded sovereign debt is to borrow from standard option pricing techniques and estimate a country's riskiness as its implied volatility. The implied volatility for a market security can be obtained by running the option pricing model with volatility as the unknown. This involves knowing the price of the underlying security, the price of the option, its exercise price, its expiry date and the risk free rate of interest. Although in the case of a country the risk free rate and, in many cases, the market value of the debt are the only parameters that are readily observable, the other parameters can easily be estimated.

The key variable and also the most difficult to estimate is the price of the underlying security. In the case of a country, it is nothing more or less than the country's international market value. From the standpoint of the international investor, he wants to know the value of the country's assets in terms of its ability to generate the foreign exchange that will enable payment of the interest and dividends on his investments in that country. In this respect, it is no different from a company. To estimate this value I use the standard financial technique of discounted cash flows. Surprisingly enough, it is easier to estimate macroeconomic cash flows than it is to estimate the cash flows of most private companies. Consider the following notation:

X = total exports not including investment income measured in USD
M = total imports not including investment income measured in USD
M^C = imports of final consumption goods measured in USD
C = local consumption measured in USD
b = total income from the sale of the economy's output of final goods and services measured in USD
t = time
$b_t = X_t + C_t - M_t^C$
a = total expenditure by the economy for the purchase of final goods and services measured in USD
$a_t = M_t + C_t - M_t^C$
$R = 1 + r$ where r represents the economy's internal rate of return
V_t = the value of the economy at the beginning of period t measured in USD

From the definitions of b and a it is clear that $b - a = X - M$. The value of the economy in USD can be written as the present value of expected macroeconomic cash flows:

$$V_t = E\lfloor (b_t - a_t) + (b_{t+1} - a_{t+1})R^{-1} + \ldots + (b_n - a_n)R^{-(n-t)} \rfloor \tag{1}$$

where all transactions take place on the first day of each period. It is interesting to see the relationship between this equation and the national accounts as they are usually presented. To see this, first calculate V_{t+1} and substitute into (1). Then multiply by $1 + r$ and rearrange. Ignoring interest on net exports, which disappears in continuous time, this gives

$$X_t - M_t + C_t + (V_{t+1} - V_t) = rV_t + C_t \tag{2}$$

We recognize the left hand side (LHS) of equation (2) as net domestic product where net investment in any year is equal to $V_{t+1} - V_t$. The right hand side (RHS) is profits rV_t plus cost (consumption) C_t.

There are many ways to get an estimate of V_t. One technique that I have used is to define a process for $(b - a)$. For reasons I won't go into here but which are linked to the balance of payments identity, a mean reverting process seems appropriate. Let $q(t) = (b_t - a_t)$ and

$$dq(t) = k\left[\frac{\alpha}{k}e^{\beta t} - q(t)\right]dt + \sigma\, dz \tag{3}$$

where

k = the speed of adjustment parameter
$(\alpha/k)e^{\beta t}$ = the long term mean of q
dz = a Wiener process with zero mean and variance equal to dt
σ^2 = the variance of $dq(t)$

Changes in $q(t)$ will tend to move toward the long term mean, $(\alpha/k)e^{\beta t}$. The long term mean can grow, decrease, or remain constant depending on whether $\beta > 0$, $\beta < 0$, $\beta = 0$.

If we note that for a national economy, the relevant time horizon is infinity so that V is a function of q and not of t and apply Ito's lemma and the boundary condition

$$\underset{q \to \pm\infty}{\text{Lim}} |D_q(q)| < \infty \tag{4}$$

we get a differential equation whose solution is:

$$V(q) = \frac{q}{r+k} + \frac{\alpha e^{\beta t}}{r(r+k)} \tag{5}$$

To get an estimate of V we have to estimate the parameters α, β and k. To do this, we solve equation (3), discretise the process and make a number of tedious manipulations, which gives an equation that can be used to estimate the parameters for the process (3):

$$\Delta q(t) = \gamma_1 + \gamma_2 q(t) + \varepsilon(t) \tag{6}$$

where

$$\gamma_1 = \frac{\alpha}{k+\beta}e^{\beta t}[e^{\beta \Delta t} - e^{-k\Delta t}]$$

and

$$\gamma_2 = [e^{-k\Delta t} - 1]$$

Once V_t has been estimated, the implied volatility can be calculated in 4 steps:

1. Calculate the nominal value of total outstanding foreign debt.
2. Calculate the market value of total outstanding foreign debt.
3. Calculate the duration of total outstanding foreign debt.
4. Plug this information into the Black–Scholes formula:

$$B_0 = V_0 N(-d_1) + K e^{-rt} N(d_2) \tag{7}$$

where B is the dollar market value of the debt, t is the duration of the debt, r is the USD risk free rate of interest, K is the nominal dollar amount of foreign debt outstanding and $N(d)$ is the value of the standardized, normal cumulative distribution evaluated at d with

$$d_1 = \frac{\ln \frac{V_0}{K} + (r + \frac{\omega^2}{2})t}{\omega\sqrt{t}} \tag{8}$$

and

$$d_2 = \frac{\ln \frac{V_0}{K} + (r - \frac{\omega^2}{2})t}{\omega\sqrt{t}} \tag{9}$$

In the examples that follow, I show how the implied country volatility can be used as a tool to measure country risk.

The Mexican peso crisis of 1994

On December 20, 1994 international investors were supposedly surprised by a 13% devaluation of the Mexican peso that fell from 3.4647 for one dollar at the end of trading on December 19 to 3.9750 at the end of trading the next day. Furious flight from Mexican assets pressured the peso to the point that two days later the authorities surrendered to the inevitable and allowed the peso to float. By March it had fallen by 54% to 7.5 for one dollar. The wicked witch of the international financial markets had waved its wand and turned Mexico, the handsome prince of the emerging markets, into the ugly frog of a risky developing country. The stock market plunged, interest rates soared, debt default was threatened and international investors (including domestic residents capable of investing abroad) were ostensibly left with staggering losses that even the loving kiss of a $50 billion rescue package organized by Princess USA was unable to reverse. Furthermore, as the crisis sharpened, it also spread worldwide to the other "emerging markets", causing falling stock markets and massive capital outflows.

The conventional post mortem has it that economic reform and the government's hard sell of a NAFTA-linked economic miracle drew foreign investors into the Mexican economy to the tune of more than $90 billion between 1990 and 1993. Aided by a newly privatized banking sector loaning with reckless abandon, Mexico went on a consumption binge. Foreign capital financed a sharp drop in the domestic savings rate and a current account deficit that grew to 6.5% of GDP in 1993 and to 7.7% in 1994. International investors supposedly did not realize that the situation was untenable over the medium term and the outgoing Salinas government doggedly refused to either curb consumption or to devalue the peso. Nevertheless, it is generally agreed that when the reckoning did come, investors were taken completely by surprise and Mexico's underlying economic and financial situation did not warrant the humiliating treatment inflicted on it by the international financial markets. It is noted that the government's budget was balanced, the economy had been opened up and deregulated, exports were growing, inflation was low at 7% and falling, economic growth was strong, and the North American Free Trade Agreement was signed and being implemented.

According to conventional wisdom, then, the peso crisis was a mindless overreaction by international investors to a bursting "speculative bubble" caused by a sudden realization of the increased political uncertainty associated with the PRI's steady loss of political control,

punctuated by the uprising in the impoverished state of Chiapas, the assassinations of Luis Donaldo Colosio, the ruling party's presidential nominee, and Jose Francisco Ruiz Massieu, a close friend and former brother-in-law of President Salinas, and a bungled devaluation. In other words, conventional wisdom has it that investors were overreacting to perceived changes in the country's political fragility.

There are several shortcomings to this attractive conventional view, which seems to fit the facts in general. First of all, it fails to explain how otherwise sophisticated international investors could have remained oblivious so long to events that were known and had been developing over an extended period. It also fails to explain what caused them to overreact when they finally did get wise. Finally, it fails to explain what caused a crisis that was uniquely Mexican in nature to spread to the other emerging markets in general, including those as far afield and economically different as Hong Kong, Thailand, Indonesia, and Malaysia - to mention only a few - with no apparent connection to Mexico.

Using the concept of implied country volatility developed in the preceding section, we answer the first two questions. We answer the third question, the contagion effect, in the following section when we discuss the Southeast Asian crisis.

Table 1 shows that international investors were not really surprised by Mexico's meltdown. The volatility implied by the observed risk premiums, which indicate Mexico's riskiness as perceived by the international financial community, increases steadily from 1991 to 1993. By the end of 1993, the eve of scheduled presidential elections, Mexico's perceived riskiness was at the extremely high level of 70.46%, 54.5% higher than it was at the end of 1991. The level of implied volatility suggests that investors were expecting a large move in Mexico's economic situation. There really was no "overreaction". This evidence runs counter to the popular argument of conventional wisdom that investors were surprised by, and overreacting to, perceived changes in the country's political fragility. It suggests that international investors were aware of Mexico's evolving economic and political situation and that they were adjusting to it in a continuous, orderly manner from as far back as 1991.

TABLE 1: MEXICO'S IMPLIED VOLATILITY

1991	1992	1993
45.60%	56.70%	70.46%

The Southeast Asian crisis of 1997

On July 2, 1997 the Thai baht was abruptly devalued by 20% despite weeks of desperate moves to prop up the currency, including central bank intervention of $8.7 billion on the spot market and $23 billion in forward contracts, interest rate increases from 12% to 18% and restrictions on foreign speculators. By the end of the year the baht crisis had spread around the world. The median devaluation of the five East Asian tigers hardest hit by the crisis – Indonesia, Korea, Malaysia, the Philippines and Thailand – was 80%. The International Finance Corporation's (IFC) emerging stock market index dropped by 20% between June and December and its Asian index fell by 53%. By the end of the year the baht had depreciated by 93%, the Hong Kong

dollar, the Korean won and the Taiwan dollar were under attack and their stock markets were nose-diving, currencies and equity prices in Eastern Europe and Latin America were falling and in November, Korea, the world's eleventh largest economy and an OECD-member country, became the recipient of the world's largest-ever rescue package.

What happened? Conventional wisdom has it that in spite of a benign international background with high rates of growth in world trade and declining spreads on international borrowing, international investors suddenly awoke to the reality of structural weaknesses in the private financial sector, including resource misallocation and maturity and currency mismatches as well as public sector economic mismanagement regarding the exchange rate, financial regulation and implicit or explicit government guarantees. The rude awakening caused a crisis of confidence that the five countries, vulnerable because of the build-up of private sector, short-term, un-hedged debt, were unable to overcome. Nevertheless, it is generally agreed that when the reckoning did come, the countries' underlying economic and financial situation did not warrant the humiliating treatment inflicted on it by the international financial markets. It is noted that public borrowing was subdued, most of the countries were running a fiscal surplus, inflation was low relative to most other developing countries and savings rates were high. With this in mind, conventional wisdom has it that the Asian crisis was another mindless overreaction by international investors.

Again, there are several shortcomings to this attractive conventional view, which seems to fit the facts in general: 1) it fails to explain how otherwise sophisticated international investors could have remained oblivious so long to events that were known and had been developing over an extended period; 2) it also fails to explain what caused them to overreact when they finally did get wise; 3) it fails to explain what caused a crisis that was uniquely Asian in nature to spread to the other emerging markets in general, including those as far afield and economically different as Latin America and Eastern Europe.

Using the concept of implied country volatility suggests that, as in the case of Mexico, markets were neither surprised nor overreacting. In Table 2, we can observe that as far back as 1993, a year before the peso crisis, the market considered these countries as extremely risky with implied volatility ranging from 41% to 78%. When the peso crisis manifested itself in 1994, the markets were already expecting a large move in the Southeast Asian countries' economies. Furthermore, by 1996 implied volatility for Indonesia and Malaysia was at the levels reached at the height of the Mexican peso crisis in 1994. For Korea, Thailand and the Philippines, it had fallen only slightly. Over the whole period from 1993 to 1996, implied volatility rose for 3 countries (Indonesia, Malaysia and the Philippines) and fell for 2 (Korea and Thailand). These two countries that experienced a fall in their implied volatility were the two that had the highest implied volatilities in 1993. The reduction only brought them to the same level as the other three countries. All this suggests that as early as 1993 the market was anticipating the potential difficulties that would eventually materialize in 1997. It is interesting to note that Indonesia had the highest implied volatility on the eve of the crisis and was the country that suffered most when it hit. On the other hand, the Philippines had the lowest implied volatility and was the least affected.

It might seem surprising to rate the Philippines as the lowest risk before 1997 but in light of the actual 1997 events, this was, on the contrary, a very shrewd assessment. Indeed, as evidenced by *a posteriori* results, even though the Philippines had not been able to achieve its neighbours' economic performance over the last decade, it did not suffer their specific weaknesses to the same extent.

TABLE 2: IMPLIED VOLATILITY OF FIVE SOUTHEAST ASIAN ECONOMIES

Year	1993	1994	1995	1996
Indonesia	55.0%	63.3%	56.2%	67.8%
Korea	77.7%	70.2%	67.9%	63.9%
Malaysia	43.6%	54.1%	47.0%	53.2%
Philippines	41.4%	49.8%	46.9%	46.2%
Thailand	63.3%	62.1%	56.5%	56.0%

Conclusion

In this paper I tried to outline the concept of implied country volatility, how it can be measured and its pertinence to country risk assessment. It has shown itself to be a reliable tool with many uses in international risk management. It is also interesting to note that the country's international market value, which makes it possible to estimate the implied country volatility, is a powerful concept with documented results in forecasting sovereign debt reschedulings and defaults as well as in the construction of international portfolios of stocks, long term government bonds and money market instruments that outperform their benchmarks by wide margins.

9

Next Generation Models for Convertible Bonds with Credit Risk

E. Ayache,*† P. A. Forsyth‡ and K. R. Vetzal§

Wilmott magazine, December 2002

Convertible bonds are hybrid securities which offer equity-like returns when the share of the issuing firm is strong, yet behave like conservative fixed-income investments when the stock market is either stagnant or negative. Indeed the convertible bond is essentially a bond that can be converted into shares, a feature which allows the equilibrium of interests between the three parties involved, the issuing company, the equity investor and the fixed-income investor to be struck more efficiently than was the case when equity and fixed-income were treated as separate investment categories, involving different, if not incompatible, standards. As the company issuing the convertible bond sells an embedded option to convert into its shares, it expects its creditors to charge a lower fee than would otherwise apply to its credit class, hence is able to pay lower coupons. Reciprocally, the fixed-income investor earning these coupons is rewarded by his upside participation in the performance of the share. The equity investor, on the other hand, whose basis of judgment is the price of the equity and its expected return, makes up for the premium paid over parity by the downside protection that the bond floor automatically provides, and by the fact that the convertible bond coupons are usually set higher than the projected dividends (hence creating the notion of a break-even date).

It is obvious, from this preliminary analysis, that the equity level will be the determining factor in the convertible bond value, where "value" means what specifically distinguishes the convertible bond from an ordinary fixed-income investment or an ordinary equity investment. This is the reason why the quantitative analysis of convertible bonds lends itself naturally to

Contact addresses: *ITO 33 SA, 36 Rue Lacépède 75005 Paris, France, numbersix@ito33.com
‡Department of Computer Science, University of Waterloo, Waterloo, ON, Canada, paforsyth@elora.math.uwaterloo.ca
§Centre for Advanced Studies in Finance, University of Waterloo, Waterloo, ON, Canada, kvetzal@watarts.uwaterloo.ca
†The author would like to thank Philippe Henrotte for very helpful and illuminating comments.

the Black–Scholes analysis where the share price is the state variable, and dynamic hedging strategies are the basis for the valuation of the embedded option. Not only will the convertible bond value depend on the volatility of the share, but we shall expect the share price itself to set the dividing line between equity behavior and bond behavior (as well as between the ensuing concerns, respectively about share price volatility and credit quality volatility). It turns out indeed that the Black–Scholes analysis provides just the right unifying framework to formulate the convertible bond pricing problem. Unification comes at a cost, however. For we can no longer ignore, once the share becomes the driving factor, what direct effect it may have on the issuer's credit quality. And if such an effect is to be assumed, we can no longer but model it explicitly.

Credit spread and the fixed-income logic

The quantitative measure of credit quality has traditionally been the credit spread. Risky bonds are priced by the market at a discount to sovereign debt, and the price difference, when expressed in terms of the excess in the implied yield, is the credit spread. Bonds maturing at different dates can imply different credit spreads, hence creating credit spread term structure. When the bonds are zero coupons, the term structure of credit spread is equivalent to giving the whole array of risky discount factors, etc. So credit spread is really a notion from fixed-income analysis, and for that reason quite foreign to a framework such as Black–Scholes, where the state variable is the underlying share. As long as bond pricing was the sole concern, all that the fixed-income analyst needed was the spot yield curve and the spot credit spread curve. The necessity of modeling stochastic credit spread, however, became evident with the emergence of credit derivatives. But when the derivative payoff did not specifically depend on the credit spread (for instance options on corporate bonds), a risky yield curve model could still be developed along the lines of the traditional yield curve models (Hull and White, Black Derman Toy, Heath Jarrow Morton), in such a way that the changes of credit spread curve and the changes of risk-free yield curve would indistinguishably be captured by the overall changes of the risky yield curve. Only when credit spread changes had to be separately modeled did the need arise to identify the real "physical" variable underlying these changes, the instantaneous probability of default of the issuer (assuming a deterministic recovery rate). Just as the instantaneous interest rate is the state variable driving the basic yield curve models (e.g. Hull and White), the instantaneous probability of default, or hazard rate, drives the stochastic credit spread models. One writes directly the stochastic process followed by the hazard rate (possibly with a time dependent drift in order to match a given spot credit spread curve) and generates different prices for risky zero coupon bonds in different states of the world, in other words, different credit spread curves.

Credit spread and the convertible bond

The relation between the convertible bond and the credit spread seemed at first to arise only from the "bond character" of the convertible. The embedded equity option would be priced in the Black–Scholes framework alright, where discounting takes place at the risk-free interest rate, but the presence of a fixed-income part of course implied that "something" had to be discounted under a risky curve, if only to be consistent with the fixed-income analysis of the issuer's debt. The difficulty, however, was that bond component and equity component were

not readily separable. On the contrary, we saw that the convertible bond could very well display a mixed behavior, now like equity, now like bond, depending on the share level. This is why the question "How exactly to apply the credit spread in the convertible bond pricing tree, and how to link that to share price?" became the central problem in convertible bond valuation.

One early attempt interpreted the mixed behavior of the convertible bond in probabilistic terms (Goldman, 1994). It is only with *some probability*, so the argument went, that the convertible bond would end up like equity or end up like pure bond, and that probability was identified with the probability of conversion. Ignoring what distortions might arise from other embedded options, such as the issuer's call or the holder's put, the suggestion was to discount the value of the convertible bond, at nodes of the pricing tree, with a weighted average of the forward instantaneous risk-free interest rate and the forward instantaneous risky rate. The delta of the convertible bond, now identified with the probability of conversion, would determine this weighting. While this approach certainly fulfilled the wish that the convertible bond be treated as equity when most likely to behave like equity, and as bond when most likely to behave like bond, it certainly did not explain the financial-theoretic meaning of the mixed discounting. Perhaps it can be argued, in a global CAPM framework, that some future cash flow ought to be discounted with some exotic mixture of some given discount rates. The problem is, no sense can be made of a situation where the mixing takes place *locally*, and the weighting varies from one state of the world to the other.

More recently, another approach (Tsiveriotis and Fernandes 1998) thought better to interpret the mixed behavior of the convertible in *actuality* rather than *potentiality*. If the convertible bond is really a combination of a bond and an equity option, why not actually treat it like one, and split it into two components, one to be discounted risk-free and the other risky? When there is involved the possibility of early call or early put, however, the two discounting procedures cannot take place completely separately, so what T&F have proposed is to throw into the bond component whatever value accrues from the issuer's liability (either promised or contingent cash flows), and into the equity component whatever value accrues from the holder's contingent claim to convert into the issuer's share or from the issuer's *own* contingent claim (the idea being that the issuer could always deliver his shares, default or no default, and that he would not exercise his option to call back the convertible in case of default or shortage of cash). The two components are priced as two distinct assets. The equity component has the equity or nothing payoff as termination value, and the bond component (or the cash-only component, as T&F labeled it) the cash or nothing payoff. The two backward recursions are then coupled through the following algorithm. In case the convertible bond checks for early conversion – or early call – in a certain state of the world, the equity component is set equal to conversion value – or early redemption value – in that state, while the cash-only component is set to zero; alternatively if the convertible bond checks for early put, the cash-only component is set equal to the put strike and the equity component is set equal to zero. The cash-only component, on the other hand, earns the coupons in all states of the world.

It is interesting to note the circularity, or self-reference, that is inherent in both approaches. The value of the convertible bond crucially depends on the proportion in which the credit spread is applied to it, yet this proportion ultimately depends on the convertible bond itself. In the first approach the proportion is determined by the delta which is itself a derivative of the convertible bond value, and in the second, the proportion is determined by the value of the cash-only component relative to the equity component, which in turn depends on the particular constraint that the convertible bond as a whole checks, the conversion constraint, the call constraint or

the put constraint. Mathematically, this translates into non-linearity. This is the reflection of the fact that the risky component of the convertible, which is the value that the holder is liable to lose in case of default – and which, by the same token, he will argue he is entitled to recover a fraction of when the assets of the defaulted company are liquidated – depends on the optimal behavior of the holder himself.[1] While the recovery entitlement of the holder of a straight bond is a straightforward fraction of the present value of the bond, the holder of an option-embedded bond, such as the convertible bond, will typically want to recover more, for he will invoke what contingent rights he was holding on top of the fixed ones. And this notably depends on his optimal exercise or conversion policy in case of no default.

The missing story of default

However, this whole explanation in terms of loss and recovery in case of default is totally missing from the T&F paper. As a matter of fact, the problem with the T&F approach is that it falls one step short of telling the whole story about the convertible bond under default risk. While it certainly proposes an *actual* splitting of the convertible bond into two distinct components, and reproduces its desired extreme behaviors (pure bond, pure equity) at the extremes of the share price range, it does not say what actually *happens* to the convertible bond in case of default. And default can take place anywhere between those two extremes. T&F's line of argument is simply to identify the cash-only component (why not through a complex procedure involving two pricing PDEs and their local coupling), then to uncontroversially apply to it the credit spread, in the old fixed-income logic.

A few paragraphs back, we argued that if one wishes to model the actual default process and not just describe its phenomenological consequence, the credit spread, one has to get hold of the real physical variable underlying it, the hazard rate. All the more so when a pricing framework, such as Black–Scholes, already imposes on us a reduction in terms of state variables. Now it certainly makes sense to choose a credit spread of some given finite maturity, say one year, as state variable, and develop a stochastic model for credit quality in the same vein as the so-called "market models" of interest rates. Furthermore, one can assume some explicit correlation between the share process and the credit spread process and complete the program that we have announced earlier, of explicitly modeling the effect of the issuer's share on his credit quality. The T&F approach would generalize to this framework, the tree of the cash-only component would become two-dimensional and discounting would take place under local credit spread. However, this would still not answer the question why the cash-only component has to be discounted with credit spread in the first place, any better than just the postulation that it somehow condenses the issuer's liability and that credit spread should mimetically apply to it.

To our eyes, a model that relies on surface resemblance and no real explanatory argument is not a satisfactory model. Lying at the crux of the T&F model is the proposition that the equity component and the cash-only component are two identifiable, if hypothetical, contingent claims, hence should follow the Black–Scholes PDE, the one with risk-free discounting and the other with risky discounting. Now the Black–Scholes PDE is not just a pricing black box. It relies on "physical" first principles which are the continuous hedge and the no-arbitrage argument. The causal explanation of the Black–Scholes PDE is the precise elaboration of *that which happens* to the hedge portfolio over the infinitesimal time increment dt. On the other hand, given that the whole idea of the T&F splitting is to model the split behavior of the convertible bond under default risk, then why not go all the way, and try to spell out exactly what can

happen to the convertible bond in the eventuality of default? For the convertible bond is the real contingent claim after all. T&F's progress relative to the Goldman paper was that they took the step from *probabilistically* mixing two distinct credit stories that are likely to happen to the same instrument, to *actually* decomposing that instrument into two distinct credit entities. So what we are now urging is that the credit story really be told about those two credit entities.

The story of default told at last

"What can happen in case of default" is a question hinging *directly* on the probability of default. It imposes the instantaneous probability of default, rather than *effects* thereof such as the credit spread, as the original cause – or true explanatory variable. The extension to the case of the convertible bond under credit risk, not of the Black–Scholes PDE, but of *the line of reasoning* underlying the Black–Scholes PDE, would then run as follows.

Calling p the instantaneous probability of default (or the intensity of the Poisson default process), what "infinitesimally" happens to the composite portfolio, convertible bond and dynamic hedge, under default risk is:

(a) with probability 1−pdt: no default. The usual Black–Scholes continuous hedge argument applies, the holdings in the underlying share are chosen is such way that the hedge portfolio is immune to market risk over the time increment dt, and infinitesimal P&L can be written as (assuming no dividends for simplicity):

$$\delta\Pi = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2}\right) dt$$

(b) with probability pdt: default. The infinitesimal P&L is literally swamped by the loss of the defaultable fraction X:

$$\delta\Pi = -X$$

The expected P&L is then expressed as follows, neglecting second order terms:

$$E(\delta\Pi) = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - pX\right) dt$$

If we now assume that the probability of default is given in the risk neutral world,[2] we can equate the above expectation with the risk-free growth of the portfolio:

$$E(\delta\Pi) = r\Pi\, dt$$

and obtain the following PDE:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} = rV + pX \tag{1}$$

Differences explained

Once the problem of convertible bond valuation under default risk is framed in such a unifying formalism, the differences between all the models that the practitioners have been using with more or less rigor find an explanation in terms of different choices of X.

Grow risky, discount risky

Let X be the whole portfolio, or in other words let us assume that both the convertible bond and the underlying share drop to zero in case of default, and the PDE will transform into:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r+p)S\frac{\partial V}{\partial S} = (r+p)V$$

This corresponds to the popular model, mnemonically known as "Grow risky, Discount risky".

The general model

A more general model is one in which the share drops to a residual value $(1-\eta)S$ upon default, and the convertible bond holder is entitled to recovering a fraction F of his investment. He would then have the option either to convert into shares at their residual value, or to recover F. In this case, X would be expressed as:

$$X = V - \max[\kappa(1-\eta)S, F] - \frac{\partial V}{\partial S}\eta S$$

where κ is the conversion ratio, and the PDE governing the convertible bond value under default risk would become:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r+p\eta)S\frac{\partial V}{\partial S} = (r+p)V - p\max[\kappa(1-\eta)S, F] \tag{2}$$

The question remains how to model F. Should it be a fraction of the face value N of the convertible bond? Or should it be a fraction of the market value of the corresponding straight bond: what the practitioners call its "investment value"? To be exact, the recovered fraction should be established by the liquidator after default has taken place. However, we can assume an *a priori* recovery rate applying uniformly to the issuer's liabilities, whatever their nature, certain or contingent. What you recover is proportional to what you are owed. The holder of a coupon-bearing bond is owed more than the holder of a zero coupon bond, hence should recover more. And the holder of a bond with an embedded option, say a put, is owed more than the holder of a bullet bond, hence should recover more. The concept of probability, according to Ian Hacking, emerged from those gambling situations where the players were for some reason prevented from pursuing the game until the end. The game had to be settled one way or the other, and the money at stake distributed according to some rationale. This is how the notion of a player's best chances of winning it first made its appearance, or in other words, his expected gain. Settling the case of default of a convertible bond issue is no different. What the holder is supposed to recover in case of early termination due to default is the recovery fraction of the expected value of the cash flows he would otherwise get in case of no default.

Modeling the cash claim of the convertible bond holder

So what is it exactly that the holder of the convertible bond is owed prior to default?

The N-model

If you say it is a fraction of the face value N, then the PDE would look something like that:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta)S\frac{\partial V}{\partial S} = rV + p(V - \max[\kappa(1-\eta)S, RN])$$

Let us call it the N-model.

The Z-model

If you say it is a fraction of the present value of the outstanding coupons and face value, then you would have to determine first whether this present value should be computed under risky or risk-free yield curve. It all depends on the interpretation of that which "the convertible bond holder is owed prior to default". Since we are in the business of building a mathematically consistent model of the fair value of the convertible bond and we believe, for that matter, that the market is the fairest dispenser of value, then a possible interpretation of "the value that the holder of a convertible bond is owed prior to default" could simply be the fair value, or market value, of the convertible bond itself! And this market value would already have the default risk factored in it. Therefore a somewhat extreme model would be one where the holder simply recovers a fraction of the convertible bond value prior to default (cf. the paper by A. Takahashi *et al.*, 2001):

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta)S\frac{\partial V}{\partial S} = rV + p(V - RV)$$

The reason why this is not satisfactory is that we had assumed on the other hand that the holder would still have the right to convert into the residual value of the underlying share upon default, so the PDE would really have to look like:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta)S\frac{\partial V}{\partial S} = rV + p(V - \max[\kappa(1-\eta)S, RV])$$

The recovery rate of the underlying share $(1-\eta)$ and the recovery rate of the convertible bond R being completely independent, we would then be faced with the possibility that $\kappa(1-\eta)S$ may be greater than RV, even though $\kappa S \leq V$ at all times. Nothing would then guarantee that the holder may not optimally elect to convert into the residual value of the share, over and above the fact that the value he is recovering anyhow, the recovery fraction of the *convertible* bond, already incorporates the value of a conversion right! While the recovery procedure is aimed at compensating the convertible bond investor in case of default, we certainly do not suppose that it ends up doubling his conversion rights! Conversely, if RV is much greater than $\kappa(1-\eta)S$, say $R = 1$ and $\eta = 1$ in an extreme case, it would also seem strange that the holder should recover the full convertible bond value (including the full value of the conversion right)

when the share has actually dropped to zero! To sum up, if we are keen on leaving the holder the right to convert right at the time of default, then the only way to avoid this conflict is to assume that the value he is likely to recover, or in other words that which "he was owed prior to default", has been stripped of the conversion rights first.

So it seems we are back to modeling F as the present value of the underlying straight bond, and the above digression would have only convinced us that the fair value of "what the holder is owed prior to default" should take into account default risk, or in other words, that the present value of the outstanding coupons and face value should be computed under the risky curve. Thus we have:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta)S\frac{\partial V}{\partial S} = rV + p(V - \max[\kappa(1-\eta)S, RZ])$$

where Z, the present value of the straight bond, solves the same PDE as the convertible bond, only without conversion rights:

$$\frac{\partial Z}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 Z}{\partial S^2} + (r + p\eta)S\frac{\partial Z}{\partial S} = rZ + p(1-R)Z$$

If the hazard rate were independent of S, the last PDE would integrate to:

$$Z(t) = \sum_{t_i \geq t}^{T} C_i e^{-\int_t^{t_i}(r+p(1-R))du}$$

where C_i are the outstanding cash-flows. (This is simply a forward calculation under the risky yield curve). Otherwise we would have to solve in parallel two full PDEs, Let us call this model the Z-model.

The P-model

So far we have considered two interpretations of the notion of recovery. These corresponded to two different interpretations of the notion of default of the convertible bond. In one case, the convertible bond was construed as a debt instrument, binding the issuer to redeem the principal at maturity and to pay interest in the meantime. Default in this case meant that the *structure* of the convertible bond as debt instrument was over, and that the investor had to be reimbursed the amount of money he had initially invested. Recovery would then simply appear as a case of early redemption, caused by default, and it would mean recovering a fraction of the principal right away. In the other case, the convertible bond was construed as a tradable asset whose fair market value represented all the value there is to consider, and default simply meant that this value had dropped to zero. Recovery would then amount to recovering part of the pre-default holdings, or in other words, a certain fixed proportion of this value. So in the one case, default is a failure of a contractual obligation while in the other, it is simply a failure of market value. Recovery is defined accordingly: in one case the holder is owed the principal while in the other he is owed this market value, and the difference between the two interpretations is further reflected in the fact that the recovered value in the first case is purely nominal and independent of market conditions, while in the other, it is itself subject to interest rates and credit spread discounting. The two models further differ in that the N-model

does not really discriminate between the holder of a zero-coupon bond and the holder of a coupon-bearing bond as far as recovery is concerned, while the Z-model does. However, the two models have in common that the holder is offered an amount of cash (RN or RZ) right after default and right before he exercises his last option to convert, and that that is the end of the story.

Now consider a refinement of the N-model where we wish to compensate the holder of a coupon-bearing bond more than the holder of a zero-coupon bond. What should he recover exactly? We cannot just pay back a fraction of the sum of the face value and the outstanding coupons because the coupons were just the reflection of the scheduling of the issuer's debt over *time*. A more appropriate model seems to be one where the holder recovers a fraction of the *present value* of the outstanding straight bond, where this present value is *discounted under the risk-free curve*. In a sense, the occurrence of default eliminates default risk, and we wake up the day after in a default-free world where this present value calculation is the only way to discriminate between the holder of a zero coupon bond, and the holder of a coupon-bearing bond. Hence the following PDE:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta) S \frac{\partial V}{\partial S} = rV + p(V - \max[\kappa(1-\eta)S, RP])$$

where:

$$P(t) = \sum_{t_i \geq t}^{T} C_i e^{-\int_t^{t_i} r(u)\,du}$$

Let us call this the P-model.

Although it looks as if the P-model is just intermediate between the N-model and the Z-model (it achieves more than the N-model in integrating the coupons but achieves less than the Z-model in not applying full discounting with the credit spread), in fact it opens a whole a new perspective for it is the first among the models we've considered so far to assume that life continues after default, and to bring the post-default world into the picture. Indeed the discounting of the recovered value under risk-free curve, simple as it may seem, is in fact an instance of a general category of models which we will examine later and which *couple the pre-default world and the post-default world*.

The optimal model

For now let us just note that a common point between all the previous models is that none of them involved non-linearity such as alluded to earlier. The situation is somewhat akin to a free-boundary problem. In all the cases above, we imposed on the convertible bond PDE that the recovery fraction be some value computed separately. Be it a fraction of the face value, or of the present value of the outstanding payments, we never let F be determined freely by the value of the convertible bond itself. As mentioned previously, the holder of a risky bond with an embedded option will want to argue that he was owed more prior to default than just the present value of the fixed income part of the bond. Having agreed to exclude the option to convert from the treatment of recovery, this means that contingent cash-flows such as puts and calls have ideally to be incorporated in the holder's claim to

recovery. The problem is that their precise value will depend on whatever optimal exercise policy the holder was supposed to follow prior to default. Just as the free-boundary problem inherent in American option pricing translates into maximizing the value of the American option and the early exercise boundary is itself part of the solution (see Wilmott, 1998), we feel that the fraction of the convertible bond with other embedded options that the holder will ideally want to claim for recovery, is the greatest such fraction subject to the constraint that it may be legally argued, once default has taken place, that this fraction was owed to the holder. We are implying, in other words, that our algorithm for computing the recovery fraction F should really act as a lawyer trying to optimize his client's interests, and that the real lawyers should perhaps equip themselves with our convertible bond pricing model under default risk, once it is completed, in order to best serve their client. And just as the free-boundary problem is essentially non linear, we should expect ours to be non linear.

Our proposed model: the AFV splitting

Trying to bring together all the desiderata and the constraints that our "philosophical" analysis of default and recovery seems so far to suggest for the case of the convertible bond, we can summarize them as follows:

- Split the convertible bond value into two components: $V = B + C$.
- B is the value that the holder will argue he was owed anyway prior to default, and consequently will claim he must recover a fraction of according to some universal recovery rate R. Hence $F = RB$, in our case.
- B will be worth at least the present value of the underlying straight bond, for the holder will typically argue that he was owed more than this present value in case of an embedded options such as a put.
- B should not include the option to convert. On the contrary, the option to convert acts "externally" to the process of recovery, for the holder will retain the right to convert at the residual value of the share once default *and* recovery have taken place.
- C would then have to incorporate this option to convert, and would consequently finish as the holder's last option to convert into the residual value of the share when default takes place.

Given our general PDE for the convertible bond:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta)S\frac{\partial V}{\partial S} = (r + p)V - p\max[\kappa(1 - \eta)S, RB]$$

subject to the constraints of early call and early put:

$$V \geq \max(B_p, \kappa S)$$

$$V \leq \max(B_c, \kappa S)$$

(where B_p is the holder's put strike price, and B_c the issuer's call price, $B_c > B_p$), the following coupled PDEs should in effect be solved in order to value the convertible bond under

default risk:

$$\frac{\partial B}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 B}{\partial S^2} + (r + p\eta)S\frac{\partial B}{\partial S} = (r + p)B - pRB$$

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + (r + p\eta)S\frac{\partial C}{\partial S} = (r + p)C - p\max[\kappa(1 - \eta)S - RB, 0]$$

with initial conditions:

$$B(S, T) = N$$

$$C(S, T) = \max(\kappa S - N, 0)$$

and subject to the following algorithm (which is the cause of non-linearity):

- If $B_p > \kappa S$ and the continuation value of $B + C$ is less than B_p then $B := B_p - C$
- Else if $B_p \leq \kappa S$ and the continuation value of $B + C$ is less than κS then $C := \kappa S - B$
- If $B_c < \kappa S$ then $C := \kappa S - B$
- Else if $B_c \geq \kappa S$ and $B + C$ is greater than B_c then $C := B_c - B$
- $B := B +$ Coupon, on coupon dates

Notice that the term that multiplies the hazard rate in the right hand side of each PDE expresses the recovery value of each one of the two components after default. For the bond component B, this is the usual term, whereas for the option to convert, or equity component C, this is the intrinsic value of the holder's last option to convert into the residual value of the share.

Interpretation of the T&F model in our framework

We argued earlier that T&F do not provide a justification of their mathematical model in terms of what happens *in effect* to the convertible bond and to its components in case of default. Their splitting is just a heuristic splitting which tries to fulfil at best the desiderata that we have listed above, to the effect that the bond component should capture the cash-flows, fixed and contingent, that the holder is owed, and the equity component should capture his right to convert, etc., only it stops short of telling *the whole, consistent story of default*. What we call "telling the whole, consistent story of default" is that we be able to write PDEs for B and C that govern their respective values prior to default *by way of explicitly stating the outcome of default for these values*. So it certainly would be interesting to try to test T&F's model against our criterion.

The general PDE that the convertible bond value solves in the T&F model is the following:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} = rV + pB$$

It falls in our general schema (1) with $X = B$, and it splits into:

$$\frac{\partial B}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 B}{\partial S^2} + rS\frac{\partial B}{\partial S} = (r + p)B$$

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} = rC$$

with initial conditions:

$$B(S,T) = \begin{cases} N & \text{if } \kappa S \leq N \\ 0 & \text{otherwise} \end{cases}$$

$$C(S,T) = \begin{cases} 0 & \text{if } \kappa S \leq N \\ \kappa S & \text{otherwise} \end{cases}$$

and subject to the following algorithm:

- If $B_p > \kappa S$ and the continuation value of $B + C$ is less than B_p then $B := B_p$ and $C := 0$
- Else if $B_p \leq \kappa S$ and the continuation value of $B + C$ is less than κS then $B := 0$ and $C := \kappa S$
- If $B_c < \kappa S$ then $B := 0$ and $C := \kappa S$
- Else if $B_c \geq \kappa S$ and $B + C$ is greater than B_c then $B := 0$ and $C := B_c$
- $B := B+$ Coupon, on coupon dates

Notice that T&F do not assume that the underlying share drops in the event of default, and that they assume zero recovery.

If we were to recount the consequences of a default event on a convertible bond holder, as this transpires through the T&F model, we would have to admit that he first loses B, *and second*, that he carries on holding the Black–Scholes asset C which is unaffected by default. In other words, life continues after default in the T&F model through the subsequent trading and hedging of the equity component C. See in comparison how life stops in the AFV model in case of default: the holder has to make a last optimal decision, either to exercise the right to convert at residual value or to recover a cash amount. *And notice that both the bond and equity component are subject to default risk in the AFV model:* they both undergo a jump, the bond component to its recovery value, and the equity component to its intrinsic value.

The coupling of pre-default and post-default worlds

Now it would certainly make sense to imagine a continuation of life after default. A softer appellation of the state of default would be "distress regime", and a more general model would be one where the holder may have to reserve until later his decision to convert at the post-default value of the share. Indeed it may not be optimal to exercise the option either to recover the cash value or to convert at residual value of the share, right after default. Cases were witnessed where the conversion ratio was revised after default. Not to mention that the volatility of the underlying share is most likely to have dramatically changed too. Therefore, a more accurate description would be one where the holder ends up holding an "ersatz-convertible bond" in case of default, with bond floor equal to RB, underlying share spot level equal to residual value, possibly a different conversion ratio (provided the issuer agrees to postpone the reimbursement of the recovered value), all of which would have to be priced in a new world (i.e. a new PDE), where volatility may be different, and, last but not least, where credit risk is different. Indeed an open question is whether the post-default world is not default-free. Could a company that

has already defaulted default once again on the recovery value of its previous debt? And if it does, wouldn't that mean that the post-default world itself has to be further coupled with a post-default-post-default world?

Assuming for simplicity that default happens only once, or in other words that the distress regime is default-free, the general convertible bond pricing model we are contemplating may now be expressed in the following terms:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta) S \frac{\partial V}{\partial S} = rV + p[V - V'(S(1-\eta), t)]$$

with the usual convertibility, puttability, and callability constraints:

$$V \geq \max(B_p, \kappa S)$$

$$V \leq \max(B_c, \kappa S)$$

and initial condition:

$$V(S, T) = \max(\kappa S, N)$$

$(V - V')$ is the jump that the convertible bond value undergoes in the event of default, and the jump into default is now generally seen as a case of switching to the distress regime.

Case of no life after default

If life ends with the default event, then $V'(S, t)$ has to assume one of the following "stopped" solutions:

- $V'(S_\tau, \tau) = \max(\kappa S_\tau, RN)$: N-model
- $V'(S_\tau, \tau) = \max\lfloor \kappa S_\tau, RZ(S_{\tau^-})\rfloor$: Z-model
- $V'(S_\tau, \tau) = \max\lfloor \kappa S_\tau, RB(S_{\tau^-})\rfloor$: AFV model

where τ is the time of default and $S_\tau = S_{\tau^-}(1 - \eta)$.

Case of life after default

Otherwise, if V' is allowed to have a life after default, we may write for it the following, default-free, PDE:

$$\frac{\partial V'}{\partial t} + \frac{1}{2}\sigma'^2 S^2 \frac{\partial^2 V'}{\partial S^2} + rS \frac{\partial V'}{\partial S} = rV' \tag{3}$$

Note that the volatility of the share in the distress regime has possibly new value σ'.

Imposing the right constraints and the right initial and boundary conditions on this PDE, will depend on the policy that the issuer wishes to pursue after default.

Suppose he agrees to pay the remaining fraction of coupons and face value at their pre-default payment dates, but grants a conversion option just at the moment of default and not after, then the ersatz-convertible bond V' will solve PDE (3) with:

- the following initial condition:

$$V'(S, T) = RN$$

– the following jump-conditions on coupon dates:

$$V'(S, t^-) = V'(S, t^+) + R \ Coupon$$

– and the following "time of default" constraint:

$$V'(S, \tau) \geq \kappa S_\tau$$

where τ is the time of default.

In other words, we would just have the P-model.

Suppose the issuer extends the conversion option and that he maintains the original scheduling of the interest payments (to be applied now to the recovered fraction). The ersatz-convertible bond V' will solve PDE (3) with, in this case,

– the following initial condition:

$$V'(S, T) = \max(\kappa S, RN)$$

– the following jump-conditions on coupon dates:

$$V'(S, t^-) = V'(S, t^+) + R \ Coupon$$

– and the following continuous constraint:

$$V'(S, t) \geq \kappa S$$

So really the ersatz-convertible bond will behave like a mini-convertible bond in this case, with a new bond floor and initial underlying value equal to the recovery value of the share $S(1 - \eta)$. (We are of course ignoring how the embedded put or call options would fare under the new distress regime). This model really looks like the P-model, only the option either to convert into the recovery value of the share or to recover a fraction of the outstanding straight bond has been given time value.

Suppose the issuer extends the conversion option but doesn't want to postpone the payment of the cash recovery fraction B. V' will now solve PDE (3) with

– the following initial condition:

$$V'(S, T) = \kappa S$$

– the following continuous constraint:

$$V'(S, t) \geq \kappa S$$

– and the following "time of default" constraint:

$$V'(S, \tau) \geq RB(S_{\tau^-}, \tau^-)$$

where τ is the time of default, and B the risky component. This constraint expresses the fact that the holder has to make the optimal decision, at the time of default, whether to accept the recovery cash value RB and end the game, or to go on holding his option to convert in the life after default. However, due to the martingale property of the underlying asset, the solution of PDE (3) with boundary conditions such as described, collapses to:

$$V'(S(1-\eta),\tau) = \max[RB(S,\tau^-), \kappa S(1-\eta)]$$

So really this case would be equivalent to the AFV model.

Finally, suppose that the issuer extends the conversion option after default, only it is an option with a very strange terminal payoff and knock-out barrier.

V' solves PDE (3)

- with initial condition

$$V'(S,T) = \begin{vmatrix} 0 & \text{if } \kappa S \le N \\ \kappa S & \text{otherwise} \end{vmatrix}$$

- and boundary condition:

$V'(S,t) = 0$ for all S and t such that it is optimal for the CB holder to exercise the put
$V'(S,t) = \kappa S$ for all S and t such that it is optimal for the CB holder to convert the bond

As for the fraction B that is lost on default and likely to be partially recovered, it solves the following risky PDE

$$\frac{\partial B}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 B}{\partial S^2} + (r + p\eta) S \frac{\partial B}{\partial S} = (r+p)B$$

Only suppose that it is subject to the following, no less puzzling, initial and boundary conditions:

$$B(S,T) = \begin{vmatrix} N & \text{if } \kappa S \le N \\ 0 & \text{otherwise} \end{vmatrix}$$

and $B(S,t) = 0$ for all S and t such that it is optimal either for the issuer to call the bond, or for the holder to convert it.

Bringing the pieces together, the convertible bond value would be governed by the following PDE

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + p\eta) S \frac{\partial V}{\partial S} = rV + p[V - \max(RB, V'(S(1-\eta),t)]$$

and it would switch to the following PDE in case of default and in case $V'(S(1-\eta),\tau) > RB(S,\tau)$ at the time of default:

$$\frac{\partial V'}{\partial t} + \frac{1}{2}\sigma'^2 S^2 \frac{\partial^2 V'}{\partial S^2} + rS\frac{\partial V'}{\partial S} = rV'$$

where V' is this strange knock-out equity option.

Conclusion: A philosophical refutation of T&F?

When $R = 0$ and $\eta = 0$ the mathematics of the last model becomes identical to T&F. They both give the same value for the convertible bond and its components. All we have done really is interpret the formalism of T&F in our general philosophy where the actual consequences of default are spelled out exactly. An interpretation cannot prove a theory right or wrong. It only gives us arguments to accept it, or to prefer another theory to it. The T&F model is mathematically consistent. It produces the kind of behavior the trader expects from the convertible bond at the extremities of the stock price range. However our general presentation of the various recovery models has convinced us by now that there is much leeway in the choice of model for B, the cash claim of the holder in the event of default.

Much as it seemed legitimate that the cash claim, or the value recovered, in the AFV model should depend on the optimal behavior of the holder in case of no default (remember the case for the cash settlement of interrupted gambling games), we see no reason why it should so dramatically depend on such a hypothetical behavior in the T&F case, as to deny him *any* cash recovery claim in those regions where he would have optimally converted. Even less so do we see the reason why the contingent claim that the holder ends up holding in the life after default should be knocked out in those regions where he would have optimally exercised the put.

T&F would of course object that we are over-interpreting their model. It is only when viewed in the perspective of the post-default world, that the Black-Scholes component C and the cash claim B look so strange! For if one were to stop at the surface, and envision the split into B and C as just a rule for varying the weight of the credit spread in the overall discounting procedure of the convertible bond value, then all that would matter is that the credit spread be applied in the "right places", and this is certainly what T&F achieves! The reason why we feel uncomfortable with this minimalist requirement, however, is that we do not think we can possibly shy away from the consequences of the default event. B is the risky component in the T&F model; hence B is the fraction that I expect to lose in case of default. And if I don't lose everything then I keep something of some value, and then I have to explain why this something has this value. The only explanation is that I can cash in immediately this something through some action (either actual cash, or conversion on the spot), or that this something is just the present value of something that lives through future actions and decisions ...

FOOTNOTES & REFERENCES

1. The delta-weighting approach is a somewhat heuristic, probabilistic expression of the same thing.

2. Alternatively, we can argue, with Wilmott (1998) and Merton (1976): that if the jump component (of the asset price or the convertible bond price) is uncorrelated with the market as a whole, then the risk in the discontinuity should not be priced into the option. Diversifiable risk should not be rewarded. In other words, we can take expectations of this expression and set that value equal to the risk-free return from the portfolio. This is not completely satisfactory, but is a common assumption whenever there is a risk that cannot be fully hedged; default risk is another example of this. (Wilmott 1998, p. 330). (Note: In this paragraph, Wilmott is in the process of deriving the option pricing PDE in the presence of jumps in the underlying; but like he says, the reasoning applies to default risk too). "To be fully satisfactory, the argument

for risk neutrality must presuppose the hedging of default risk. Indeed the derivation would be valid if it were implicit that a credit default swap (CDS), for instance, was being used as a dynamic hedge against default, in conjunction with the underlying share as dynamic hedge against market fluctuations. This would also imply that the risk-neutral probability of default is extracted from the market prices of the CDS."

■ Goldman Sachs (1994). Valuing convertible bonds as derivatives. *Quantitative Strategies Research Notes*, Goldman Sachs.
■ Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press.
■ Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.
■ Takahashi, A., Kobayashi, T. and Nakagawa, N. (2001). Pricing convertible bonds with default risk. *Journal of Fixed Income* 11, 20–29.
■ Tsiveriotis, K. and Fernandes, C. (1998). Valuing convertible bonds with credit risk. *Journal of Fixed Income* 8, 95–102.
■ Wilmott, P. (1998). *Derivatives: The Theory and Practice of Financial Engineering*. Wiley, Chichester.

10
First to Default Swaps

Antony Penaud and James Selfe

Wilmott magazine, January 2003

Default dependence is one of the biggest issues in quantitative finance at the moment. One of the most popular products in which default dependence occurs is the first to default swap. In this short article we are going to review different pricing methodologies for this product.

The product. In a first to default swap (FTDS) the buyer has to pay a fixed rate at each payment date until the default of any of the reference names defined in the swap or until the maturity date of the swap – whichever comes first. In return for this payment the buyer receives a one off payment on the first occurrence of default experienced by any of the names defined in the swap contract. The contract has very similar payoffs to a vanilla credit default swap (CDS). The main difference is in what constitutes an event causing payout for the contract (in one case it is default of a single name and in the other it is the default of any of a list of names).

Pricing via the risk neutral survival curve. The correspondence between the fixed rate of a CDS and the risk neutral survival curve of a single reference name is known (assuming the writer cannot default). Similarly there is a correspondence between the fixed rate of a FTDS and the synthetic survival curve of the first to default. In this paper we are generally concerned with calculating this synthetic first to default survival curve from the single name survival curves. We know from the correspondence above that we can determine a fixed rate for the FTDS if we have this. All we need to do is to find the first to default survival curve

$$S^1(t) = \text{Prob}(\min_i t_i \geq t) \tag{1}$$

where t_i is the default time for asset i (which is in the basket). The difficulty is that there is no unique way to make the first to default survival curve consistent with the single name survival curves that we get from market data. There are in fact an infinite number of candidates for the first to default survival curve.

Bounds for the first to default survival curve. There are two useful bounds for the first to default survival curve that we can consider. Firstly at each time t the minimum of all individual survival probabilities is an upper bound. Secondly if we assume negative correlation between

Contact address: Mitsubishi Securities International plc, 6 Broadgate, London ECZM 2AA
E-mail: antony.penaud@mitsubishi-sec-intl.com

assets is not possible then the product of all individual survival probabilities at time t is a lower bound.

The copula method

The Gaussian copula method (see Li, 2000). Suppose there are α assets in the basket. For each simulation we do the following:

- Simulate α correlated $N(0,1)$ random variables ϕ_i, $i = 1 \ldots \alpha$.
- Find x_i's such that $CND(\phi_i) = x_i$ for $i = 1 \ldots \alpha$.
- Find the default times t_i's by $S_i(t_i) = x_i$ for $i = 1 \ldots \alpha$.
- Find the first default time $t^* = \min_i t_i$.

In the above, *CND* is the cumulative normal distribution function and $S_i(t)$'s are the marginal survival curves. Repeating this procedure allows us to construct the first to default survival curve.

Example 1 Let's assume that there are four assets in the basket and that the four hazard rates are constant: $h_1 = 0.3\%, h_2 = 0.8\%, h_3 = 1.4\%, h_4 = 2\%$. We plot the first to default survival curve for different correlation matrices (Figure 1). For each correlation matrix all correlation coefficients are equal. We plot both the lower bound and the upper bound for the first to default survival curve as well.

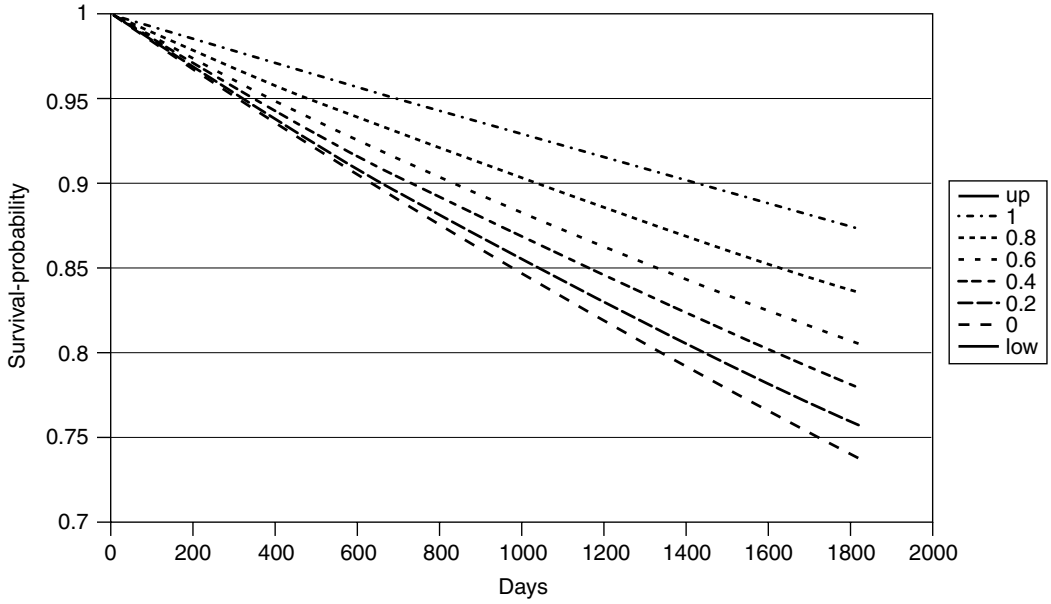


Figure 1: First to default survival curves for different correlation matrices

Note that the effect of the correlation coefficient is to rotate the survival curve. Both bounds are attainable: the upper bound is attained for correlation 1 and the lower bound for correlation

0. Assuming maturity 5 years, zero interest rate, 25% recovery and CDS frequency 4, we can find the ongoing premiums and see the effect of correlation:

Correlation	0	0.2	0.4	0.6	0.8	1
Price	4.5164	4.1451	3.7176	3.2395	2.6803	2.0017

For correlation 1, the theoretical result is 2. Indeed when correlation is 1 the asset with the highest spread always defaults first.

Note as well that the shorter the maturity the higher the ongoing fee for the first to default swap. In the table below we compute first to default ongoing premiums for different maturities (correlation 0.5).

Maturity	1	2	3	4	5
Rate	3.8360	3.6875	3.6116	3.5349	3.4844

In the copula method one doesn't have to use Gaussian distribution. One could use a t-distribution, for instance.

t-copulas (see Mashal and Naldi, 2002). Using the same four marginal survival curves as in the previous paragraph, let's plot (Figure 2) the first to default survival curves produced by Gaussian copula ("Gaussian") and t-copula with degree of freedom 5 ("Student") when correlation coefficients are 0.5. The first to default survival curve produced by t-copula is higher than the one produced by Gaussian copula. We have plotted as well the first to default survival curve produced by the t-copula method in the case of zero correlation ("Student0" on the graph). Note that it is well above the lower bound.

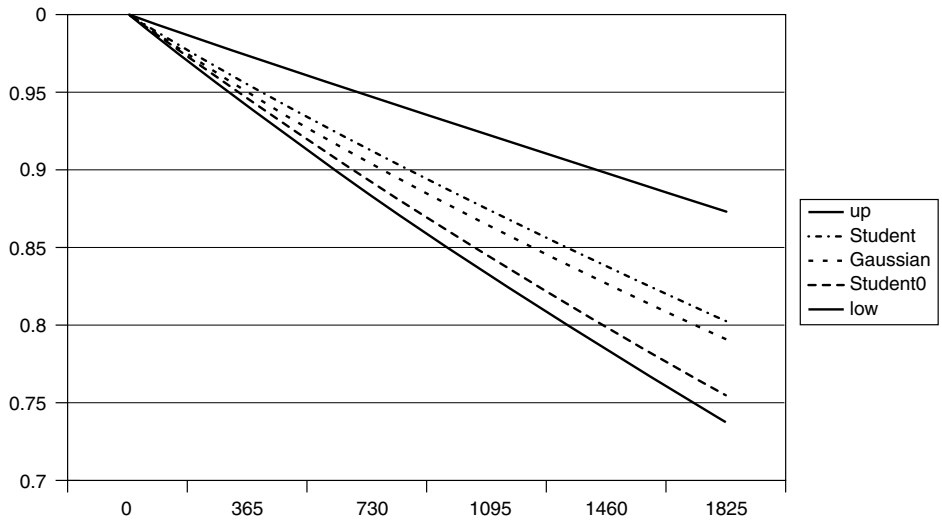


Figure 2: First to default survival curves for Gaussian copula, Student and Gaussian copula with higher correlation

Under the same assumptions as before (correlation 0.5), the ongoing premium for Gaussian copula is 3.4914 and it is 3.2768 for t-copula. The t-copula is computationally more expensive than the Gaussian copula as it requires – for each first default time simulation – the simulation of a chi-square distribution. Furthermore it is the cumulative distribution of the t-distribution that needs to be computed this time.

The Hull–White model

In this model the credit index x of a company follows a standard Brownian motion. Default can only occur at discrete times T_n's, $n = 1, \ldots, N$ where T_N is the maturity of the contract. At each time T_n there's a *wall* and defaults occur when the credit index hits the *wall*. The *height* $B(T_n)$ of the wall is such that the probability for the credit index for not hitting any of the walls up to time T_n is equal to $S(T_n)$, the survival probability of this company at that time. In their paper Hull and White (2000) find $B(T_1), \ldots, B(T_N)$ by inverting integrals. Here we prefer to invert the N backward Kolmogorov partial differential equations below. For $n = 1, \ldots, N$ we're looking for $B(T_n)$ such that the solution $V(0,0)$ (the probability of not hitting any of the future barriers) of the problem below is $V(0, 0) = S(T_n)$:

$$V_t + \frac{1}{2} V_{xx} = 0 \tag{2}$$

with boundary conditions:

$$\forall\ m < n, V(T_m, x) = 0 \text{ for } x \leq B(T_m) \tag{3}$$

and final condition:

$$V(T_n, x) \quad \begin{cases} 1 & \text{if } x > B(T_n) \\ 0 & \text{otherwise.} \end{cases}$$

We do not want N to be too large as computation time would be too long and it could introduce numerical problems near time 0 (for N very large we almost have a continuous default barrier – for which the slope at time zero is minus infinity). Once we have found the $B(T_n)$'s we find the first to default survival probabilities by simulating correlated discrete Brownian motions. The first to default survival curve is found by loglinear interpolations and the pdf for first default time is therefore slightly affected.[1] In terms of computation time, for each first default simulation we need to simulate $N \times \alpha$ normal variables (sometimes less when default occurs and we're not interested in knowing the 2nd to default survival curve).

In Figure 3 we plot the Hull–White barriers as well as the thresholds of the Gaussian copula method corresponding to spread 3% and recovery rate 25%.

In this model the correlation has a better intuitive interpretation and evaluation of the correlation inputs is easier to do. In the Hull–White model correlation can be a function of time too.

Example 2 Let's find the ongoing premiums for both models. We are going to price different first to default swaps. The base scenario is maturity 5 years, annual frequency, correlation 0.5, 4 assets with constant spreads 30 bps, 80 bps, 140 bps and 200 bps, zero interest rate. Each scenario (from 1 to 10) has only one input differing from the base scenario. For scenario 1, correlation coefficients are 1, scenario 2 correlation coefficients are 0, scenario 3 interest rate is 10%, scenario 4 maturity is 10 years, scenario 5 maturity is 1 year, scenario 6 all four spreads

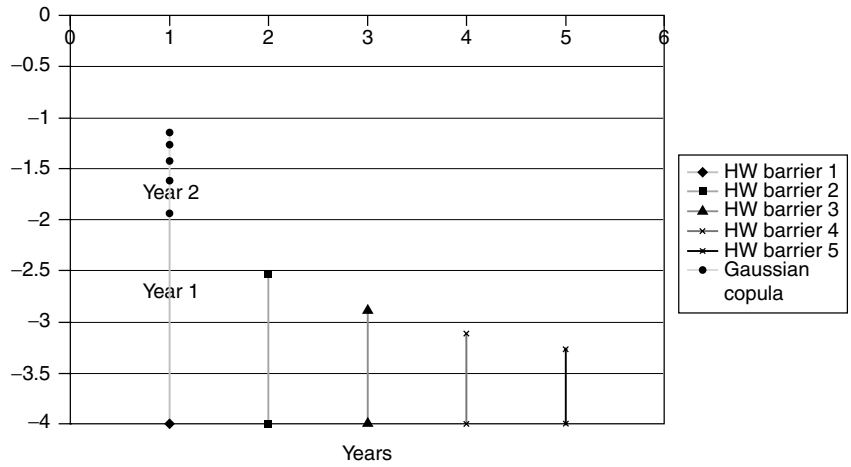


Figure 3: Hull–White walls and Gaussian copula thresholds

are 20 bps, scenario 7 all four spreads are 500 bps, scenario 8 frequency is biannual, scenario 9 there are 10 assets and the spreads are 30 bps, 30 bps, 50 bps, 50 bps, 80 bps, 80 bps, 100 bps, 100 bps, 120 bps and 120 bps. For scenario 10 the hazard rate is not constant and the term hazard rate is in the Appendix.

	Copula	HW	Copula-HW	(C-HW)/HW	Scenario
base	3.517	3.596	−7.9%	−2.197%	
1	2	2	0%	0%	correlation 1
2	4.585	4.583	0.2%	0.04%	correlation 0
3	3.56	3.621	−6.1%	−1.68%	high interest rate
4	3.343	3.416	−7.3%	−2.13%	long maturity
5	3.9	3.894	0.6%	0.15%	short maturity
6	0.666	0.684	−1.8%	−2.63%	low spreads
7	13.604	13.952	−34.8%	−2.49%	high spreads
8	3.496	3.563	−6.7%	−1.88%	high frequency
9	4.473	4.659	−18.6%	−3.99%	many assets
10	5.01	5.084	−7.4%	−1.45%	non-constant hazard rate

For scenarios 1, 2 and 5 both models give the same result. When correlation is 1 (scenario 1) the two models give the same result because in our example $\max_i B_i(T_n)$ is attained by the same asset for each time T_n. For correlation 0 the result is true in general. For scenario 5 both models give the same result because we have used $N = 1$, and in this case both models are equivalent.

As for the copula method, the Hull–White method can be used with non-Gaussian distributions.

Summary and conclusion

There are different methods to go from a set of single name survival curves to a multiple asset survival curve that can be used to price an Nth to default swap. We have reviewed briefly and compared results for three of these methods (a) Gaussian copula (b) t-distribution copula and (c) Gaussian Hull–White method. We have shown that Gaussian copula and Gaussian Hull–White give, over a range of scenarios, rather similar results and in some cases by construction identical results. The Gaussian copula method is easy to implement and requires less computational time than the Hull White method. Therefore for practical purposes the Gaussian copula model is the more attractive to implement of these two.

The Gaussian distribution has thin tails compared to other distributions. Equity and other markets have been shown by many studies to have fatter tails than those implied by the normal distribution. As we are measuring events of default that are by nature tail events this is of some concern to us. An alternative is to use a distribution with fat tails such as the Student's t distribution. We note that this change in assumption changes our results more significantly.

Appendix

Spreads for scenario 9.

	Asset 1	Asset 2	Asset 3	Asset 4
1 year	0.5	4	1	0.5
2 years	1	3.5	1	0.5
3 years	1.5	3	1	0.5
4 years	2	2.5	1	0.5
5 years	2.5	2	1	0.5

Acknowledgement

The authors would like to thank Michal Tomek and YanMin Li for helpful comments. The opinions expressed are those of the authors and not necessarily those of Mitsubishi Securities International plc.

FOOTNOTE & REFERENCES

1. The loglinear interpolation makes the average default time larger.

■ Hull, J. and White, A. (2000) Valuing credit default swaps II: modeling default correlations, *Journal of Derivatives* 8(3), Spring 2001. Also available on www.defaultrisk.com
■ Li, D. (2000), On default correlation: a copula function approach. www.riskmetrics.com
■ Mashal, R. and Naldi, M. (2002) Pricing multiname credit derivatives: heavy tailed hybrid approach. www.defaultrisk.com

11
Taken to the Limit: Simple and Not-so-simple Loan Loss Distributions

Philipp J. Schönbucher

Wilmott magazine, January 2003

In his influential papers, Vasicek (1987, 1997) showed that in a simplified multi-obligor version of the Merton (1974) credit risk model, the distribution of the losses of a large loan portfolio can be described by the inverse Gaussian distribution function. In his setup, the probability that the fraction L of defaults in the portfolio is less than a given level q is given by

$$\mathbf{P}[L \leq q] = \Phi\left[\frac{1}{\sqrt{\varrho}}(\sqrt{1-\varrho}\ \Phi^{-1}(q) - \Phi^{-1}(p))\right] \tag{1}$$

where p is the default probability of any individual obligor in the portfolio, and ϱ is the asset value correlation between any two obligors ($\Phi(\cdot)$ denotes the cumulative standard normal distribution function).

Usually the loss distribution of a credit risk model can only be determined using lengthy numerical simulations, thus a simple closed-form solution like (1) which involves just two parameters has a lot of appeal: It can be very useful to understand the behaviour of the more complex variants of the model, to find benchmark parameter values that can be fitted to historical observations, or simply as a "quick and dirty" first approximation in all situations when setting

Contact address: Eidgenössische Technische Hochschule Zürich, D-MATH, ETH Zürich, CH 8092 Zurich, Switzerland. E-mail: philipp.schoenbucher@math.ethz.ch http://www.schonbucher.de

This paper was written while the author was at the Department of Statistics, Bonn University.
Financial assistance by the DFG is gratefully acknowledged. Comments and suggestions are welcome. All errors are my own.

up a full credit risk model would take too long. Furthermore, as the Credit Metrics model (Gupton *et al*., 1997) is also based upon the Merton (1974) firm's value setup, Vasicek's result can also be regarded as a limiting case of this very popular model, and many of the qualitative features of the Credit Metrics model can be analysed in closed-form without having to resort to the usual lengthy simulations. The accuracy of the approximation is remarkably good, the approximation error becomes unacceptable only when very low asset value correlations $\varrho < 1\%$, very few obligors (<20) or extremely heterogeneous exposure sizes (e.g. one dominating obligor) are considered.

Thus it is no surprise that the Vasicek model has been quickly adopted in practice, large portfolios are managed on the basis of (1) and the relationship is also used in a regulatory context to set risk measures for credit exposures.

Despite its widespread use, the Vasicek model does have some shortcomings beyond the obvious over-simplification of identical default risks and exposure sizes of the obligors. For instance, there are significant difficulties replicating the *qualitative* shape of the loss distribution. For a given default probability parameter p, one can only vary ϱ to fit both the main part of the loss distribution and the tail of the distribution. If a manager of a CDO wants to calibrate ϱ in such a way that tails of the distribution are fitted well (i.e. such that the senior and super-senior tranche of his CDO are priced correctly), then he may experience serious mispricing of the main body of the distribution (i.e. the mezzanine and equity tranches). Furthermore, the shape of the distribution changes significantly when different time horizons (and thus default probabilities) are considered. Both problems have their roots in the implicit assumption of a joint Gaussian distribution of the obligors' asset value processes which imply a very specific transition from the limiting case of independence ($\varrho = 0$, all probability at $L = p$) to the fully dependent case ($\varrho = 100\%$, all probability at $L = 0$ and $L = 100\%$).

In this paper we will give a class of similar approximative loss distributions of large portfolios where we do not use multivariate normally distributed random variables to trigger defaults. Instead we model the dependencies between the defaults using *Archimedean copula functions*. Archimedean copula functions are a tractable class of joint distribution functions with characteristics that can be significantly different from the characteristics of a multivariate normal distribution function.

The aim of this exercise is to analyse the effects that the specification of the dependency structure between the individual default events can have on the loss distribution of the whole portfolio. Therefore we can use a very stylised model and we can keep all other parameters fixed, such as individual default probabilities, exposure size, and even the pairwise default correlations, in order to isolate the effects of the dependency structure.

The main tool that we use to achieve this result is an algorithm for the generation of random variates with a given Archimedean copula function as joint distribution function which was first proposed by Marshall and Olkin (1988). This algorithm in itself may already be a valuable tool for the reader, because a major obstacle against the wider adoption of alternative dependency models in risk management to date was the lack of efficient numerical implementation schemes for large-scale simulations.

In the next section we will introduce the notion of a copula function and motivate why it is a good idea to check the results of a Gaussian model against alternative specifications of the dependency structure. Then we will introduce the special class of Archimedean copulae and give the algorithm to generate random variates whose dependency structure is described by an Archimedean copula. We specify the algorithm for the Clayton, the Gumbel and the Frank

copulae, three special cases where the data generating process can be given in closed-form. Then, we enter the portfolio credit risk model and derive closed-form formulae for the loss distribution of a large loan portfolio whose defaults are driven by random variables with this type of dependence. The numerical implementation and comparison of these loss distributions will show that the nature of the dependency structure does have a significant effect on the loss distribution, even if the default correlation between any two obligors is held fixed.

The literature on dependency modelling and copula functions has grown substantially in recent years, therefore we cannot give a full overview here. The reader is referred to the excellent (but slightly technical) textbook by Joe (1997) for the basics; another popular textbook is Nelsen (1999). The standard reference for the generation of non-uniform random variates is Devroye (1986) which contains hundreds of algorithms. For the application of copula functions to credit risk modelling we should mention Li (2000) and Schönbucher and Schubert (2001) and Frey and McNeil (2001). Frey and McNeil (2001) give an overview of the different approaches to portfolio credit risk modelling with a particular focus on the dependency structures implied by the models. They also analyse the loss distribution of large loan portfolios in the more general setup of Bernoulli Mixture models, and compare the differences between a Gaussian (e.g. Credit Metrics) and t-Copula dependency structure.

Copula functions and Laplace transforms

Copula functions

Whenever several dependent dimensions of uncertainty have to be modelled, the standard (and often also the only) approach is to somehow transform the problem in such a way that a multivariate normal distribution can be used to model the uncertainty. In the modelling of equities, exchange rates and interest-rates, multivariate lognormal distributions are used (i.e. exponentials of normals), squared Gaussian and related models are also popular, and in portfolio credit risk modelling the Credit Metrics model is driven by a multivariate normal distribution of the obligor's asset value processes.

Using a multivariate normal distribution as driver of the model leads to a so-called *Gaussian dependency structure* of the key variables of the resulting model. This can be a restrictive modelling choice; it is just one out of an infinite number of possible joint distribution functions. The full set of all possible dependencies between I random variables is given by the set of all I-dimensional copula functions.

So what is a copula? Roughly speaking:

> An I-dimensional copula is a distribution function on $[0, 1]^I$ with uniform marginal distributions.

That is all. Copulas concentrate on the dependency, so the marginal distribution is irrelevant. It is set to a uniform distribution because this makes the later incorporation of other marginal distributions straightforward, and we recover the benchmark case of the uniform distribution on [0, 1] if we ignore the other $I - 1$ random variables.

The technical definitions of copulas that are given in the mathematical literature often look quite different, but to a financial modeller, this is the definition to build an intuition from. The reason why copulae provide a useful framework to analyse dependencies between

random variables is the fact that to every multivariate distribution function there is a copula which contains all information on dependence. This is the essence of the following theorem by Sklar:

Theorem 1 (Sklar). *$X_1, \ldots, X_I$ with marginal distribution functions $F_1, F_2, \ldots, F_I$ and joint distribution function F. Then there exists an I dimensional copula C such that:*

$$F(x_1, \ldots, x_I) = C(F_1(x_1), F_2(x_2), \ldots, F_N(x_I)) \qquad \forall \boldsymbol{x} \in \mathbb{R}^I,$$
$$C(u_1, \ldots, u_I) = F(F_1^{[-1]}(u_1), \ldots, F_1^{[-1]}(u_I)).$$

If $F_1, F_2, \ldots, F_N$ are continuous, then C is unique.

In particular, the copula $C(\cdot)$ is the distribution function of the transformed random variables $U_1 = F_1(X_1), \ldots, U_I = F_I(X_I)$.

So, to every distribution function on $\mathbb{R}^I$, there is a corresponding copula function. For example, if the random variables X_i are independent, then the *independence copula* is just the product of the u_i

$$C(u_1, \ldots, u_I) = u_1 \cdot u_2 \cdot \ldots \cdot u_I.$$

If $X_1, \ldots, X_I$ have a multivariate normal distribution with covariance matrix Σ and mean zero (for simplicity), then the *Gaussian copula* is reached:

$$C(x_1, \ldots, x_I) = \Phi_{\Sigma,0}(\Phi^{[-1]}_{\sigma^2_{11}}(x_1), \ldots, \Phi^{[-1]}_{\sigma^2_{II}}(x_I)),$$

where $\Phi_{\sigma^2}()$ is the univariate cumulative normal distribution function with variance σ^2 and mean zero, and Φ_Σ the multivariate cumulative normal distribution function with covariance matrix Σ.

As the next section will show, there are even more possibilities.

Archimedean copula functions

Copula functions do not impose any restrictions on the model at all, so in order to reach a model that is to be useful in practical applications, a particular specification of the copula must be chosen. As we want to provide an alternative to the Gaussian model, we use the Archimedean copula functions as a benchmark model.

Definition 1 (Archimedean copula)

(i) An Archimedean copula function $C : [0, 1]^I \to [0, 1]$ is a copula function which can be represented in the following form:

$$C(\mathbf{x}) = \phi^{[-1]}\left(\sum_{i=1}^{I} \phi(x_i)\right), \tag{2}$$

with a suitable function $\phi : [0, 1] \to \mathbb{R}_+$ with $\phi(1) = 0, \phi(0) = \infty$.

(ii) The function $\phi : [0, 1] \to \mathbb{R}_+$ is called the generator of the copula.

Not every function ϕ is a suitable generator for a copula function; there are restrictions on the signs of the derivatives of ϕ which become more stringent with increasing dimension I. But in the following case the existence of the copula can be ensured:

If $F(x)$ is a distribution function of a positive random variable with $F(x = 0) = 0$ and

$$\hat{F}(y) = \int_0^\infty e^{-yx} dF(x)$$

is its Laplace transform, then $\phi(t) := \hat{F}^{[-1]}(t)$ is the generator of a Archimedean copula of dimension I for every $I > 0$. (In fact, $\phi^{[-1]}()$ must be a Laplace transform if it is to be an admissible generator for all dimensions $I > 0$.)

From equation (2) we can see that Archimedean copula models are *exchangeable*, i.e. the dependency between any two (or i) different risk factors does not depend on the question *which* two (or i) risk factors were chosen. For our aim of assessing portfolio credit risk in large, homogeneous portfolios this does not pose a major restriction, in fact it is a desirable property. (For other applications this may not be the case.)

In Table 1 we give some popular specifications of the generator functions ϕ and their inverses $\phi^{[-1]}$, together with the inverse Laplace transform of the inverse generator $\psi(s) = \mathcal{L}^{[-1]}_{\phi^{[-1]}}(s)$. We will need Laplace transform and inverse Laplace transforms later on, so this may be a good place to define them:

TABLE 1: SOME GENERATORS FOR ARCHIMEDEAN COPULAS, THEIR INVERSES AND THEIR LAPLACE TRANSFORMS. SOURCE: MARSHALL AND OLKIN (1988)

1.	Name: Clayton $\phi(t) = (t^{-\theta} - 1)$ $\phi^{[-1]}(s) = (1 + s)^{-1/\theta}$ Parameter: $\theta \geq 0$ γ-Distribution: Gamma $(1/\theta)$ Density of γ : $\dfrac{1}{\Gamma(1/\theta)} e^{-y} y^{(1-\theta)/\theta}$
2.	Name: Gumbel $\phi(t) = (-\ln t)^{\theta}$ $\phi^{[-1]}(t) = e^{(-s^{1/\theta})}$ Parameter: $\theta \geq 1$ γ-Distribution: α-stable, $\alpha = 1/\theta$ Density of γ: (no closed-form is known)
3.	Name: Frank $\phi(t) = -\ln \dfrac{e^{-\theta t} - 1}{e^{-\theta} - 1}$ $\phi^{[-1]}(t) = -\dfrac{1}{\theta} \ln[1 - e^{-s}(1 - e^{-\theta})]$ Parameter: $\theta \in \mathbb{R} \backslash [0]$ γ-Distribution: Logarithmic series on $\mathbb{N}_+$ with $\alpha = (1 - e^{-\theta})$ Distribution of γ: $\boldsymbol{P}[\gamma = k] = \dfrac{-1}{\ln(1-\alpha)} \dfrac{\alpha^k}{k}$

Definition 2 (Laplace transform) Let Y be a non-negative random variable with distribution function $G(y)$ and density function $g(y)$ (if a density exists). Then:

(i) The Laplace transform of Y is defined as:

$$\mathcal{L}_Y(t) := \mathbf{E}[e^{-tY}] = \int_0^\infty e^{-ty} dG(y) = \int_0^\infty e^{-ty} g(y)\, dy =: \mathcal{L}_g(t), \quad \forall t \geq 0. \tag{3}$$

(ii) Let $\psi : \mathbb{R}_+ \to [0, 1]$. If a solution exists, the inverse Laplace transform $\mathcal{L}_\psi^{[-1]}$ of ψ is defined as the function $\chi : \mathbb{R}_+ \to [0, 1]$ which solves:

$$\mathcal{L}_\chi(t) = \int_0^\infty e^{-ty} \chi(y)\, dy = \psi(t), \qquad \forall t \geq 0.$$

(iii) The distribution of Y is uniquely characterized by its Laplace transform.

Generation of copula-dependent random numbers

Despite the importance of an accurate model for the dependency structure of the returns of the assets in a portfolio, an obstacle for practical implementation of any copula-based model was the absence of an efficient method for generating copula-dependent random variates. These dependent random variates are essential for the simulation of the portfolio's risk/return profile, and also for the development and testing of estimation methods for the parameters of these distributions.

The most frequently used method is the conditional distributions method which involves a differentiation step for each dimension of the problem. For this reason it is not practical in dimensions larger than ten. As an alternative we propose a method which is based upon the representation of a large class of copula functions with Laplace transforms and mixtures of powers as described in Joe (1997).

Our strategy for the sampling of a random vector X with the distribution function above is the following algorithm by Marshall and Olkin (1988).

Proposition 1 (Marshall/Olkin 1988). Let $\phi^{[-1]} : \mathbb{R}_+ \to [0, 1]$ and $\phi : [0, 1] \to \mathbb{R}_+$ be continuous, strictly decreasing functions. Follow the following algorithm:

(a) Draw $U_1, \ldots, U_I$ i.i.d. uniformly distributed on $[0, 1]$.
(b) Draw the mixing variable Y with the following properties:
- We call Y's distribution function G (and its density g if a density exists).
- Y is independent of $U_1, \ldots, U_I$
- Y's Laplace transform is $\phi^{[-1]}(\cdot)$

$$\mathcal{L}_G(s) := \mathbf{E}[e^{-sY}] = \int_0^\infty e^{-sy} dG(y) = \phi^{[-1]}(s). \tag{4}$$

(c) Define

$$X_i := \phi^{[-1]}\left(-\frac{1}{Y} \ln U_i\right) \qquad 1 \leq i \leq I. \tag{5}$$

Then the joint distribution function of the X_i, $1 \leq i \leq I$ is

$$\mathbf{P}[\mathbf{X} \leq \mathbf{x}] = \phi^{[-1]}\left(\sum_{i=1}^{I} \phi(x_i)\right),$$

the X_i have the Archimedean copula function with generator $\phi(\cdot)$ as distribution function.

From (4) follows that the density of Y is the inverse Laplace transform of $\phi^{[-1]}$. In many cases the distribution of Y can already be identified by looking at $\mathcal{L}^{[-1]}(\phi^{[-1]})$ and an efficient simulation algorithm may already be available. Otherwise Y can also be generated using a uniform random variable V as follows:

$$Y := G^{[-1]}(V) \qquad \text{where} \quad G = \mathcal{L}^{[-1]}(\phi^{[-1]}).$$

Proof First note that:

$$\mathbf{P}[X_i \leq x_i | Y] = \exp\{-\phi(x_i)Y\},$$

and that the unconditional distribution function of Y is G. The claim of the proposition follows by using iterated expectations:

$$\begin{aligned}
\mathbf{P}[\mathbf{X} \leq \mathbf{x}] &= \mathbf{E}\left[\prod_{i=1}^{I} \mathbf{P}[X_i \leq x_i | Y]\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{I} \exp\{-\phi(x_i)Y\}\right] \\
&= \mathbf{E}\left[\exp\left\{-Y\sum_{i=1}^{I} \phi(x_i)\right\}\right] \\
&= \mathcal{L}_G\left(\sum_{i=1}^{I} \phi(x_i)\right) = \phi^{[-1]}\left(\sum_{i=1}^{I} \phi(x_i)\right). \qquad (6)
\end{aligned}$$

The key point about the algorithm shown above is that *conditional on the realisation of Y, the random variables X_i are independent*. This conditional independence property was exploited in the proof of the algorithm, and it will also drive the results in the credit risk model.

The portfolio credit risk model

We now have a recipe (a set of recipes) to generate a set of I dependent random variables with uniform marginal distributions. Let us use this recipe to define a simple portfolio default risk model. The model setup is as follows:

Assumption 1 (Finite Portfolio):

- There are I obligors, we consider defaults up to a fixed time-horizon T.

- All obligors have the same exposure size and the same loss in default. Thus, the number D of defaults is sufficient to determine the loss of the portfolio.
- Obligor i has the default probability p_i until T.
- Obligor i defaults, if and only if $X_i \leq p_i$, where X_i is generated by the algorithm of proposition 3.1.

The loan loss distribution for a finite portfolio

In this setup, the loan loss distribution can be easily derived by conditioning on the mixing variable Y. *Conditional* on $Y = y$, the default probability of an obligor i is:

$$\begin{aligned} p_i(y) := \mathbf{P}[X_i \leq p_i | Y = y] &= \mathbf{P}\left[\phi^{[-1]}\left(-\frac{1}{y}\ln U_i\right) \leq p_i\right] \\ &= \mathbf{P}\left[-\frac{1}{y}\ln U_i \geq \phi(p_i)\right] = \mathbf{P}[\ln U_i \leq -y\phi(p_i)] \\ &= \mathbf{P}[U_i \leq \exp\{-y\phi(p_i)\}] = \exp\{-y\phi(p_i)\} \end{aligned}$$

If all obligors have the same unconditional default probability $p = p_i, \quad \forall i \leq I$, then $p(y) = \exp\{-y\phi(p)\}$ and the probability of k defaults in the portfolio is

$$\mathbf{P}[D = k] = \int_0^\infty \binom{I}{k} p^k(y)(1 - p(y))^{I-k} G(dy). \tag{7}$$

The large portfolio approximation

Assumption 2 (Large Portfolio) In addition to assumption 1 we assume:

- All obligors have the same unconditional default probability p.
- The number of obligors I is very large ($I \to \infty$), the relevant quantity for the portfolio risk is the fraction L of defaulted obligors in the portfolio.

By the law of large numbers, the fraction L of defaults will almost surely be $p(y)$ in the limit of the very large portfolio, whenever the mixing variable Y has taken the value of y. Thus, the probability of having more than a fraction q of defaults in the portfolio is:

$$\begin{aligned} \mathbf{P}[L \leq q] = \mathbf{P}[p(Y) \leq q] = \mathbf{P}[\exp\{-Y\phi(p)\} \leq q] &= \mathbf{P}[-Y\phi(p) \leq \ln q] \\ = \mathbf{P}\left[-Y \leq \frac{\ln q}{\phi(p)}\right] = \mathbf{P}\left[Y \geq -\frac{\ln q}{\phi(p)}\right] &= 1 - G\left(-\frac{\ln q}{\phi(p)}\right). \end{aligned}$$

The distribution F and the density f of the limiting loss distribution are thus

$$F(q) = 1 - G\left(-\frac{\ln q}{\phi(p)}\right), \tag{8}$$

$$f(q) = \frac{1}{q\phi(p)} g\left(-\frac{\ln q}{\phi(p)}\right), \tag{9}$$

where we assumed in (9) that the mixing variable has a density.

Some examples

Armed with the large portfolio loss distribution functions (8) and (9), we can now analyse the effects of using different dependency specifications (i.e. different copulae, generated by different generator functions ϕ) and compare them to the standard Gaussian specification used by Vasicek.

We only compare the Clayton and the Gumbel model to the Gaussian case, and excluded the Frank copula. This was done because the mixing variable in the Frank copula does not have a density (it is integer-valued), and because the main effects should already become clear with these two comparison cases.

We use the following case as our benchmark:

Assumption 3 (Benchmark Case):

- The individual default probability is $p = 5\%$.
- The linear correlation between two default events is $\rho = 10\%$.

The Vasicek set-up

First, we define our benchmark case, the Gaussian model used by Vasicek.

Assumption 4 (Vasicek model):

(1) The default of each obligor i is triggered by the realisation of the value V_i of the assets of its firm.
(2) V_i is normally distributed. Without loss of generality[1], the V_i are standardised $V_i \sim \Phi(0, 1)$.
(3) Obligor i defaults if its firm's value V_i is below a barrier K, i.e. if $V_i \leq K$. K is chosen such that the individual default probability p is matched: $p = \Phi(K)$.
(4) The values of the assets of the obligors are driven by: one common factor Y, and an idiosyncratic standard normal noise component ε_i

$$V_i(T) = \sqrt{\varrho}\, Y + \sqrt{1 - \varrho}\, \varepsilon_i \qquad \forall i \leq I,$$

where Y and $\varepsilon_i, i \leq I$ are i.i.d. $\Phi(0, 1)$-distributed.

Again, conditional on the realisation of the systematic factor Y, the firm's values and the defaults are *independent*, only now the default risk enters additively as a systematic factor Y in the evolution of the firm's asset values. The *asset correlation* between two obligors is $\varrho = \mathbf{E}[V_i(T)V_j(T)]$. This model is very similar to the JPMorgan Credit Metrics model; it can be transformed to our copula-setup by defining:

$$X_i := \Phi(V_i) \qquad \text{and} \qquad p := \Phi(K).$$

Thus, the Vasicek model can be written as a copula model with a Gaussian copula function. The resulting loss distribution in the large portfolio setup is:

$$F(q) := \mathbf{P}[L \leq q] = \Phi\left(\frac{1}{\sqrt{\varrho}}(\sqrt{1 - \varrho}\ \Phi^{-1}(q) - \Phi^{-1}(p))\right).$$

The probability density function $f(q)$:

$$f(q) = \sqrt{\frac{1-\varrho}{\varrho}} \exp\left\{\frac{1}{2}(\Phi^{-1}(q))^2 - \frac{1}{2\varrho}(\Phi^{-1}(p) - \sqrt{1-\varrho}\ \Phi^{-1}(q))^2\right\}.$$

Figure 1 shows the limiting large portfolio loss distribution for various values of the asset correlation. In order to be able to distinguish the tail behaviour, the figure is also shown in log-scale. For relatively small values of the asset correlation, the loss distribution is peaked

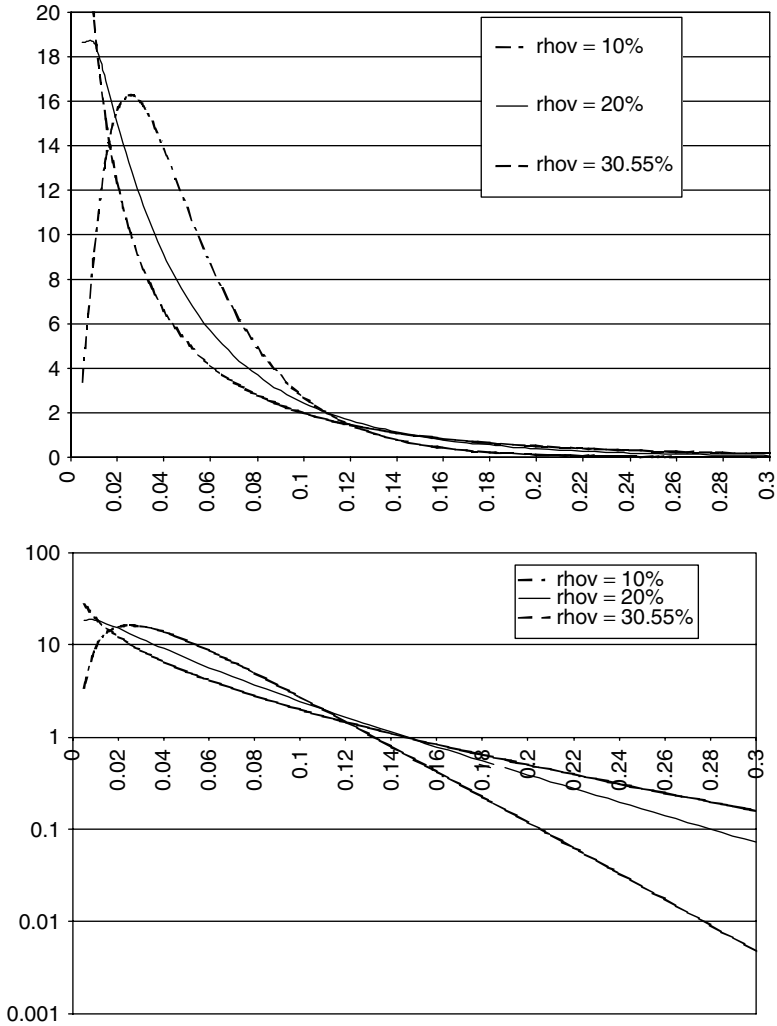


Figure 1: Loan loss distributions for the Vasicek model, $p = 5\%$, various asset correlations ($\varrho_V = 10\%, 20\%, 30, 55\%$)

around the average default rate 5%, but with increasing asset value correlation (and thus default correlation), the distribution skewed and has a single (but finite) peak at zero. This is a "balancing" effect that compensates the shift of probability mass into the tail of the loss distribution. Figure 2 shows the development of the loss distribution if the asset correlation is increased to extremely large levels. In extreme cases, the distribution exhibits a second peak at $q = 1$, i.e. 100% losses. The loss distribution approaches the scenario with the highest default dependency, this is the scenario in which either all firms default (with 5% probability), or none (with 95% probability).

The Clayton copula

The density of the loss distribution in the Clayton copula model is given by (9), where $\phi()$ is the generator of the Clayton copula and $g()$ the density of a gamma distribution with parameter $1/\theta$.

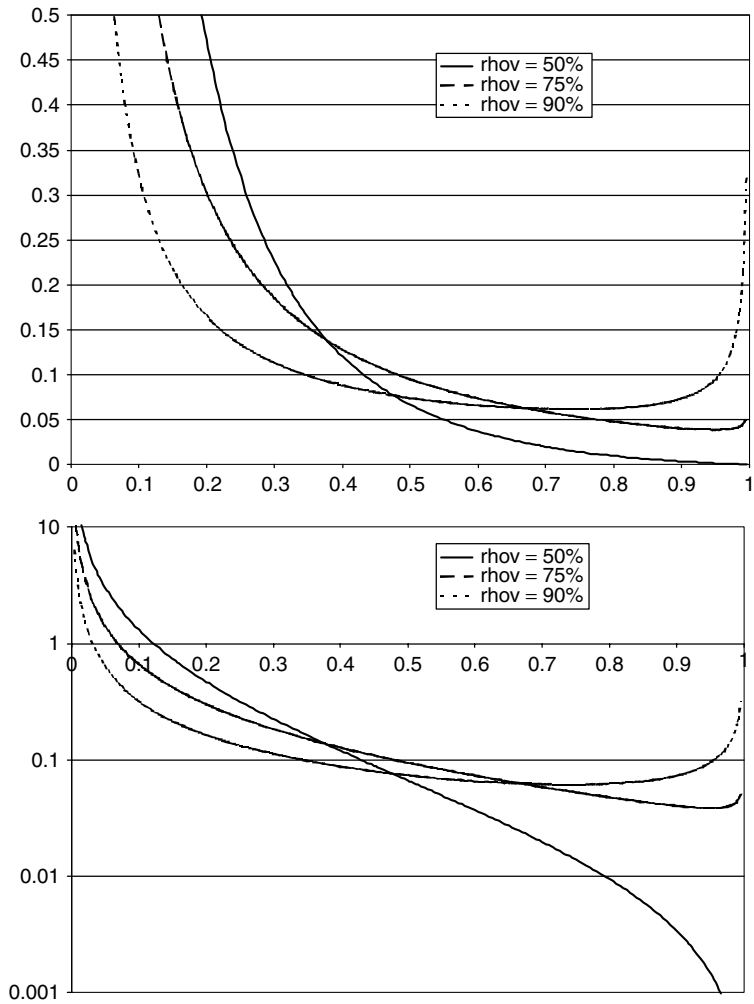


Figure 2: Loan loss distributions for the Vasicek model, $p = 5\%$, various large asset correlations ($\varrho_V = 50\%, 75\%, 90\%$)

Figures 3 and 4 show the loan loss distributions in the Clayton model for small ($\theta = 5\%$, 10%, 18.12%) and large ($\theta = 50\%$, 100%, 200%) parameter values. (The parameter θ in the Clayton model is unbounded.) The qualitative behaviour of the model resembles that of the Gaussian/Vasicek model: again we have a single peak around the average loss rate in the cases of low default dependency, the distribution shifts to skewed distribution with a single peak at $q = 0$ as the dependency is increased, and finally, for very large dependencies, we approach the extreme case with a second peak at 100% defaults.

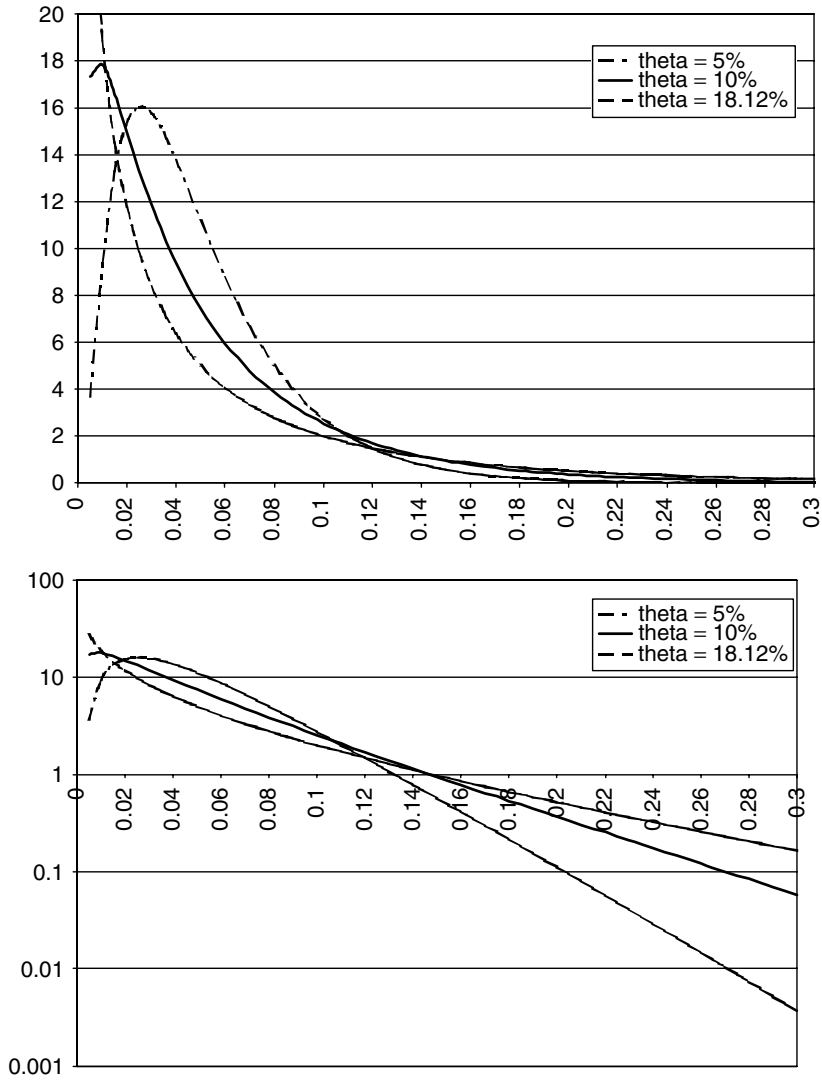


Figure 3: Loan loss distributions for the Clayton model, $p = 5\%$, parameter ($\theta = 5\%$, 10%, 18.12%). ($\theta = 0$ corresponds to independence, dependence increases with θ)

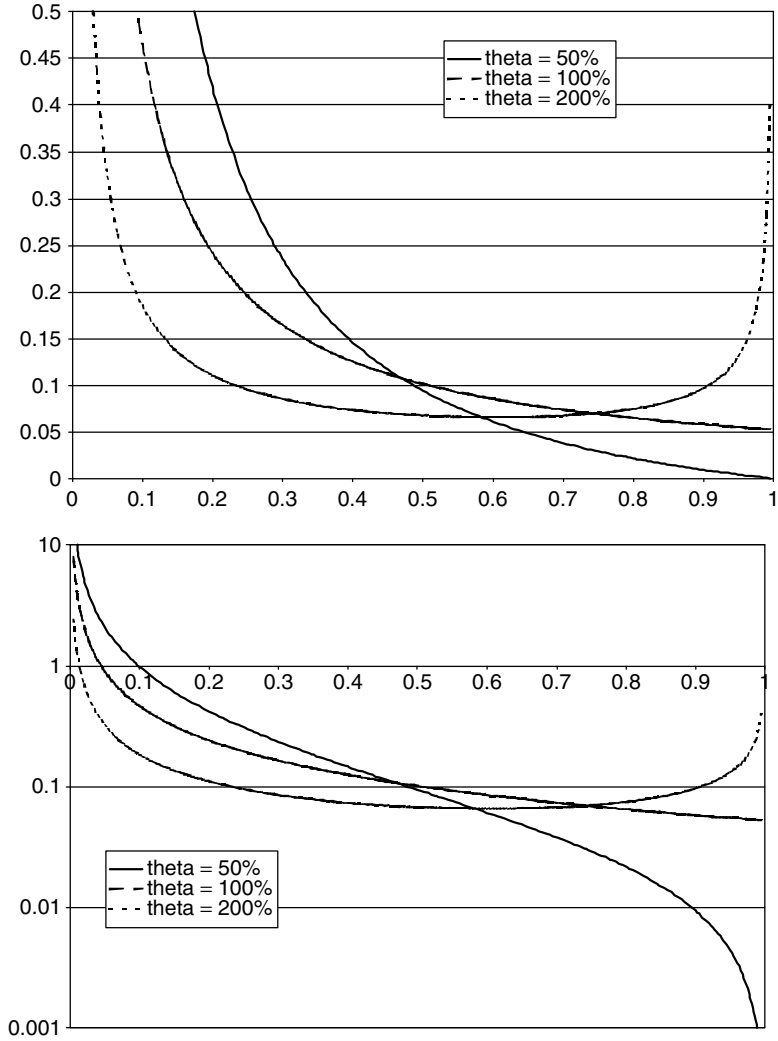


Figure 4: Loan loss distributions for the Clayton model, $p = 5\%$, large dependency parameter ($\theta = 50\%, 100\%, 200\%$)

The Gumbel copula

For the Gumbel copula we again must substitute in (9), but now $\phi()$ is the generator of the Gumbel copula and $g()$ the density[2] of a α-stable distribution with parameter $\alpha = 1/\theta$.

Figures 5 and 6 show the loan loss distributions in the Gumbel model. We see a markedly different behaviour from the Gaussian and the Clayton case. This behaviour is driven by the special properties of the Gumbel copula: The Gumbel copula exhibits strong *upper tail dependency*. This means that – even if linear correlations or similar measures look the same – the Gumbel distribution implies far more joint extreme events than for example the Gaussian copula

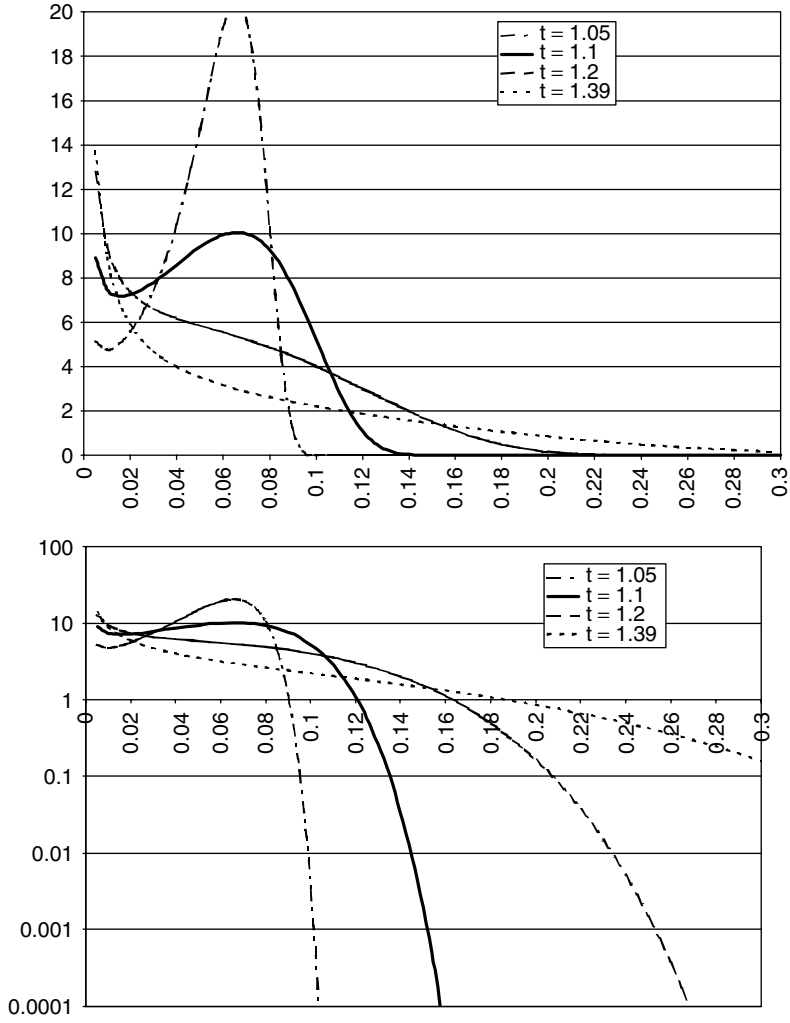


Figure 5: Loan loss distributions for the Gumbel model, $p = 5\%$, parameter ($\theta = 1.05, 1.1, 1.2, 1.39$). ($\theta = 1$ corresponds to independence, dependence increases with θ)

which has no tail dependency. (The Clayton copula exhibits lower tail dependency, i.e. extreme movements only cluster in one direction.)

Figure 5 shows very clearly the implications of the strong tail dependency: as we move away from the independence case ($\theta = 1$, all mass at $q = p = 5\%$), the probability mass is not simply flattening out and widening a bit as it did in the Gaussian case or the Clayton copula. No: the probability mass moves directly to the extreme event, here the "no defaults at all" ($q = 0$) event. Therefore, a second peak appears at $q = 0$ with a trough of probability for the intermediate events, in fact, at $q = 0$ there is a singularity in the density of the loss distribution. This can be seen nicely at the plot for $\theta = 1.1$. As dependency is further increased, the peak

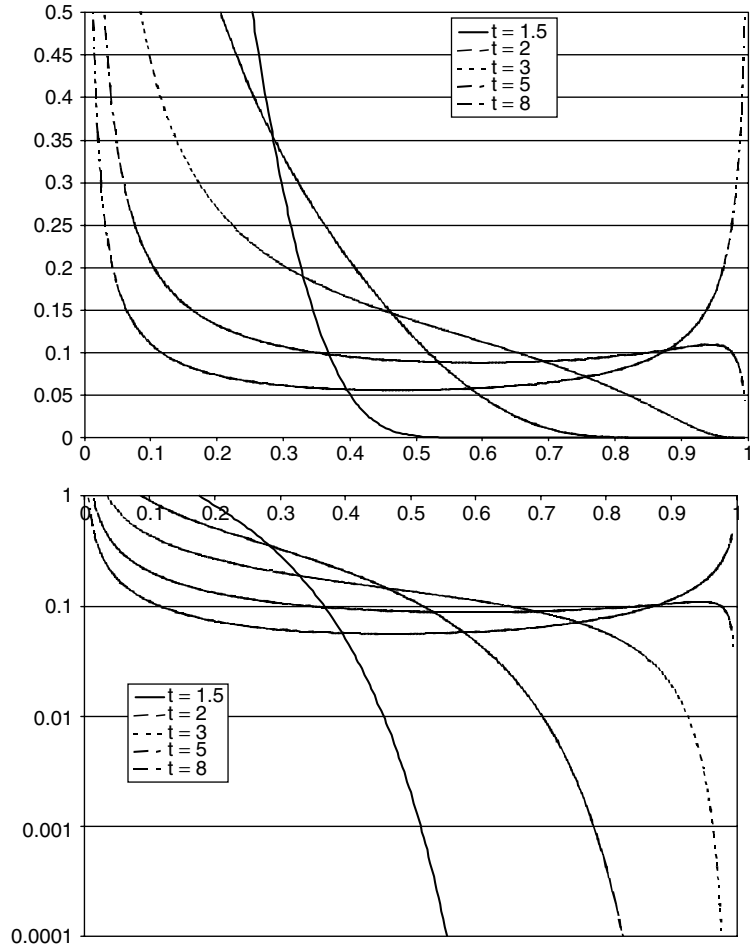


Figure 6: Loan loss distributions for the Gumbel model, $p = 5\%$, large dependency parameter ($\theta = 1.5, 2, 3, 5, 8$)

around $q = 5\%$ disappears, and in the end the distribution slowly approaches the perfect positive dependence limit with mass only at $q = 0$ and $q = 100\%$.

Comparison for constant bivariate default correlation

If the default correlation (i.e. the linear correlation coefficient) between two default events is fixed at $\rho = 10\%$, we have to choose the following model parameters: $\varrho_V = 30.55\%$ in the Vasicek model, $\theta = 18.12\%$ in the Clayton model, and $\theta = 1.39$ in the Gumbel model. This level of bivariate default correlation corresponds to a high, but not unrealistic level of default correlation in the loan portfolio.

In general, just specifying the default correlation between any *two* obligors does not completely determine the loss distribution of the *whole* portfolio. There are many more defaults involved in the events that we are considering here. Even in a portfolio of just 100 obligors,

a 10% loss rate would amount to 10 defaults, and it requires a lot of faith to assume that the joint default probabilities of any two obligors gives us much information on the probability of this event. (It does give some information, though.) For the comparison, we used the default correlation primarily to fix the last remaining parameter in the models, and we are interested how large the differences between the different models may be.

Figure 7 shows the result of the model comparison. There are two surprises: first, the Gaussian (Vasicek) model and the Clayton model imply almost identical loss distributions,

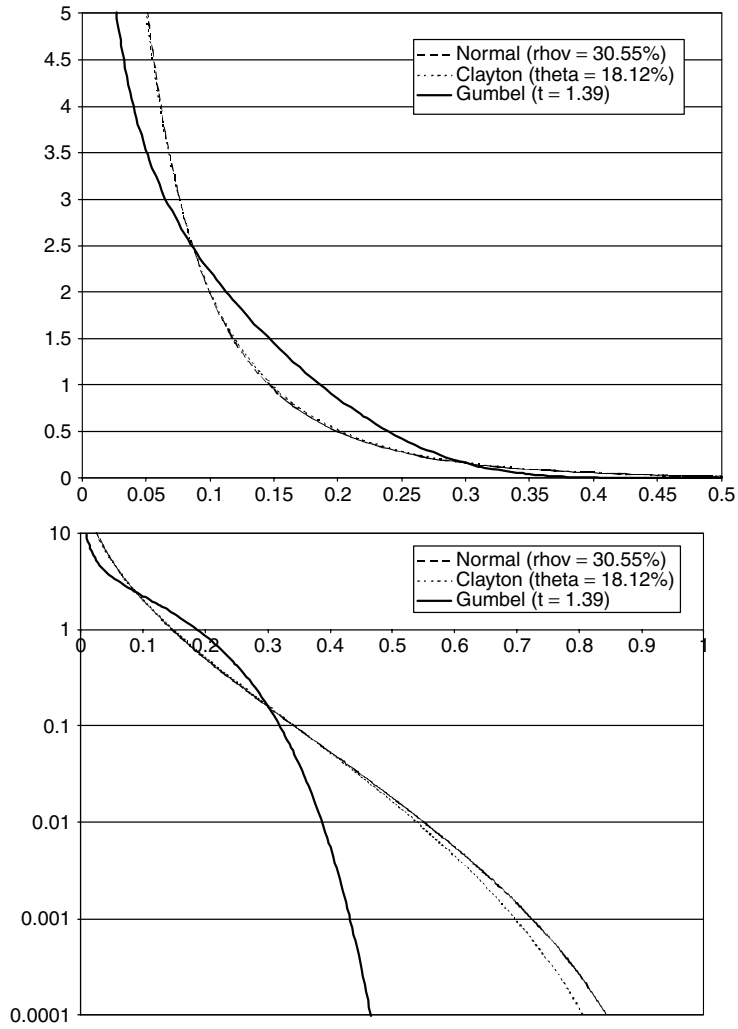


Figure 7: Loan loss distributions for the Gaussian (Vasicek), the Clayton and the Gumbel model. Default correlation between two default events is fixed at $\rho = 10\%$. Individual default probability $p = 5\%$. The parameter values are $\varrho_v = 30.55\%$ (Vasicek); $\theta = 18.12\%$ (Clayton); and $\theta = 1.39$ (Gumbel)

and it seems that for these models, the bivariate default correlations do have very similar implications for the loss distribution.

The second surprise is the large deviation of the Gumbel model from the other two. The Gumbel model has significantly more probability mass for losses between 10% and 30% of the portfolio (and for zero losses: the density is infinite there), but then again it has significantly less probability mass for higher default events (losses larger than 30%). This result cannot be driven by any correlation-type measure that only measures the dependency between two obligors' defaults, it is driven by higher order moments.

Finally, we should mention that the Frank copula would imply a loss distribution that is yet again fundamentally different. The mixing variable in the Frank copula only takes values on the positive integers. Thus, the large portfolio losses can only take a countable number of values. The loss distribution will be discrete, with discrete steps for different values of Y.

Conclusion

This paper has shown three things. First, modelling joint distributions in a different way than just using a variant of the multivariate normal distribution function is feasible. In particular, there are algorithms (such as the one by Marshall and Olkin, 1988) that allow the efficient generation of dependent random numbers in high dimensions. We gave a few examples, but the class of Laplace transforms of positive random variables (and thus of possible dependency structures) that can be generated with the Marshall and Olkin (1988) algorithm is much larger.

Second, it is worthwhile to investigate the effect of the implicit assumption of a Gaussian dependency structure on the risk measures and the returns distribution of the portfolio. As we have seen in the credit risk case, this effect can be either minor (if one only compares the Vasicek model to the Clayton-dependent model) or significant (if one thinks the Gumbel copula is a realistic alternative).

And finally, we have provided an application of this modelling strategy to the field of credit risk modelling. Credit risk is a particularly interesting application because here the consequences of extreme events are large, and much less data is available than for example for equity returns. Yet, the simple 0-1 structure of default-survival allowed us the derivation of some closed-form solutions for the loss distributions of large portfolio loan losses, and we could compare the implications of these models without having to resort to lengthy simulations.

FOOTNOTES & REFERENCES

1. This amounts to a shift in the coordinate system and a subsequent linear scaling. As the default barrier K will be chosen to fit the default probability (and thus implicitly follows the same transformation), this transformation does not change the structure of the model.
2. There is no closed-form solution for the density of a α-stable distribution, but it can be readily evaluated from its Fourier transform.

■ Devroye, L. (1986) *Non-uniform Random Variate Generation*. Springer-Verlag, Berlin.
■ Frey, R. and McNeil, A. J. (2001) Modelling dependent defaults. Working paper, Department of Mathematics, ETH Zürich.
■ Gupton, G., Finger, C. and Bhatia, M. (1997) Credit metrics – technical document, Risk Metrics Group.

■ Joe, H. (1997) *Multivariate Models and Dependence Concepts*, vol. 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.
■ Li, D. X. (2000) On default correlation: a copula function approach. Working paper 99-07, Risk Metrics Group.
■ Marshall, A. W. and Olkin, I. (1988) Families of multivariate distributions. *Journal of the American Statistical Association* 83, 834–841.
■ Merton, R. C. (1974) On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* 29, 449–470.
■ Nelsen, R. B. (1999) *An Introduction to Copulas*, vol. 139 of *Lecture Notes in Statistics*. Springer, Berlin.
■ Schönbucher, P. J. and Schubert, D. (2001) Copula-dependent default risk in intensity models. Working paper, Department of Statistics, Bonn University.
■ Vasicek, O. (1987) Probability of loss on loan portfolio. Working paper, KMV Corporation.
■ Vasicek, O. (1997) The loan loss distribution. Working paper, KMV Corporation.

12

Sovereign Debt Default Risk: Quantifying the (Un)Willingness to Pay

Ephraim Clark

Wilmott magazine, May 2003

The creditworthiness of a corporate borrower depends, for all practical purposes, on its ability to pay. Sovereign borrowers generally have the power to unilaterally abrogate contractual obligations, and, thus, besides the ability to pay, their creditworthiness depends on the government's willingness or unwillingness to pay even if it has the ability.

The literature on country risk has recognized the importance of the willingness factor. Eaton, Gersovitz and Stiglitz (1986), for example, argued that because a country's wealth is always greater than its foreign debts, the real key to default is the government's willingness to pay. Borensztein and Pennacchi (1990) suggest that besides other observable variables that are tested, the price of sovereign debt should be related to an unobservable variable that expresses the debtor country's willingness to pay. Clark (1991) suggests that the price of sovereign debt is related to a country's willingness to pay which is motivated by a desire to avoid the penalties of default.

From a practical point of view for default risk analysis, the problem with the concept of the willingness (unwillingness) to pay is that it is not readily observable. However, although it is not directly observable, it can be measured, as we shall see, with a high degree of accuracy as an American style call option where the decision to default depends on the government optimizing the trade-off between the gains to be reaped through non payment and the costs associated with not paying. Furthermore, the model generates estimates of the parameters that make it possible to make reasonable forecasts of sovereign debt defaults.

The model: unwillingness as an option

The argument for modelling the unwillingness to pay as an American style call option goes as follows. Based on the generally accepted concept of national sovereignty, a government has

Contact address: Middlesex University, The Burroughs, London NW4 4BT, UK
E-mail: e.clark@countrymetrics.com Telephone: (44) 0181 411 5130 www.countrymetrics.com

an ongoing de facto right to repudiate or default on its foreign debt, if this is deemed in the national interest. There is, however, no obligation on the part of the government to default. Thus, it is an option. It is an American style option because the government can default at any time it chooses. As we mentioned, this it will do when deemed in the national interest and the national interest is when the benefits from defaulting are large enough to offset the costs of doing so. Hence, if we measure the relative value of default with respect to the costs of defaulting, we are in effect measuring the degree of the government's unwillingness to honour its contractual debt obligations. The higher the value of the option to default, the less willing is the government to pay.

To value this option, let x represent the nominal amount of foreign debt outstanding. This is the amount at stake if the government decides to default.[1] To the extent that autonomous commercial and capital transactions are not perfectly offsetting, the nominal amount of sovereign foreign debt outstanding has a random element that can be represented by geometric Brownian motion:[2]

$$dx(t) = \alpha x(t)\,dt + \sigma x(t)\,dz(t) \tag{1}$$

where α = the growth rate of the foreign debt, which depends on the economy's requirements for external financing.
$\mathrm{d}z(t)$ = a Wiener process with zero mean and variance equal to dt.
σ^2 = the variance parameter of $dx(t)/x(t)$.

The instantaneous dividend rate or convenience yield is equal to $R - \alpha = \psi$.

When the government defaults, the net value that it receives is equal to the amount of the debt outstanding less the penalties and other costs associated with its action. In the literature, these penalties and costs take two forms. The first revolves around the costs associated with the loss of access to capital markets (Eaton and Gersovitz, 1981). The second concerns the costs due to direct sanctions such as the elimination of trade credits or the seizure of assets (Bulow and Rogoff, 1989).

Let C represent the indemnities and costs associated with default. These costs will be influenced by the economy's overall performance, which varies stochastically over time, and the extent to which foreign resources, both imported and borrowed, which also vary stochastically over time, play a role in the economy. They will also be influenced by the reactions to the default of a wide range of players including politicians, businessmen, bankers, international civil servants, and consumers. Typically these reactions vary according to circumstances and current perceptions surrounding them. Finally, perceptions themselves are likely to vary according to the evolution of a complex set of economic, political, social, environmental, etc. variables at the local, regional, and international levels. In short, the sources of variation are numerous and unpredictable enough that there should be a considerable random element in variations of C.

Since default costs cannot be negative and since the sources of variation are numerous and random, the stochastic element of C can be represented by geometric Brownian motion. If there is a long term trend, due, for example, to a relationship between C and gross domestic product (GDP) or C and outstanding debt or to ongoing initiatives of the IMF and the World Bank that make it more or less difficult to default, the following process will describe the evolution of C through time:

$$dC(t) = \pi C(t)\,dt + \omega C(t)\,dw(t) \tag{2}$$

where π is the trend parameter, ω^2 is the variance parameter of the percentage change in $C(t)$ and $dw(t)$ is a Wiener process with zero mean and variance equal to dt, with $dz(t)dw(t) = \rho\, dt$ where ρ is the instantaneous correlation coefficient between x and C.

Let Y represent the value of the government's option to default. It is a function of x and C:

$$Y = Y(x(t), C(t)) \tag{3}$$

Since this option can be exercised at any time its value depends on when the option is exercised. The government will want to exercise at the point that maximizes its value. To solve this problem, consider a new variable $g = x/C$, the value of the investment per dollar of default cost, where the time arguments have been dropped for simplicity of notation. Using (1), (2) and Ito's lemma gives:

$$dg = \mu g\, dt + \delta g\, ds \tag{4}$$

where:

$$\mu = \alpha - \pi - \sigma\omega\rho + \omega^2$$
$$\delta^2 = \sigma^2 - 2\sigma\omega\rho + \omega^2$$
$$ds = \frac{\sigma\, dz - \omega\, dw}{\delta}$$

Make the change of variables $y(g, 1) = Y(x, C)/C$. Use the capital asset pricing model to find R_g, the required rate of return on g, so that the instantaneous payout rate κ is equal to $R_g - \mu = \kappa$. Then, with the instantaneous payout or convenience yield equal to $\kappa g\, dt$, going through the well known steps of setting up a riskless hedge consisting of one unit of the option and $-y'(g)$ units of the investment and applying Ito's Lemma gives the following differential equation:

$$\tfrac{1}{2}\delta^2 g^2 y'' + (r - \kappa) g y' - r y = 0 \tag{5}$$

The general solution to (5) is:

$$y = K_1 g^{\eta_1} + K_2 g^{\eta_2} \tag{6}$$

where $\eta_1 > 1$ (because $\kappa > 0$) and $\eta_2 < 0$ are the roots to the quadratic equation in η:

$$\eta_1, \eta_2 = \frac{-(r - \kappa - \delta^2/2) \pm \sqrt{(r - \kappa - \delta^2/2)^2 + 2\delta^2 r}}{\delta^2}$$

The particular solution depends on the boundary conditions. When g goes to zero, the option has no value. Thus, the first boundary condition is:

$$y(0) = 0 \tag{7}$$

which makes $K_2 = 0$.

When the government defaults, it will receive $x - C$, that is, the amount of debt outstanding less the indemnities and other costs associated with the act of default. In terms of equation 6, this implies that default will yield $g - 1$. However, there will be a level of g, designated by g^*, where it will be optimal for the government to act. At values of g lower than g^*, the value of the right to default will be higher than the net value of defaulting and, consequently, it will be in the government's interest to put off defaulting until the net value of defaulting is at least as high as the value of the right to default that will be lost. At the boundary, then, the value of the right to default is just equal to the net value to be obtained through defaulting:

$$y(g^*) = g^* - 1 \tag{8}$$

The smooth pasting condition that makes it possible to find g^* jointly with $y(g)$ is:

$$y'(g^*) = 1 \tag{9}$$

Thus, the solution to (6) is:

$$y = K_1 g^{\eta_1} \tag{10}$$

where:

$$K_1 = \frac{1}{\eta_1 - 1} g^{*-\eta_1}$$

and:

$$g^* = \frac{\eta_1}{\eta_1 - 1}$$

and the government's option to default can be evaluated as:

$$Y = C K_1 g^{\eta_1} \tag{11}$$

Equation 11 expresses the government's unwillingness to honor its international debt obligations. The unwillingness grows as the gains from default grow with respect to the costs of defaulting. In this context, default itself is the result of a rational welfare optimizing decision based on relative costs and benefits.

Implementing the methodology

To implement the foregoing methodology, we need estimates of the parameters $\alpha, \pi, \sigma, \omega, \rho$ and κ(r can be observed). These parameters can be estimated directly from the times series of D and C. The World Bank and a number of private services provide data on D, total outstanding country debt. The time series for C must be estimated. Estimating C involves estimating the costs pointed out by Eaton and Gersovitz (1981) that are associated with the loss of access to capital markets as well as the costs pointed out by Bulow and Rogoff (1989) that are due to direct sanctions such as the elimination of trade credits or the seizure of assets. To

do this, we can apply the countrymetrics methodology (www.countrymetrics.com) developed by Clark (1991) and summarized in Clark (2002). This methodology involves two steps. First, the country's international market value is estimated year by year. Second, an algorithm is generated that estimates the percentage of this value that would be lost in the case of default.

Once the time series for D and C and their parameters have been estimated, they can be applied in equation 11 to calculate the value of the unwillingness to pay. The resulting estimate of the unwillingness to pay can then be used in default risk analysis. One way that the information can be used is to forecast sovereign debt discounts on the secondary market. Clark and Zenaidi (1999), for example, showed that the option value of the unwillingness to pay is a significant explanatory variable for the sovereign debt discounts on the 21 countries they studied from 1986–1994, when standing alone and when combined with the other major variables outlined in the literature.

The information can also be used to forecast debt defaults. From boundary conditions 8 and 9 we can calculate the value of g^*. This value represents the ratio of debt to the cost of default that triggers a default. This figure can be compared with the actual ratio to measure the "distance to default" from which a forecast can be derived. For example, in my talk at the GARP Credit and Counterparty Risk Summit, I used this methodology to forecast an Argentine default within the year. True to form, Argentina defaulted on its foreign debt and, what is even more interesting, the default is entirely due to its "willingness to pay". In fact, Argentina's foreign exchange reserves are many times higher than the payment that has been defaulted.

I will conclude this paper with another forecast. Much has been written on Brazil's election results and its willingness to continue paying on its sizeable foreign debt. I have run the figures on this and find that the distance to default is actually longer with the incoming populist government than it was with the outgoing conservative government. The bad news is that Brazil has already entered the default zone, and unless a large amount of new money is forthcoming, default will be triggered. Thus, for Brazil, I am looking either for a default, a major rescheduling or a bailout within the next 12 months.

FOOTNOTES & REFERENCES

1. As in corporate defaults, sovereign default on a debt service payment puts the total debt outstanding in default through *pari passu* and cross default clauses that are routinely written into the debt contracts. In practice, once default has occurred and the government has demonstrated its willingness to suffer the costs this entails, a bargaining process begins, usually within the Paris and London Clubs, whereby the government enters negotiations with its creditors to trade the value of the exercised default option by recommencing payments in exchange for concessions such as forgiveness, reschedulings, etc. Our analysis is limited to the initial decision to default.

2. The continuous, random element stems from the balance of payments identity and the random, continuous nature of autonomous commercial and capital transactions where sovereign capital transactions make up the difference. We consider jumps to new levels of nominal debt outstanding through forgiveness, rescheduling, Brady deals or the like as part of the negotiation process that occurs subsequent to the act of *de facto* or *de jure* default. This will be the subject of another paper.

■ Borensztein, E. and Pennacchi, G. (1990) Valuing interest payment guarantees on developing country debt, *IMF Staff Papers*.
■ Bulow, J. and Rogoff, K. (1989) Sovereign debt: is not to forgive to forget? *American Economic Review* 79(1), 43–51.
■ Clark, E. (1991) *Cross Border Investment Risk*. Euromoney Publications, London.
■ Clark, E. (2002) *International Finance*. Thomson Learning, London.
■ Clark, E. and Zenaidi, A. (1999) Sovereign debt discounts and the unwillingness to pay. *Finance*, 20(2), 185–199.
■ Eaton, J. and Gersovitz, M. (1981) Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies* 48(152), 289–309.
■ Eaton, J., Gersovitz, M. and Stiglitz, J. (1986) A pure theory of country risk. *European Economic Journal*.

13
Chord of Association

Aaron Brown[1]

Quantitative Finance Review 2003

> A writer expresses himself in words that have been used before because they give his meaning better than he can give it himself, or because they are beautiful or witty, or because he expects them to touch a *chord of association* in his reader, or because he wishes to show that he is learned and well read. Quotations due to the last motive are invariably ill-advised; the discerning reader detects it and is contemptuous; the undiscerning is perhaps impressed, but even then is at the same time repelled, pretentious quotations being the surest road to tedium.[2]

I like the opening quotation because it contains a double metaphor for association. It discusses the effect of a quotation on the reader's opinion of the author. Of course, the author's use of a quotation also changes its meaning. When John Donne wrote the phrase "for whom the bell tolls", it meant one thing. When Ernest Hemmingway chose it as the title of a novel, it changed both how readers interpreted his novel, and how everyone thought about the phrase afterwards. When Metallica later used it as a song title, the meanings shifted again. I use the word "association" to refer to interaction among factors in which all of them are affected.

"Chord of association" is an evocative phrase for this definition. If the strings of a piano are undamped (which can be done with one of the foot pedals) the vibration of any string causes the other strings to vibrate in proportion to the least common denominator of their lengths. Striking one note produces a chord, and the chords reinforce each other in infinite progression. Random vibrations will cause the piano to sing. It is not that one string causes another to vibrate, it's that all the strings together interact to make the song. This is a good metaphor for memory, and also for financial market price movements.

A more precise related concept is *correlation*. For two standardized[3] random variables, the correlation coefficient is the expectation of their product. Correlation is a useful concept, but it also can be a misleading one. I offer the following quiz to start you thinking about the issues I want to discuss. Spend a few minutes pondering these questions and settle on your best answers, then read on to see my opinion.

Contact address: Morgan Stanley, 750 7th Avenue, 11th Floor, New York, NY 10019.
E-mail: AC.Brown@MorganStanley.com

This article represents the personal opinion of the author and does not necessarily reflect the views of Morgan Stanley or any other entity.

Correlation quiz

1. You lend money to 10 different entities, each with a 10% probability of default. Defaults are uncorrelated. What is the probability that all 10 loans will default?
2. The manager of a large cap equity mutual fund wants to keep his portfolio standard deviation within 10% of the S&P 500 index standard deviation. How many stocks of average volatility and beta must he buy to accomplish this?
3. The Sharpe ratio of a portfolio is defined as:

$$\frac{\mu_p - r_0}{\sigma_p}$$

   where μ_p is the expected return on the portfolio, σ_p is the standard deviation and r_0 is the risk-free rate of interest. For a manager who holds the S&P 500 index, which will improve her Sharpe ratio more, finding a stock with:
   (a) Average volatility and correlation with the S&P 500, but 2% more expected return, or:
   (b) Average volatility and expected return, but zero correlation with the S&P 500?
4. Which financial institution will have the lower firmwide 99% value-at-risk (VaR), a firm with:
   (a) 100 different businesses, each with $1 million VaR on a stand-alone basis and average correlation of 0.1 between businesses, or:
   (b) 10 different businesses, each with $10 million VaR on a stand-alone basis and no correlation between businesses?

It takes two to tango, but it takes more to make money

One basic problem with correlation is that it is a pairwise concept. A correlation matrix can show interactions among more than two variables, but each matrix element is defined by a pairwise relation. Think of quotations or piano strings to realize how limited pairwise analysis can be.

For a more practical example, consider the correlation of short-term interest rates and long-term rates. Suppose short-term rates are cut by central bank action. This will cause borrowers to shorten their maturities in order to get cheaper funds and lenders to lengthen their maturities to get higher yields. Fewer borrowers and more lenders in the long end will cause long-term interest rates to fall, arguing for a positive correlation between short-term and long-term rates.

The lower interest rates will stimulate economic activity, lead to an increase in the money supply and weaken the currency. All three factors will fuel inflation. Inflation will push long-term interest rates up. So we can equally well argue that the correlation should be negative, a decrease in short-term rates leads to an increase in long-term rates. And I have only covered two out of hundreds of economically important interactions, and only for one type of shock.

There is no meaningful general answer to the question of the correlation between short-term and long-term interest rates. Of course I can measure a correlation coefficient using historical data, but the answer I get will depend on the time period, sampling interval and lag. It will not be useful either for understanding the relation between short-term and long-term rates, nor predicting future movements.

Even when people do not explicitly use correlation coefficient in an argument, they may have an underlying assumption that pairwise information is enough to make a judgment about the joint probability of more than two events.

Quiz answer 1

If the defaults are independent then the probability of all 10 loans defaulting is the product of their individual probabilities, or 10^{-10}, one chance in ten billion. But all we are told is that the defaults are uncorrelated, which means the chance of any pair of them defaulting is the product of their individual probabilities. So, the chance of any two loans defaulting is one in one hundred.

Suppose there are two possible cases. One percent of the time, all loans default. The other 99% of the time, the loans are placed into a hat with one blank piece of paper, and one is drawn at random to default. If the blank paper is drawn, no loan defaults.

It's easy to verify that the probability of any individual loan defaulting is 10%, composed of 1% chance of all loans defaulting and $99\%/11 = 9\%$ of being drawn from the hat in the other case. It's even easier to see that the chance of any two specific loans defaulting is 1%, because it can only happen in the first case. So the loans have the specified chance of default and zero correlation. But the chance of all 10 loans defaulting is one in one hundred, not one in ten billion. I could easily make other assumptions that make the chance of all 10 loans defaulting zero.

This is not simply a textbook illustration. Although the correlation between two unrelated borrowers defaulting is typically low, we observe occasional spikes in default rates that cannot be explained by pairwise analysis. Most investment grade defaults happen during these spikes, making them more important for estimating credit risk than the normal periods in between. The effect is less pronounced in equity markets, but it is well known that when the market goes down sharply, individual stock correlations go up. The worst days in the stock market are much worse than can be explained by average pairwise correlations between stocks and the market.

Pairwise correlations cannot tell us much about probabilities that involve more than two events. In principle, we could get around this problem by looking at third and higher cross-moments, but this quickly becomes impractical. There are too many of them to estimate reliably, and small errors in even one cross moment can lead to significant overall errors.

A non-financial example of this issue is the 1996 fire in the Channel Tunnel. An expensive professional safety design study estimated that a serious fire would occur once every 300 years and that a series of safeguards reduced the risk of major damage to negligible levels. The basic argument was that a long string of individually unlikely and uncorrelated events would have to occur before a fire could kill anyone or threaten the tunnel itself. Yet the fire occurred in only 2 years, and all the safeguards failed (although fortunately no-one died and the tunnel was repaired).

My advice in this situation is to estimate the probability of disaster as the product of the probabilities of the two most unlikely events, rather than all the events. Since correlation is pairwise, that's all we can safely assume. It's also an example of the famous applied mathematics counting rule of thumb, 0, 1, ∞. Things either don't exist, or are unique, or there are an infinite number of them. Either none of your safeguards will fail, or one will, or all will. So once two go, you might as well assume they all go. All bets are off. You're already in such unlikely territory that something you didn't factor into your calculations probably happened.

The Chunnel fire began on a train in France, before it entered the tunnel. The smoke detectors on the train failed, and five of the six detectors in the tunnel. Guards saw the fire but were unable to contact the central command station because of communication problems and absent personnel. The detector and communications failures can be reasonably considered to be uncorrelated random events, although arguably both stemmed from complacency and sloppiness. However, all the failures after that point were results of the damage caused by the fire (among other things, it destroyed the power and communications facilities and triggered a false derailment alarm) and panic.

Conditional expectation

People often use the correlation coefficient to form a conditional expectation. For example, if scores on the math and verbal SAT tests have a correlation of 0.75, and we know that a student scored two standard deviations above the mean on the math test, we predict a score of $0.75 \times 2 = 1.5$ standard deviations above the mean on the verbal test. When people make this argument, they are thinking of a graph like Figure 1. The points are randomly scattered around a line with slope of 0.75. If you observe an X value, the corresponding point on the line seems to be a pretty good guess of what Y will be.

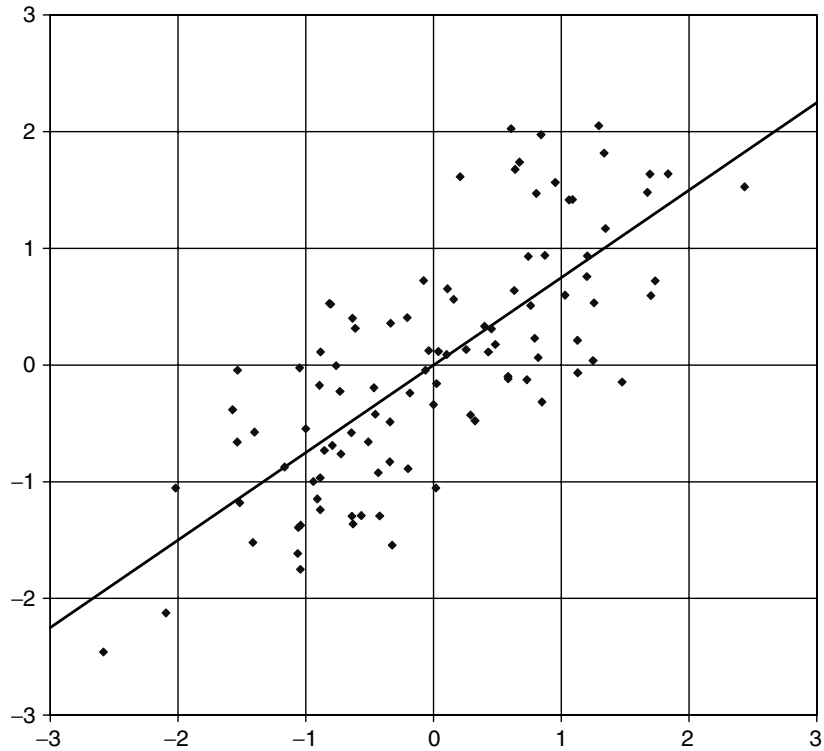


Figure 1: Two variables with correlation 0.75

But Figure 2 shows another graph of two variables with a correlation of 0.75. Now the line gives absurd predictions. If we observe an X value not near -1 or $+1$, we can't say anything about the conditional expectation of Y because we have no data. If X is near -1, the one thing we know is Y is *not* likely to be near the line. For positive X, the relation between X and Y is negative, not positive. If this were the relation between math and verbal SAT scores, and we knew someone scored two standard deviations above the mean on the math test, it would be silly to predict a verbal score 1.5 standard deviations above the mean. We don't have any data for that observation, but extrapolation suggests a verbal score a little below the mean.

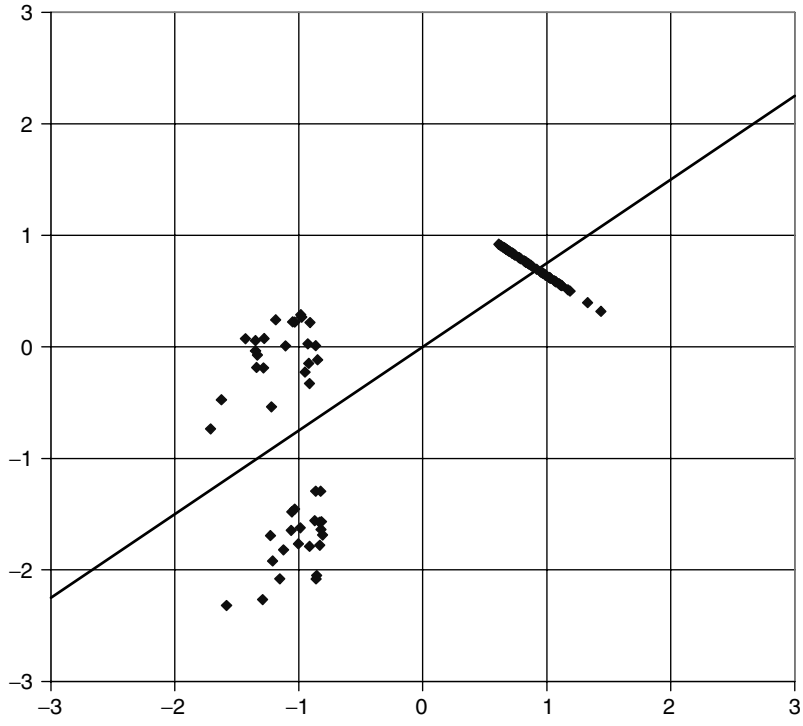


Figure 2: Another two variables with correlation 0.75

The data suggest that there are three types of students: one below average on both tests, one below average in math but about average on verbal, and one above average in math with a strict negative relation between math and verbal scores. This pattern cannot be described by any one statistic. The problem isn't that correlation coefficient is a bad measure of conditional expectation, it's that conditional expectation is too complicated to pack into any one number. In cases like this, which are very common in finance, you first have to figure out what cluster of points you're in, then you can make meaningful predictions from simple statistics.

Figure 3 shows the standardized daily returns of Johnson & Johnson stock vs. Procter & Gamble from January 2000 to January 2004. The solid line showing correlation of 0.4 clearly does not describe the data. The correlation between these two similar companies is much higher, about 0.75, except for 7 March 2000, when P&G stock fell 31 per cent on bad earnings news

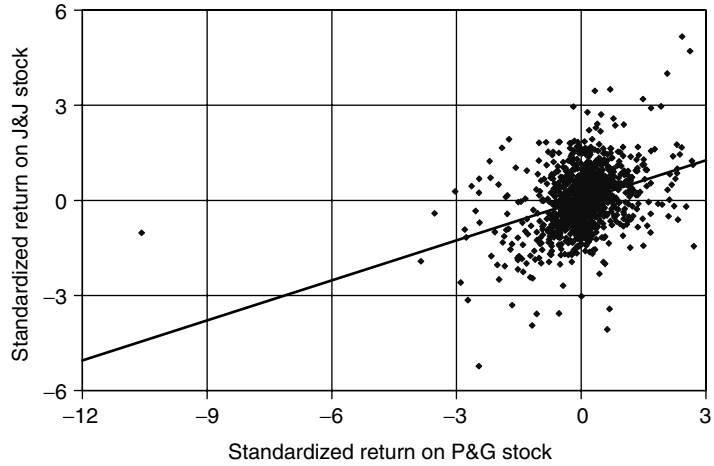


Figure 3: What is the correlation between J&J and P&G stock returns?

(in retrospect, it was an early warning of the market-wide crash to come 3 days later). Neither 0.75 nor 0.4 describes the relationship between the returns on these two stocks.

Quiz answer 2

Before tackling this question, let's review some basic correlation mathematics. The standard deviation of log return of a portfolio is equal to:

$$\sqrt{\sum_{i=1}^{n}\omega_i^2\sigma_i^2+\sum_{i=i}^{n}\sum_{j\neq i}^{n}\rho_{ij}\omega_i\omega_j\sigma_i\sigma_j}$$

where ω_i is the weight of security i in the portfolio, σ_i is the standard deviation of log return of security i and ρ_{ij} is correlation coefficient between the log returns of security i and security j.

If all the weights, standard deviations and correlations are equal, a little algebra reduces this expression to:

$$\sigma\sqrt{\rho+\frac{1-\rho}{n}}$$

This expression is very important in finance. It tells us that diversification can only help us reduce the standard deviation of a portfolio to the square root of the average correlation coefficient times the average standard deviation of individual security (we have to be loose about the definition of "average" here). A lot of people have in the backs of their minds the zero correlation case, like casino gambling or life insurance (approximately anyway), where infinite diversification can eliminate all risk.

For large capitalization, US stocks volatilities and correlations vary quite a bit over time, but long-term average values of 32% annual volatility and pairwise correlation of 0.4 are reasonable. That implies a market standard deviation of $0.32\sqrt{0.4}=0.2$, which is about right.

The quiz question asks how many stocks we need to get within a given range of the standard deviation of the market. So we have to solve for n in:

$$\frac{\sigma\sqrt{\rho}}{\sigma\sqrt{\rho + \frac{1-\rho}{n}}} = k$$

That gives us:

$$n = \frac{k^2}{1-k^2}\frac{1-\rho}{\rho}$$

The first term depends only on how close we want to get to the standard deviation of the market, for 10 per cent (i.e. $k = 0.9$) it is about 4.5. The second term depends only on the average pairwise correlation among stocks, and is 1.5 for $? = 0.4$. So seven average stocks will get us a standard deviation within about 10 per cent of what we can get by buying the entire market.

This assumes the individual stocks are average. It is not hard to find four stocks with average pairwise correlations of 0.2; they combine to give a portfolio standard deviation of:

$$\sqrt{0.2 + \frac{0.8}{4}} = \sqrt{0.4} \text{ times the average standard deviation of the individual stocks}$$

This is the same diversification benefit as you get by holding the entire market. If you can find more than four such stocks, you can get better diversification than the holding the market.

The idea that small portfolios can have more diversification benefit than the entire market is surprising to many people, because they are thinking of the uncorrelated case. With uncorrelated returns, the more securities you buy, the lower your portfolio standard deviation. This is why most mutual funds hold 100 or more stocks and diversification rules from the Securities and Exchange Commission (SEC), Internal Revenue Service and pension fund organizations encourage excessive diversification. The SEC rules for a well-diversified fund[4] are the most stringent; no stock can make up more than 5 per cent of the portfolio. While you could meet that test with 20 stocks, you would have to hold exactly equal shares in each, and rebalance constantly, selling your winners and buying more of your losers. The practical minimum is 40 stocks to meet all diversification tests without special effort.

At that level, each additional stock reduces standard deviation by less than one basis point, assuming it has average correlation and volatility. If stocks are selected and monitored carefully, holding the extra stock almost certainly adds more cost than diversification benefit, except perhaps in the largest funds. Excessive faith in diversification thus contributes to mutual funds that pay no serious attention to governance at their portfolio companies and rely on sell-side analysts for research. Another problem is the rules force a high correlation with the market, which means portfolio managers can distinguish themselves only through high turnover trading, derivative bets or other techniques far riskier than portfolio concentration.

The Capital Asset Pricing Model (CAPM) has also contributed to misunderstanding correlation. The CAPM tells us the expected excess return on any security above the risk-free rate of interest is proportional to its correlation with the market. That suggests relying on low correlation among stocks to reduce risk will give lower expected returns. But the standard deviation of the portfolio depends on the pairwise correlation of its constituent stocks with each other,

not with the market. It is easy to find efficient portfolios in the CAPM sense with a dozen or fewer stocks. The CAPM says no-one can beat the market, but it doesn't say you can't do just as well with only few stocks, thus saving on research, monitoring and trading costs. Moreover, the empirical evidence for the CAPM, while strong in general, leaves plenty of room for the possibility of small portfolios with significantly above-market returns. The recent evidence from hedge funds suggests that superior Sharpe ratios are achievable through judicious management of correlation.

Causation and the central limit theorem

The first thing students learn about correlation in statistics class is "correlation does not imply causation". The dictum is promptly forgotten. That it needs to be said at all shows how easy it is to slide from one concept to the other. If A is correlated with B, it is hard to resist the idea that A causes B or B causes A or A and B are caused by some third factor.

Without getting into metaphysics, the problem with this assumption in finance is that the correlation between A and B is generally caused by many factors. Consider two stock returns, such as Johnson & Johnson and Procter & Gamble above. Some events, like J&J winning market share from P&G, will induce a negative correlation. Others, like a general increase in consumer spending, will affect both in the same direction. When P&G announces worse than expected earnings, investors may react by switching to J&J stock, or by selling both on the assumption that J&J will have similar news. When the market in general goes down investors might rationally decide that stocks are cheaper, and buy more, or that the expected return of the market has fallen, so they should sell stocks.

With all these unmeasurable effects relating thousands of securities, it is not surprising that correlations are unstable. But the central limit theorem appears to offer a way out. In the long run, all the higher order effects will diversify away, and all I need to know is the average mean, standard deviation and correlation coefficients to specify the joint distribution of my variables.

This is an "asymptotic" argument, meaning it relies upon something going to infinity. There are two schools of thought about asymptotics. One holds that proving something asymptotically is like a newspaper reporting that something is "rumoured". When you read that something is rumoured, you know the paper could find no-one to say it was true, or might be true, or even that it was rumoured to be true. So it's probably not true. Similarly, if something can be proved only asymptotically, it's probably false.

The other side has two subschools. The low self-esteem crowd figures that if an estimator is asymptotically optimal, hell, if it's any kind of optimal, it's good enough for me. The mathematical purists delight in asymptotics that don't kick in until unimaginably high numbers of observations. This means they have found something ideally true that contradicts all possible experience.

I'm in the middle on this one, I say do the asymptotics. Sometimes they help you with your second observation, sometimes they are worthless with all the data in the universe. So how does the asymptotic argument for correlation coefficient stack-up?

Suppose I flip a fair coin 1000 times, winning \$1 for every head and paying \$1 for every tail. If I use the Normal approximation to the binomial distribution, my 99% VaR for this is \$74; 1% of the time I will lose more than \$74. The actual probability of losing more than \$74 is 0.88%. An error of this size will not faze risk managers; there are much bigger errors in any practical VaR calculation. But it does surprise a lot of people to learn that 1000 independent

Bernoulli trials, about the friendliest test imaginable for the central limit theorem, have a 12% error solely from assuming Normality.

Now suppose there is a pairwise correlation among the flips of 0.003. The Normal approximation says VaR doubles. A pairwise correlation of −0.001 cuts the Normal approximation to VaR to $2. When you learn that the standard error of measuring the correlation coefficient in this case is 0.004, and changes of that much in the correlation coefficient move VaR from $2 to $150, it's clear that correlation coefficient is not a useful statistic in this context.

It gets worse. The Normal approximation is useless for setting VaR. If all the coin flips have zero pairwise correlation with each other, the true VaR can be anywhere from zero to $316, more than four times the Normal approximation VaR. If the central limit theorem doesn't work for 1000 uncorrelated Bernoulli trials, it is not reliable for much financial data.

Quiz answer 3

The very best large cap equity mutual funds seem to outperform index funds, on a risk-adjusted basis, by about 2% per year. This is the difference between a 5-star fund that attracts lots of new capital, and one that never gets enough assets under management to break even. Michael Price managed three funds with 2% outperformance over 20 years, and sold them for $610 million in 1996.

Using 4% for the long-term expected excess return on the S&P 500 (the return on the index minus the risk-free rate of interest) and 20% for the standard deviation, an S&P 500 index fund has a Sharpe ratio of $0.04/0.2 = 0.2$. It takes $0.06/0.2 = 0.3$ to get to Michael Price territory.

A stock with an excess expected return of 6% and average correlation (0.4 with other stocks, $\sqrt{0.4} = 0.63$ with the S&P 500) and volatility (32%) can help us get there. We get the best Sharpe ratio by putting one-third of the portfolio in this stock and two-thirds in the market. That gives us an excess return of 0.0467, with a standard deviation of $\sqrt{0.467} = 0.216$. The Sharpe ratio is 0.216.

A stock with average expected return (4%, the same as the market), and average volatility (32%) but zero correlation should be weighted two-sevenths, with five-sevenths in the market. Our excess return remains at 4%, of course, but the standard deviation falls to $\sqrt{1/35} = 0.169$. That gives a Sharpe Ratio of 0.237, more than double the improvement of the high expected return stock.

Low correlation stocks have another advantage. Each one you find helps more than the last one, three zero correlation stocks are enough to get our Sharpe ratio up to the target 0.3. Each high expected return stock helps less than the previous one; we can never get up to 0.3, however many we find.

This raises the question of why there is a massive industry looking for stocks with better than average expected returns, and little public effort expended on finding stocks with low correlations. The question is even stranger when we consider that no one has ever succeeded demonstrably at the first task, while the second is easy. Hedge funds exploit it every day, but regulation discourages people from selling it to the public.

Suppose you did find a stock you thought had an excess expected return of 2%. To verify that it had positive excess expected return at the conventional two standard deviation level of statistical confidence would take 1000 years. It takes only 10 observations to get the same confidence for a zero correlation stock. For large cap stocks, you could do that in one day. That

makes the search for high expected return stocks a matter of pure faith. Precise measurement can quickly find low correlation stocks, and just as quickly alert the manager when things change.

The correlation problem

To move from the theoretical to the practical, consider an investor planning some international diversification. I took data for the 36 national equity indices with good quality daily price quotes. Then I computed the level of investment in each that led to a $1 million 99% 1-day VaR.

Figure 4 shows the average VaR that results from combining different numbers of these portfolios, with equal weights. For one portfolio, the VaR is $1 million by construction. For two portfolios the average VaR should have fallen to $830,000, based on the average correlation among indices; in fact, it fell to only $890,000. However, this is not what I call the correlation problem. There are several reasons why the actual result will not be as good as the theoretical prediction. This sort of result is about as good as you ever get with real financial data.

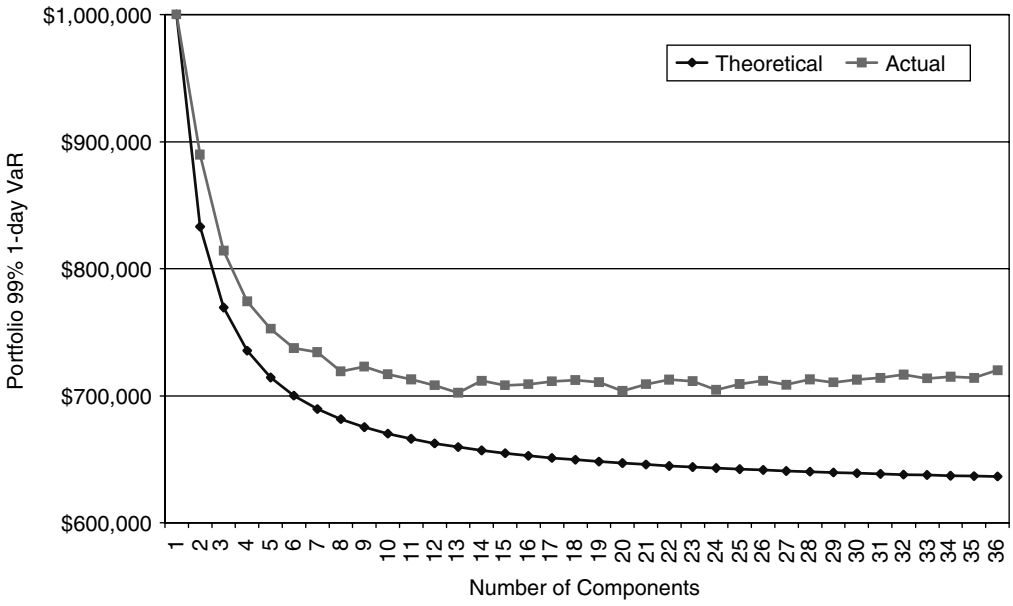


Figure 4: Diversification benefits in international investing

As the number of international indices added goes up, the actual results follow the theoretical curve, just a little less steep. But when you add the ninth index, the VaR actually goes up. After that point, further diversification does not seem to help. The theoretical VaR drops another $50,000, but the actual VaR does not change.

This is something you will see a lot in financial data. Diversification effects are similar to what is predicted by the correlation matrix until you get more than a certain number of securities, then they stop helping. If you are looking at things near the centre of the distribution, standard deviation for example, diversification will help for more securities than if you are looking at

the tails, as we do in risk management. If you go far enough out in the tails, diversification does not even help with two securities.

I consider this to be the oldest problem in finance. A lot of people would say standard deviation is a bigger problem than correlation. I think standard deviation is like energy, it's what makes finance work. If prices didn't change over time, there would be no reason to store or plan. If prices didn't change over space, there would be no incentive to trade. There would be no dynamic business models; the only job would be taking 12 eggs and selling them as a dozen. So standard deviation is an essential to economic activity, not a financial problem.

Although standard deviation is good, like energy, it can shock you, burn you, blow you up. Correlation offers to tame standard deviation. If correlation is low we can diversify, if it's high we can hedge. But as we've seen, its value is limited.

The oldest known diversification advice comes from Ecclesiastes, chapter 11 verse 2 (in the King James translation): "Give a portion to seven, and also to eight; for thou knowest not what evil shall be upon the earth". It is significant that the Bible does not say "diversify as much as possible" or "buy hundreds of individual stocks". Ecclesiastes wrote at a time of sophisticated trade throughout the Mediterranean. His readers certainly knew there were bigger numbers than eight. But they had already observed that diversification doesn't help much beyond that number. Dividing into more portions increases cost and reduces oversight without protecting against the unknown evils of the earth.

Two thousand years later, in another sophisticated Mediterranean trading environment, Shakespeare[5] has Shylock say:

> "Ho, no, no, no, no; my meaning in saying he is a good man is to have you understand me that he is sufficient ... he hath an argosy bound to Tripolis, another to the Indies; I understand, moreover, upon the Rialto, he hath a third in Mexico, a fourth for England? And other ventures he hath ... ships are but boards, sailors but men; there be land-rats and water-rats, water-thieves and land-thieves? I mean pirates; and then there is the peril of waters, winds and rocks. The man is ... sufficient. Three thousand ducats – I think I may take his bond."

Shylock is obviously putting some faith in the power of diversification based on the number of Antonio's ventures, and their geographic dispersion. The next part of the speech is often misinterpreted as a counterargument, as if Shylock is weighing the pros and cons of the loan. However, it is clearly meant as additional support. This shows a sophisticated statistical understanding, that for a given total amount of risk the more varieties there are the less the chance that any single risk will increase in probability enough to frustrate all the ventures.

I don't want to give away too much of the plot for those who haven't read it, but the story turns on the possibility that none of Antonio's ships come in (we later learn there are seven altogether). Although probability theory was undeveloped at the time, marine insurance was an important industry. Underwriters had already observed, in modern language, that failures of two ships on unrelated voyages were close to independent; but that failures of larger numbers of ships were far more common than extrapolation of individual probabilities would predict. Thus, Shakespeare's audience was familiar with the argument that, while diversification helped for a few ships, it was plausible that seven ships to seven different destinations could fail for seven different reasons.

Quiz answer 4

Let's first tackle this problem assuming the firm uses a Normal approximation to VaR. In that case, VaR is proportional to standard deviation. Total firmwide VaR will be the sum of the stand-alone VaR ($100 million in both cases) times $\sqrt{\rho + (1-\rho)/n}$. In that case $\rho = 0.1$ and $n = 100$ implies a VaR of $33.0 million, while $\rho = 0$ and $n = 10$ means that VaR is lower, only $31.6 million. The lesson is that even small correlations among businesses limits the value of diversification. There is no practical way to distinguish between businesses with correlations of 0 and 0.1. Large financial institutions cannot assume that diversification among markets, businesses and regions reduces risk capital requirements beyond a certain point, even if there are no measurable correlations among the businesses.

If VaR is not computed with a Normal approximation, we cannot compute the firmwide VaR. However, we can still analyse the question. Zero correlation means the probability of simultaneous VaR breaks in two specific businesses is 0.0001. A correlation of 0.1 raises that probability by a factor of 11 to 0.00109. The firm with 10 zero correlation businesses cannot have two simultaneous VaR breaks with probability more than 0.0045 which is less than 1%. So the firmwide VaR must be set by a scenario in which only one business loses more than $10 million.

With bad risk controls, it's possible for the probability distribution of losses to pile up just before the VaR point, and spread out afterwards. So with a VaR of $10 million, for example, you get a lot of days with losses between $9 and $10 million, and once you exceed $10 million you have expected losses of $100 million or more. That is why we cannot compute the firmwide VaR. However, with good risk controls the pathological distributions should not occur and we can set reasonable upper bounds for firmwide VaR.

One VaR break at the firm with 10 businesses should not cause the firmwide VaR to exceed $50 million. The nine businesses that do not have VaR breaks shouldn't lose more than the square root of nine times their VaRs (remember, these businesses are uncorrelated, so half of them make money on average). The one that breaks VaR should not do worse than lose twice the VaR amount.

The firm with 100 businesses could see 34 VaR breaks on one day in 100, that is consistent with the individual probability of 1 per cent and the correlation of 0.1. Using the same assumptions as above, that sets a reasonable upper limit on firmwide VaR of $76 million.

Of course there's quite a bit of guesswork in these upper bound calculations. But all risk managers I know are more comfortable contemplating 1 business in 10 breaking VaR than 34 businesses in 100.

The future

The monuments of ancient Egypt are the most spectacular buildings of the ancient world. They are covered with intricate hieroglyphic writing that seems to offer insight into the wisdom of the ancients. However for a millennium and a half, all knowledge of how to read the hieroglyphics was lost.

In 1799, a French engineer discovered the Rosetta stone, which led to decipherment of the ancient texts. Imagine the excitement of seeing the solution to such a problem, knowing that soon the long-dead voices of ancient builders would be heard again.

We now find ourselves in a similar position with respect to the much older mystery of correlation. It has bedeviled investors since before there was money, yet it remains our best friend against excessive standard deviation. It behaves in strange and unpredictable ways. Correlation assumptions are found in many pricing models, but there is not yet any consensus model for what it is, or how it evolves.

For the first time liquid markets are emerging in correlation products. This happened with volatility in the 1970s and 1980s. When Black and Scholes announced their famous model, volatility was an unobservable. It showed up in lots of financial models, but no one knew what it was. For example, there was disagreement about whether the occasional large movements in security prices were caused by fat-tailed distributions of constant standard deviation or Normal distributions with varying standard deviation.

Liquid option markets not only answered this question (standard deviation varies), but proved to everyone's surprise that there is a term and moneyness structure to volatility that transcends individual securities and markets. Volatility has gone from something entirely unobservable to something we understand and can measure better than expected return.

It took 21 years to analyse the Rosetta stone, and about the same amount of time to really figure out volatility. There's no reason to assume correlation will be faster. But I can guess about the result.

In the first place, we don't need and won't get a theory of every pairwise correlation in financial markets. The most important feature of correlation is diversification potential: the risk of the market divided by the average risk of the constituent securities. In theory, this is the square root of the average pairwise correlation coefficient (again, being sloppy about the definition of "average"). In practice it probably has nothing to do with that. For different definitions of risk and different markets we will get different answers, but I predict we will find implied correlation, defined as the square of the implied volatility of an option on the market divided by the weighted average implied volatility of its components, will turn out to have a rich tenor and moneyness structure with common features across markets. I further predict that the definition of prudent investment management will be fundamentally altered as a result.

Another important use of correlation is in hedging. Here the key is computing the lowest risk combination of securities. In theory this is found by inverting the covariance matrix, which adds a numerically unstable process to a statistically unreliable result. And that's only if the underlying theory is sound, which it isn't. I predict that we will find statistically and numerically reliable techniques for dynamic risk management that will significantly reduce the total risk in the economy and significantly improve the efficiency of capital allocation. True innovation flourishes when it gets full credit for its correlation advantages over doing things the safest way.

Finally, correlation is essential to risk management. We need to find a way to get more than two layers of protection on things, and that will require more subtle understanding of correlation. Here I predict that we will find that existing risk management is far more expensive and stifling than it need be. With a better correlations we can tolerate less diversification, less intrusive monitoring, smaller amounts of capital and faster innovation.

Put it all together and the next 20 years have the potential to give us a sound theory of association that will relegate correlation to financial museums. The oldest financial problem will be solved forever.

FOOTNOTES

1. I would like to thank Paul Wilmott for inviting me to the 2003 London Quantitative Finance Review conference at which this paper was first presented. It was improved considerably as a result of questions and comments from the audience in London and at subsequent presentations at Citigroup's Fixed Income Research Seminar and Brown Bag Lunch talk. I particularly want to credit the insightful discussions with Kent Osband (whose book *Iceberg Risk* gave me many of the ideas presented here, and explains them in much more depth), Ed Thorpe, Deborah Pastor, John Adams, Bryce Ferguson, Evan Picoult, Richard Brandt, Jorge Sobehart, Domenic Conner and Jack Fuller.

2. Henry Watson Fowler and Francis George Fowler, *A Dictionary of Modern English Usage*, 1926, "Quotation".

3. Random variables with mean zero and standard deviation one. I can standardize a random variable by subtracting the mean, then dividing by the standard deviation.

4. A fund need not be well-diversified, but the designation helps enormously in marketing. I once ran a public mutual fund that held only four stocks. I had to plan carefully to preserve the fund's tax exemption and to qualify it to be held in pension plans. I failed to persuade major fund rating services to report on it, or brokerage firms to offer it. Virtually no-one, from SEC lawyers to investment management professionals, believed you could prudently hold fewer than 20 stocks.

5. *The Merchant of Venice*, Act I, Scene 3.

14
Introducing Variety in Risk Management

Fabrizio Lillo,* Rosario N. Mantegna,* Jean-Philippe Bouchaud† and Marc Potters****

Wilmott magazine, December 2002

"Yesterday the S&P500 went up by 3%". Is this number telling all the story if half the stocks went up 5% and half went down 1%? Surely one can do a little better and give two figures, the average and the dispersion around this average, that *two* of us have recently christened the *variety* (Lillo and Mantegna, 2000).

Call $r_i(t)$ the return of asset i on day t. The variety $\mathcal{V}(t)$ is simply the root mean square of the stock returns on a given day:

$$\mathcal{V}^2(t) = \frac{1}{N}\sum_{i=1}^{N}(r_i(t) - r_m(t))^2, \tag{1}$$

where N is the number of stocks and $r_m = (1/N)\sum_i r_i$ is the market average. If the variety is, say, 0.1%, then most stocks have indeed made between 2.9% and 3.1%. But if the variety is 10%, then stocks followed rather different trends during the day and their average happened to be positive, but this is just an average information.

The variety is *not* the volatility of the index. The volatility refers to the amplitude of the fluctuations of the index from one day to the next, not the dispersion of the result *between different stocks*. Consider a day where the market has gone down 5% with a variety of 0.1% – that is, all stocks have gone down by nearly 5%. This is a very volatile day, but with a low variety. Note that low variety means that it is hard to diversify: all stocks behave the same way.

Contact addresses: *Observatory of Complex Systems, Istituto Nazionale per la Fisica della Materia, University of Palermo, Italy.
** Science and Finance, Capital Fund Management, 6–8 Bvd Haussman, 75009 Paris, France
E-mail: bouchau@drecam.saclay.cea.fr http://www.sciencefinance.fr
† Service de Physique de L'Etat Condensé, Centre d'Études de Saclay, Orune des Merisiers, 91191 Gif-Sur-Yvette cedex, France.

The intuition is however that there should be a correlation between volatility and variety, probably a positive one: when the market makes big swings, stocks are expected to be all over the place. This is actually true. Indeed the correlation coefficient between $\mathcal{V}(t)$ and $|r_m|$ is 0.68.[1] The variety is, on average, larger when the amplitude of the market return is larger (see the discussion below). Very much like the volatility, the variety is correlated in time: there are long periods where the market volatility is high and where the market variety is high (see Figure 1). Technically, the temporal correlation function of these two objects reveal a similar slow (power-law like) decay with time.

A theoretical relation between variety and market average return can be obtained within the framework of the one-factor model, that suggests that the variety increases when the market volatility increases. The one-factor model assumes that $r_i(t)$ can be written as:

$$r_i(t) = \alpha_i + \beta_i \mathcal{R}_m(t) + \varepsilon_i(t), \tag{2}$$

where α_i is the expected value of the component of security i's return that is independent of the market's performance (this parameter usually plays a minor role and we shall neglect it),

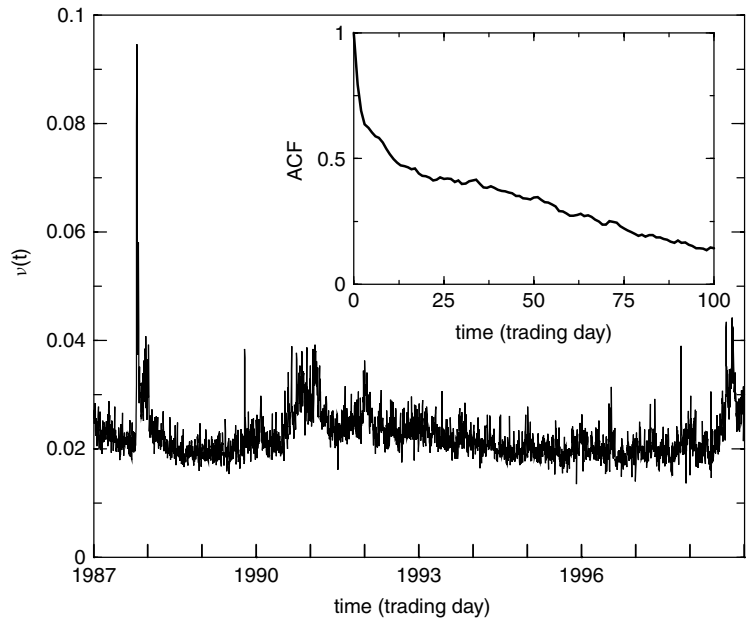


Figure 1: Time evolution of the daily variety $\mathcal{V}(t)$ of 1071 NYSE stocks continuously traded from January 1987 to December 1998. The time evolution presents slow dynamics and several bursts. The highest peak is observed at and immediately after the Black Monday. The highest value corresponds to the day after Black Monday. In the inset we show the autocorrelation function (ACF) of $\mathcal{V}(t)$. The autocorrelation has a slow decay in time and is still as high as 0.15 after 100 trading days

β_i is a coefficient usually close to unity that we will assume to be *time independent*, $\mathcal{R}_m(t)$ is the market factor and $\varepsilon_i(t)$ is called the idiosyncratic return, by construction uncorrelated both with the market and with other idiosyncratic factors. Note that in the standard one-factor model the distributions of $\mathcal{R}_m$ and ε_i are chosen to be Gaussian with constant variances; we do not make this assumption and let these distributions be completely general including possible volatility fluctuations.

In the study of the properties of the one-factor model it is useful to consider the variety $v(t)$ of idiosyncratic part, defined as

$$v^2(t) = \frac{1}{N}\sum_{i=1}^{N}[\varepsilon_i(t)]^2. \tag{3}$$

Under the above assumptions the relation between the variety and the market average return is well approximated by (see Box 1 for details):

$$\mathcal{V}^2(t) \simeq v^2(t) + \Delta\beta^2 r_m^2(t), \tag{4}$$

where $\Delta\beta^2$ is the variance of the β's divided by the square of their mean.

Therefore, even if the idiosyncratic variety v is constant, Eq. (4) predicts an increase of the volatility with r_m^2, which is a proxy of the market volatility. Because $\Delta\beta^2$ is small, however, this increase is rather small. As we shall now discuss, the effect is enhanced by the fact that v itself increases with the market volatility.

In its simplest version, the one-factor model assumes that the idiosyncratic part ε_i is *independent* of the market return. In this case, the variety of idiosyncratic terms $v(t)$ is constant in time and independent from r_m. In Figure 2 we show the variety of idiosyncratic terms as a function of the market return. In contrast with these predictions, the empirical results show that a significant correlation between $v(t)$ and $r_m(t)$ indeed exists. The degree of correlation is different for positive and negative values of the market average. In fact, the best linear least-squares fit between $v(t)$ and $r_m(t)$ provides different slopes when the fit is performed for positive (slope $+0.55$) or negative (slope -0.30) value of the market average. We have again checked that these slopes are not governed by outliers by repeating the fitting procedure in a robust way. The best fits obtained with this procedure are shown in Figure 2 as dashed lines. The slopes of the two lines are -0.25 and 0.51 for negative and positive value of the market average, respectively. Therefore, from Eq. (4) we find that the increase of variety in highly volatile periods is stronger than what is expected from the simplest one-factor model, although not as strong for negative (crashes) as it is for positive (rally) days. By analysing the three largest crashes occurred at the NYSE in the period from January 1987 to December 1998, we observe two characteristics of the variety which are recurrent during the investigated crashes: (i) the variety increases substantially starting from the crash day and remains at a level higher than typical for a period of time of the order of sixty trading days; (ii) the highest value of the variety is observed the trading day immediately after the crash.

An important quantity for risk management purposes is the degree of correlation between stocks. If this correlation is too high, diversification of risk becomes very difficult to achieve. A natural way (Cizeau *et al*, 2001) to characterize the average correlation between all stocks i,

Box 1: Proof of Eqs. (4) and (6)

Here we show that if the number of stocks N is large, then up to terms of order $1/\sqrt{N}$, Eqs. (4) and (6) indeed hold. We start from Eq. (2) with $\alpha_i \equiv 0$. Summing over $i = 1, \ldots, N$ this equation, we find:

$$r_m(t) = \mathcal{R}_m(t)\frac{1}{N}\sum_{i=1}^{N}\beta_i + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(t) \tag{7}$$

Since for a given t the idiosyncratic factors are uncorrelated from stock to stock, the second term on the right hand side is of order $1/\sqrt{N}$, and can thus be neglected in a first approximation, giving:

$$r_m(t) \simeq \overline{\beta}\mathcal{R}_m(t), \tag{8}$$

where $\overline{\beta} \equiv \sum_{i=1}^{N}\beta_i/N$. In order to obtain Eq. (4), we square Eq. (2) and summing over $i = 1, \ldots, N$, we find:

$$\frac{1}{N}\sum_{i=1}^{N} r_i^2(t) = \mathcal{R}_m^2(t)\frac{1}{N}\sum_{i=1}^{N}\beta_i^2 + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2(t) + 2\mathcal{R}_m(t)\frac{1}{N}\sum_{i=1}^{N}\beta_i\varepsilon_i(t) \tag{9}$$

Under the assumption that $\varepsilon_i(t)$ and β_i are uncorrelated the last term can be neglected and the variety $\mathcal{V}(t)$ defined in Eq. (1) is given by:

$$\mathcal{V}^2(t) \simeq v^2(t) + (\overline{\beta^2} - \overline{\beta}^2)\mathcal{R}_m^2(t) \tag{10}$$

This is the relation between $\mathcal{V}(t)$ and the market factor $\mathcal{R}_m(t)$. By inserting Eq. (8) in the previous equation one obtains Eq. (4).

Now consider Eq. (5). Using the fact that $\sum_{i=1}^{N} r_i/N = r_m$, we find that the numerator is equal to r_m^2 up to terms of order $1/N$. Inserting Eq. (2) in the denominator and again neglecting the cross-product terms that are of order $1/\sqrt{N}$, we find:

$$\mathcal{C}(t) \simeq \frac{r_m^2(t)}{\overline{\beta^2}\mathcal{R}_m^2(t) + v^2(t)} \simeq \frac{1}{\chi + F(t)}, \quad F(t) \equiv \frac{v^2(t)}{r_m^2(t)}, \tag{11}$$

where $\chi \equiv \overline{\beta^2}/\overline{\beta}^2$. This quantity is empirically found to be $\simeq$1.05 for the S&P 500, and we have therefore replaced it by 1 in Eq. (6).

j on a given day $\mathcal{C}(t)$ is to define the following quantity:

$$\mathcal{C}(t) = \frac{\dfrac{1}{N(N-1)}\sum_{i\neq j} r_i(t)r_j(t)}{\dfrac{1}{N}\sum_i r_i^2(t)}. \tag{5}$$

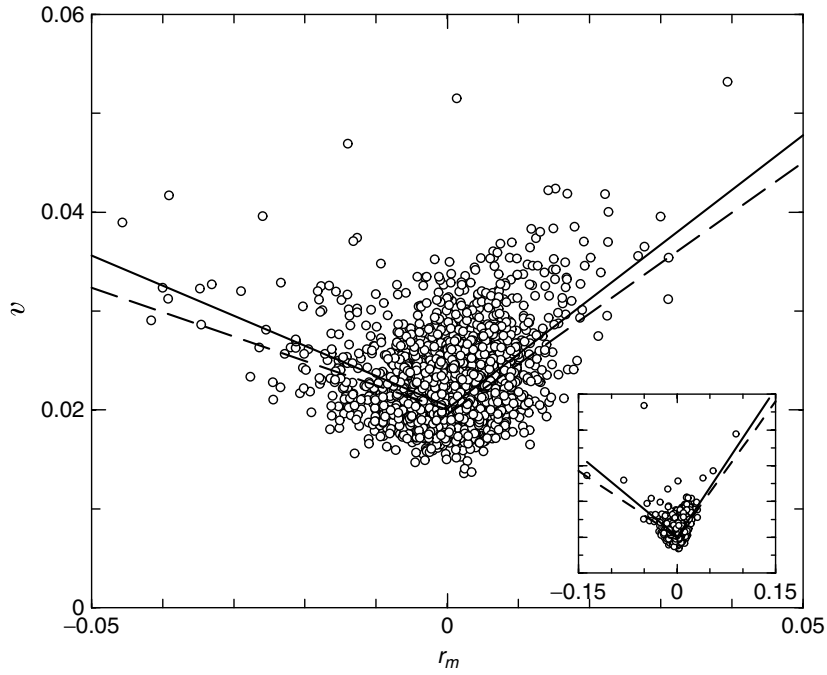


Figure 2: Daily variety v of idiosyncratic terms of the one-factor model (Eq. (3)) as a function of the market average r_m of the 1071 NYSE stocks continuously traded from January 1987 to December 1998. Each circle refers to one trading day of the investigated period. In the main panel we show the trading days with r_m belonging to the interval from -0.05 to 0.05, whereas in the inset we show the whole data set including five most extreme days. The two solid lines are linear fits over all days of positive (right line) and negative (left line) market average. The slope of the two lines are $+0.55 \pm 0.02$ (right) and -0.30 ± 0.02 (left). The tick distance in the ordinate of the inset is equal to the one of the main panel. The two dashed lines are linear fits obtained with a robust local M-estimates minimizing the absolute deviation. The slope of the two lines are $+0.51$ (right) and -0.25 (left). These values are quite close to the previously obtained ones, showing that the role of outliers is minor

As shown in Box 1, to a good approximation one finds:

$$\mathcal{C}(t) \simeq \frac{1}{1 + F(t)}, \qquad F(t) \equiv \frac{v^2(t)}{r_m^2(t)}, \tag{6}$$

As mentioned above, the variety of the idiosyncratic terms is *constant* in time in the simplest one-factor model. The correlation structure in this version of the one-factor model is very simple and time independent. Still, the quantity $\mathcal{C}$, taken for a proxy of the correlations on a given day, increases with the 'volatility' r_m^2, simply because F decreases. As shown in Figure 3, the simplest one factor model in fact overestimates this increase (Cizeau *et al*, 2001). Because the

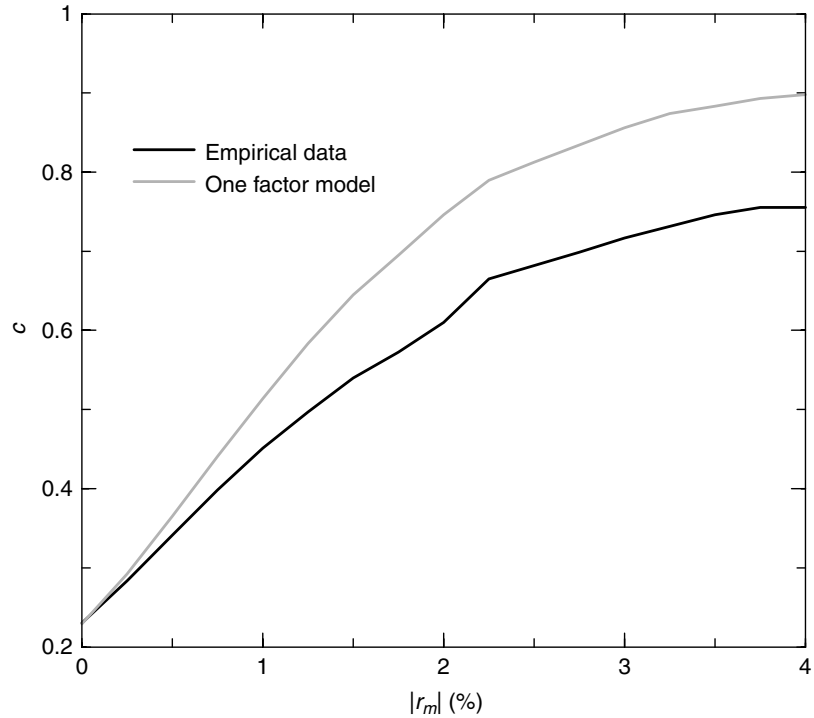


Figure 3: Correlation measure $\mathcal{C}$ conditional to the absolute market return to be larger than $|r_m|$, both for the empirical data and for surrogate data using a (non Gaussian) one-factor model. Note that both show a similar apparent increase of correlations with $|r_m|$. This effect is actually overestimated by the one-factor model with fixed residual volatilities. $|r_m|$ is in percents

idiosyncratic variety $v^2(t)$ tends to increase when $|r_m|$ increases (see Figure 2), the quantity $F(t)$ is in fact larger and $\mathcal{C}$ is smaller. This may suggest that, at odds with the common lore, correlations actually are less effective than expected using a one-factor model in high volatility periods: the unexpected increase of variety gives an additional opportunity for diversification. Other, more subtle indicators of correlations, like the exceedance correlation function defined in Box 2 and shown in Figure 4 (see Longin and Solnik, 2001), actually confirm that the commonly reported increase of correlations during highly volatile bear periods might only reflect the inadequacy of the indicators that are used to measure them.

Therefore, the idiosyncrasies are by construction uncorrelated, but not independent of the market. This shows up in the variety; does it also appear in different quantities? We have proposed above to add to the market return the variety as a second indicator. One can probably handle a third one, which gives a refined information of what happened in the market on a particular day. The natural question is indeed: what fraction f of stocks did actually better than the market? A balanced market would have $f = 50\%$. If f is larger than 50%, then the majority of the stocks beat the market, but a few ones lagging behind rather badly, and vice versa. A closely related measure is the *asymmetry* $\mathcal{A}$, defined as $\mathcal{A}(t) = r_m(t) - r^*(t)$, where

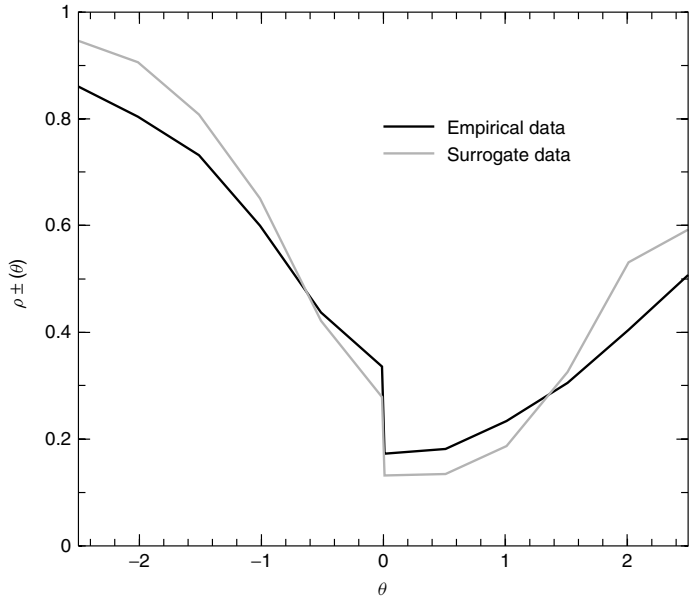


Figure 4: Average exceedance correlation functions between all pair of stocks as a function of the level parameter θ, both for real data and the surrogate non Gaussian one-factor model

Box 2: Exceedance Correlations

In order to test the structure of the cross-correlations during highly volatile periods, Longin and Solnik (2001) have proposed to study the 'exceedance correlation', defined for a given pair *ij* of stocks as follows:

$$\rho_{ij}^{+}(\theta) = \frac{\langle \tilde{r}_i \tilde{r}_j \rangle_{>\theta} - \langle \tilde{r}_i \rangle_{>\theta} \langle \tilde{r}_j \rangle_{>\theta}}{\sqrt{(\langle \tilde{r}_i^2 \rangle_{>\theta} - \langle \tilde{r}_i \rangle_{>\theta}^2)(\langle \tilde{r}_j^2 \rangle_{>\theta} - \langle \tilde{r}_j \rangle_{>\theta}^2)}}, \tag{12}$$

where the subscript $> \theta$ means that both normalized returns are larger than θ, and $\tilde{r}_i$ are normalized centred returns. The negative exceedance correlation $\rho_{ij}^{-}(\theta)$ is defined similarly, the conditioning being now on returns smaller than θ. We have plotted the average over all pairs of stocks $\rho^{+}(\theta)$ for positive θ and $\rho_{-}(\theta)$ for negative θ both for empirical data and for surrogate data generated according to a non-Gaussian one factor model Eq. (2), where both the market factor and the idiosyncratic factors have fat tails compatible with empirical data (Cizeau *et al*, 2001). Note that empirical exceedance correlations grow with $|\theta|$ and are strongly asymmetric. For a Gaussian model, $\rho^{\pm}(\theta)$ would have a symmetric tent shape, i.e. it would decrease with $|\theta|$!

In conclusion, most of the downside exceedance correlations seen in Figure 4 can be explained if one factors in properly the fat tails of the unconditional distributions of stock returns and the skewness of the index (Cizeau *et al*, 2001), and does not require a specific correlation increase mechanism.

the median r^* is, by definition, the return such that 50% of the stocks are above, 50% below. If f is larger than 50%, then the median is larger than the average, and vice versa. Is the asymmetry $\mathcal{A}$ also correlated with the market factor? Figure 5 shows that it is indeed the case: large positive days show a positive skewness in the distribution of returns – that is, a few stocks do exceptionally well – whereas large negative days show the opposite behaviour. In the figure each day is represented by a circle and all the circles cluster in a pattern which has a sigmoidal shape. The asymmetrical behaviour observed during two extreme market events is shown in the insets of Figure 5 where we present the probability density function of returns observed in the most extreme trading days of the period investigated by Lillo and Mantegna (2000). This empirical observation cannot be explained by a one-factor model. This has been shown by two different approaches: (i) by comparing empirical results with surrogate data generated by a one-factor model by Lillo and Mantegna (2000) and (ii) by considering directly the asymmetry of daily idiosyncrasies (Cizeau *et al*, 2001). Intuitively, one possible explanation of this anomalous skewness (and a corresponding increase of variety) might be related to the existence of sectors which strongly separate from each other during volatile days.

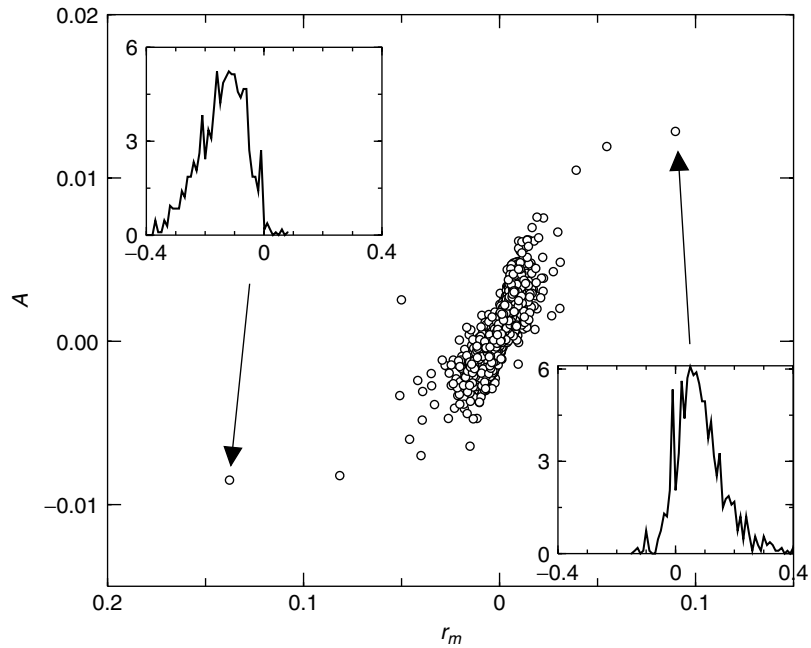


Figure 5: Daily asymmetry $\mathcal{A}$ of the probability density function of daily returns of a set of 1071 NYSE stocks continuously traded from January 1987 to December 1998 as a function of the market average r_m. Each circle refers to one trading day of the investigated period. In the insets we show the probability density function of daily returns observed for the two most extreme market days of the period investigated. Specifically, the left inset refers to Black Monday (19 October 1987) and the right inset refers to 21 October 1987. The negative (left inset) and positive (right inset) skewness of the distribution is clearly seen in both cases

The above remarks on the dynamics of stocks seen as a population are important for risk control, in particular for option books, and for long-short equity trading programs. The variety is in these cases almost as important to monitor as the volatility. Since this quantity has a very intuitive interpretation and an unambiguous definition [given by Eq. (1)], this could become a liquid financial instrument which may be used to hedge market neutral positions. Indeed, market neutrality is usually insured for 'typical' days, but is destroyed in high variety days. Buying the variety would in this case reduce the risk of these approximate market neutral portfolios.

FOOTNOTES & REFERENCES

1. This value is not an artifact due to outliers. In fact an estimation of the Spearman rank-order correlation coefficient gives the value of 0.37 with a significance level of 10^{-35}.

■ Cizeau, P., Potters, M. and Bouchaud, J.-P. (2001) Correlation structure of extreme stock returns. *Quantitative Finance* 1, 217–222.
■ Lillo, F. and Mantegna, R. N. (2000) Symmetry alteration of ensemble return distribution in crash and rally days of financial market. *European Physical Journal B* 15, 603–606.
■ Lillo, F. and Mantegna, R. N. (2000) Variety and volatility in financial markets. *European Physical Review* 62, 6126–6134.
■ Longin, F. and Solnik, B. (2001) Extreme correlation of international equity markets. *Journal of Finance* 56, 649–676.

15
Alternative Large Risks Hedging Strategies for Options

F. Selmi* and Jean-Philippe Bouchaud**

Wilmott magazine, March 2003

It is well known that the perfect Black–Scholes hedge only works in the ideal case of a continuous time, log-Brownian evolution of the price of the underlying. Unfortunately, this model is rather remote from reality: the distribution of price changes has "fat tails", which persist even for rather long time lags (see, e.g. Granger and Ding, 1997; Guillaume *et al*, 1997; Gopikrishnan *et al*, 1998; Plerou *et al*, 1999; Bouchaud and Potters, 2000). This makes the whole idea of zero-risk strategies and perfect replication shady. An alternative view was proposed in Bouchaud and Potters (2000), Schweizer (1995), where one accepts from the start that the risk associated with option trading is in general non zero, but can be *minimized* by adopting an appropriate hedging strategy (this is also discussed in Wilmott, 1998). If the risk is defined as the variance of the global wealth balance, as was proposed in Schweizer (1995) and Bouchaud and Potters (2000), one can obtain a expression for the optimal hedge that is valid under rather mild assumptions on the dynamics of price changes. This optimal strategy allows one to compute the "residual" risk, which is in general non-zero, and actually rather large in practice. For typical one month at the money options, the minimal standard deviation of the wealth balance is of the order of a third of the option price itself! This more general theory allows one to recover all the Black–Scholes results in the special case of Gaussian returns in the continuous time limit, in particular the well-known "Δ-hedge", which states that the optimal strategy is the derivative of the option price with respect to the underlying.

However, as soon as the risk is non zero, the various possible definitions of "risk" become inequivalent. One can for example define the risk through a higher moment of the wealth balance distribution – for example the fourth moment (whereas the variance is the second moment). This

Contact addresses: *CEMEFG, Université Paris II, 92–96 rue d'Assas, 75 270 Paris cedex 06, France.
**Science & Finance Capital Fund Management, 6–8 Boulevard Haussmann, 75009 Paris, France.
E-mail: jean-philippe-bouchaud@science-finance.fr http://www.science-finance.fr; and Service de Physique de l'État Condensé, Centre d'Études de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette cedex, France

is interesting since higher moments are more sensitive to extreme events. The minimization of the fourth moment of the distribution therefore allows one to reduce the probability of large losses, which is indeed often a concern to risk managers. One could also consider the 'value-at-risk' (defined as the loss level with a certain small probability) as the relevant measure of large risks (Wilmott (1998)), and aim at minimizing that quantity: this is a generalization of the present work which is still in progress (Pochart and Bouchaud, in preparation; Bouchaud and Potters, 2000). However, our main conclusions remain valid for this case as well.

Our results can be summarized as follows: the optimal strategy obtained using the fourth moment as a measure of risk varies much less with the moneyness of the underlying than both the Black–Scholes Δ-hedge and the optimal variance hedge. This is very interesting because it means that when the price of the underlying changes, the corresponding change in the hedge position is reduced. Therefore, the transaction costs associated to option hedging decrease as one attempts to hedge away large risks. Our numerical estimates show that this reduction is substantial. This result is also important for the global stability of markets: it is well known that the hedging strategies can feedback on the dynamics of the markets, as happened during the crash of October 1987, where the massive use of the Black–Scholes hedge (through 'Insurance Portfolio' strategies) amplified the drop of the market. Therefore, part of the 'fat-tails' observed in the dynamics of price changes can be attributed to this non-linear feedback effect. By reducing the sensitivity of the hedge on the price of the underlying, one can also hope to reduce this destabilizing feedback.

Let us present our mathematical and numerical results in a rather cursory way (a more detailed version will be published separately (Selmi, in preparation; Selmi and Bouchaud, in preparation). In order to keep the discussion simple, we will assume that the interest rate is zero. In this case, the global wealth balance ΔW associated to the writing of a plain vanilla European call option can be written as:

$$\Delta W = \mathcal{C} - \max(x_N - x_s, 0) + \sum_{i=1}^{N} \phi_i(x_i)[x_{i+1} - x_i], \qquad N = \frac{T}{\tau} \tag{1}$$

where $\mathcal{C}$ is the option premium, x_i the price of the underlying at time $t = i\tau$, $\phi(x)$ the hedging strategy, T the maturity of the option, x_s the strike and τ the time interval between re-hedging. Previous studies focused on the risk defined as $\mathcal{R}_2 = \langle \Delta W^2 \rangle$, while the fair game option premium is such that $\langle \Delta W \rangle = 0$ (here, $\langle \ldots \rangle$ means an average over the *historical* distribution). As stated above, this allows one to recover precisely the standard Black–Scholes results if the statistics of price returns is Gaussian and one lets τ tend to 0 (continuous time limit). This is shown in full detail in Bouchaud and Potters (2000).

Here, we consider as an alternative measure of the risk the quantity $\mathcal{R}_4 = \langle \Delta W^4 \rangle$. The corresponding optimal hedge is such that the functional derivative of $\mathcal{R}_4$ with respect to $\phi_i(x)$ is zero. This leads to a rather involved cubic equation on $\phi_i(x)$ (whereas the minimization of $\mathcal{R}_2$ leads to a simple linear equation on $\phi_i(x)$). Further insight can be gained by first assuming a static (time independent) strategy, i.e. $\phi_i(x) \equiv \phi_0(x)$. The corresponding cubic equation only depends on the terminal price distribution and can then be solved explicitly, leading to a unique real solution ϕ_4^* between 0 and 1. We show in Figure 1 the evolution of the optimal strategy ϕ_4^* as a function of the moneyness, in the case where the terminal distribution is a symmetric exponential (which is often a good description of financial data), for $T = 1$ month and a daily volatility of 1%. The corresponding price of the at-the-money call is 2.73. On the same figure,

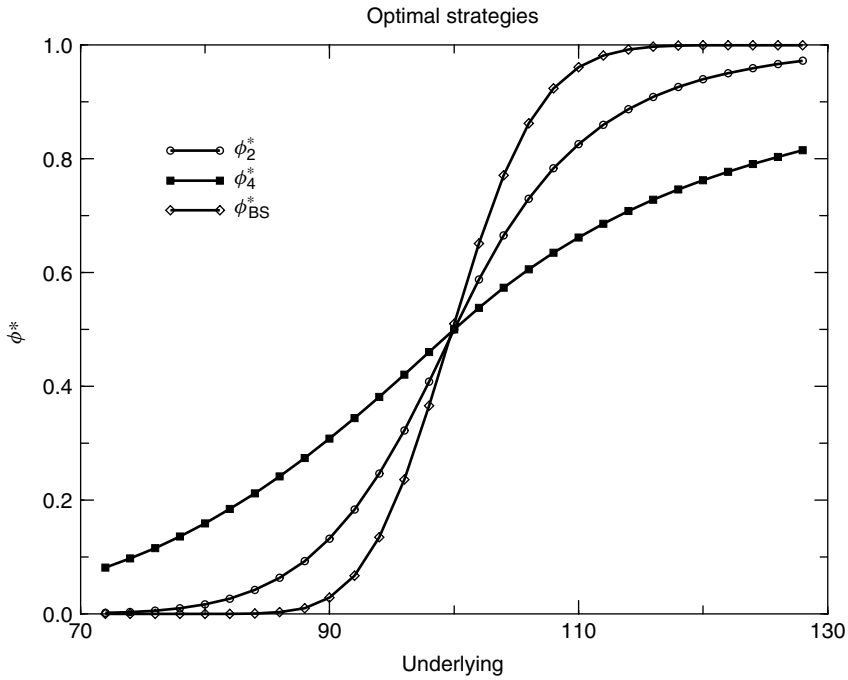


Figure 1: Three different strategies: ϕ_2^* minimizes the variance of the wealth balance, ϕ_4^* minimizes its fourth moment, whereas ϕ_{BS}^* is the Black–Scholes Δ-hedge. The strike price is 100, the maturity equal to one month, the daily volatility is 1% and the terminal price distribution is assumed to be a symmetric exponential, with an excess kurtosis of 3. The three strategies are equal to 1/2 at the money. Note that ϕ_4^* varies much less than the other two with moneyness

we have also plotted the Black–Scholes Δ-hedge, and the hedge ϕ_2^* corresponding to the minimization of $\mathcal{R}_2$. All these strategies vary between zero for deeply out of the money options to one for deeply in the money options, which is expected. However, as mentioned above, the variation of ϕ^* with moneyness is much weaker when $\mathcal{R}_4$ is chosen as the measure of risk. For example, for in the money options (resp. out of the money), ϕ_4^* is smaller (resp. greater) than the Black–Scholes Δ or than ϕ_2^*. This is because a possible large drop of the stock, which would suddenly drive the option out of the money and therefore lead to large losses due to the long position on stocks, is better taken into account by considering $\mathcal{R}_4$. One can illustrate this result differently by plotting the derivative of ϕ^* with respect to the price of the stock, which is the 'Gamma' of the option – see Figure 2. One sees that in our example the at-the-money Gamma is decreased by a factor 3.5 compared to the Black–Scholes Gamma. The corresponding average transaction costs for re-hedging are therefore also expected to decrease by the same amount. This is confirmed by numerical tests on market data (Selmi, in preparation; Selmi and Bouchaud, in preparation).

It is interesting to study the full probability distribution function of ΔW for the following three cases: un-hedged, hedged *à la* Black–Scholes or hedged following ϕ_4^*. Of particular

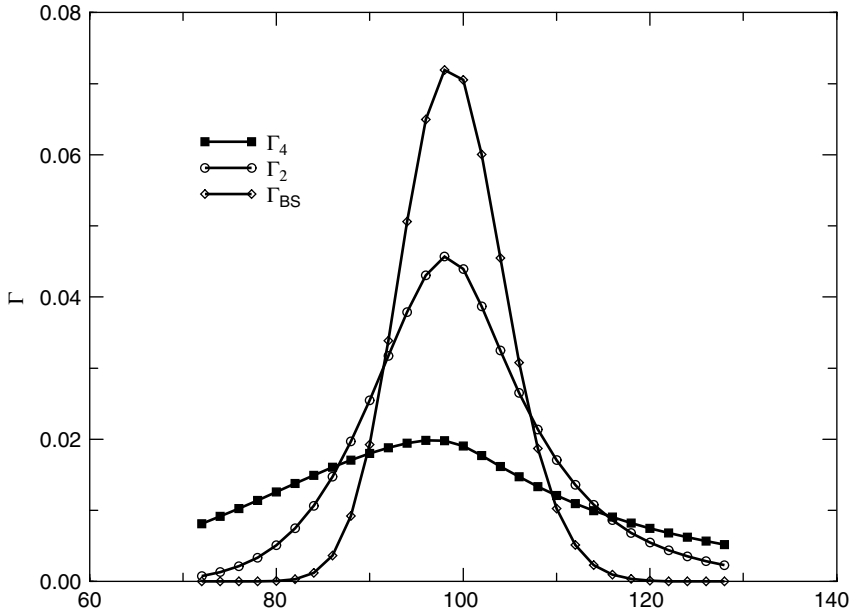


Figure 2: The three corresponding 'Gamma's', defined as the derivative of the strategy ϕ^* with respect to the price of the underlying. This quantity is important since the transaction costs for at the money options are proportional to Γ (100)

interest is the probability p of large losses – for example, the probability of losing a certain amount $|\Delta W^*|$, defined as:

$$p = \int_{-\infty}^{-|\Delta W^*|} d\Delta W \; P(\Delta W) \tag{2}$$

The results for $|\Delta W^*| = 10$ (which is four times the option premium) are shown in Figure 3 for different values of the strike price. One sees that even in the restrictive framework of a purely static hedge, ϕ_4^* allows one to decrease substantially the probability of large losses. For $x_s = 110$, this probability is decreased by a factor 3 to 4 as compared to the Black–Scholes hedge! For at-the-money options, since the static strategies are identical ($\phi_{BS}^* = \phi_4^* = 1/2$), one finds no difference in p.

We have up to now considered the simple case of a purely static strategy. In the case of the minimization of $\mathcal{R}_2$, one can show that the fully dynamical hedge can be obtained by a simple time translation of the static one, that is, one can compute ϕ_{2i}^* by again assuming a static hedge, but with an initial time translated from 0 to $t = i\tau$. This can be traced back to the fact that if the price increments are uncorrelated (but not necessarily independent), the variance of the total wealth balance is the sum of the variances of the 'instantaneous' wealth balances $\Delta W_i = W_{i+1} - W_i$. This is no longer true if one wants to minimize $\mathcal{R}_4$. However, we have shown for $N = 2$ that the simple 'translated' strategy ϕ_4^* is numerically very close to (but different from) the true optimum. Since we are in the neighbourhood of a quadratic minimum,

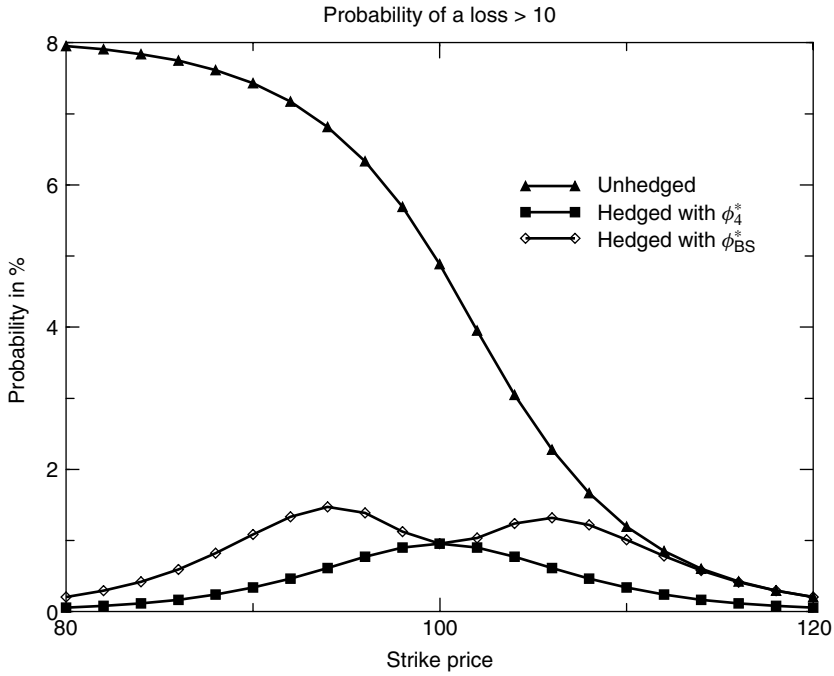


Figure 3: The probability p of losing four times the premium, as a function of the strike price, for the three different strategies: un-hedged, Black–Scholes, and ϕ_4^*. Note the substantial decrease of p for out and in the money options, even in this restrictive case where the strategy is purely static

an error of order ε on the strategy will only increase the risk to order ε^2 and is therefore often completely negligible (note that a similar argument also holds in the case of ϕ_2^*: even if the Black–Scholes Δ is in general different from ϕ_2^*, the difference is often small and leads to a very small increase of the risk – see the discussion in Bouchaud and Potters, 2000).

Finally, it is important to note that the optimal ϕ_4^* hedge for a book of options on the same underlying is not the simple linear superposition of the optimal hedge for the individual options in the book, whereas this is indeed the correct result for variance hedging. However, we have found in the case of a book containing two options with different strikes but the same maturity, that the difference between the optimal hedge and the simple linear prescription is again numerically very small.

As a conclusion, we hope to have convinced the reader that as soon as one accepts to abandon the zero-risk paradigm of Black–Scholes, very interesting issues concerning risk control arise because different definitions of the risk become inequivalent (in the Black–Scholes world, the risk is zero, whatever the definition of risk!). Therefore, optimal hedges depend on the quantity one wishes to minimize. We have shown here that a definition of the risk more sensitive to the extreme events generically leads to a decrease of the sensitivity of the hedge on the price of the underlying (the 'gamma'). Therefore, both the transaction costs and the impact of hedging on the price dynamics of the underlying are reduced.

Appendix: technical details

We give in this section the explicit form of the equations obeyed by the optimal strategies ϕ_2^* and ϕ_4^*. We introduce the following notation: $dP_N(x')$ is the probability to find the final price at x' within dx', conditioned to the present value x of the underlying asset. We introduce the following notations:

$$\mu_n = \int dP_N(x')(x'-x)^n;$$

$$\nu_{n,m} = \int_{x'>x_s} dP_N(x')(x'-x)^n(x'-x_s)^m, \tag{3}$$

and consider here the simple case where $\mu_1 = 0$. In this case, the option price is given by $\nu_{0,1}$. The minimization equation giving ϕ_2^* then reads (Bouchaud and Potters, 2000):

$$\mu_2\phi_2^* - \nu_{1,1} = 0. \tag{4}$$

Similarly, the equation for ϕ_4^* reads (Selmi, in preparation; Selmi and Bouchaud, in preparation):

$$a_4\phi_4^{*3} + 3a_3\phi_4^{*2} + a_2\phi_4^* + a_1 = 0, \tag{5}$$

with:

$$a_4 = \mu_4; \quad a_3 = \nu_{0,1}\mu_3 - \nu_{3,1}; \quad a_2 = 3\nu_{0,1}^2\mu_2 - 6\nu_{0,1}\nu_{2,1} + \nu_{2,2}; \tag{6}$$

and:

$$a_1 = -3\nu_{0,1}^2\nu_{1,1} + 3\nu_{0,1}\nu_{1,2} - 4\nu_{1,3}. \tag{7}$$

Acknowledgment

We wish to thank M. Potters, A. Matacz, A. Scannavino and the students of 'Option Math. Appli.' of Ecole Centrale Paris for useful discussions.

REFERENCES

■ Bouchaud, J.-P. and Potters, M. (2004) *Theory of Financial Risks and Derivative Pricing*. Cambridge University Press.
■ Gopikrishnan, P. Meyer, M. Amaral, L. A. and Stanley, H. E. (1998) *E.P.J.* B 3, 139.
■ Guillaume D. M., *et al*, From the bird's eye to the microscope, *Finance and Stochastics* 1, 2.
■ Granger, C. W. J. and Ding, Z. X. (1997) Stylized facts on the temporal distributional properties of daily data from speculative markets.
■ Plerou, V. Gopikrishnan, P. Amaral, L. A. Meyer, M. and Stanley, H. E. (1999) *Phys. Rev.* E 60, 6519.

■ Pochart, B. and Bouchaud, J.-P. (2004) Option pricing and hedging with minimal expected shortfall, to appear in *Qualitative Finance*: see also Föllmer, H. and Leukert, G. (2000) *Finance and Stochastics*; and Bouchaud and Potters (2004).
■ Schweizer, M. (1995) *The Mathematics of Operations Research*, 20, 1–32.
■ Selmi, F. Dissertation, PhD (in preparation).
■ Selmi, F. and Bouchaud, J.-P. (in preparation).
■ Wilmott, P. (1998) *Derivatives: The Theory and Practice of Financial Engineering* Wiley, Chichester.

16
On Exercising American Options: The Risk of Making More Money than You Expected

Hyungsok Ahn* and Paul Wilmott**

Wilmott magazine, March 2003

The price of an American option is dictated by the concept of optimal, exercise. But optimal is defined from the perspective of the option writer, who is assumed to be able to delta hedge. This theory for when to exercise the option is well known. However, buyers of American options may, and do, exercise early or late for a variety of reasons.

Suppose you are long an in-the-money American call, and you are concerned that the market may collapse. What can you do? You may close the position by selling the option, you may start delta hedging, or you may exercise the option. When the contract is OTC and there are significant costs in trading the underlying then two of these possibilities disappear. Here we present some ideas on trading options containing embedded decisions, such as those found in American options. We consider reasons why the option holder may make "suboptimal" decisions and also see what effect these decisions have on the profit made by the option writer.

1 Introduction

The American option is a contract that allows the holder to exercise before the contract's maturity, should she[1] so desire. Because of the flexibility of choosing the exercise time, the fair value of the contract is calculated as the value of the option *in the worst case for the*

Contact address: *E-mail: Hyunksok.Ahn@CommerzbankIB.com
**E-mail: paul@wilmott.com

issuer among all feasible exercise strategies that the option holder may choose. This is the price-maximization strategy.

The fundamental concept of the absence of arbitrage is still an integral part of determining the price. The issuer can construct a hedged portfolio involving the trading of the underlying asset in such a way that the value of the replicating portfolio (i.e., the up-front premium for the option plus the result of trading) is not less than his liability even when his counterparty exercises the option at the least favorable time.

Typically the price-maximizing exercise time, and hence the least favorable exercise time for the issuer, is described as an optimal stopping time, and the resulting pricing equation becomes a partial differential equation involving a free boundary.

This valuation is undoubtedly fair. If the value were less than this 'optimal' then the holder of the option could, if she so desired and were able, delta hedge her long position and then exercise her option at the price-maximizing exercise time. In this case, the balances of both the issuer and the holder would be zero at exercise. However, the option holder would have locked in a profit equal to the difference between the price at which the option was sold and its price-maximized fair value.

How to price American options

Part of the problem in valuing options with early exercise is to decide when the option will be exercised. Is there, in some sense, a best or optimal time for exercise? To correctly price American options we must place ourselves in the shoes of the option writer. We must be clear about the principles behind his strategy: the writer is hedging his option position by trading in the underlying asset. The hedging strategy is dynamic, referred to as "delta hedging". The position in the underlying asset in maintained delta neutral so as to be insensitive to movement of the asset.

By maintaining such a hedge, the writer does not care about the direction in which the underlying moves, he eliminates all asset price risk. However, he does remain exposed to the exercise strategy of the option holder.

Since the writer cannot possibly know what the holder's strategy will be, how can the writer reduce his exposure to this strategy? The answer is simple. *The writer assumes that the holder exercises at the worst possible time for the writer*. He assumes that the option is exercised at the moment that gives the writer the least profit. This is referred to as the optimal stopping time although as far as the writer is concerned it is the last thing he wants to happen.

So, out of all the possible exercise strategies we must find the one that gives the option the least value to the writer or equivalently the highest value to the holder.

This sounds very complicated but anyone who has implemented the binomial or finite-difference method knows that it is just a matter of adding one line of code to the program. That line of code simply tests at each node in the tree whether the theoretical option value is greater than the payoff, if it is not then the payoff is used instead, and this corresponds to a time at which the option should be exercised.

We will review the mathematics of this problem in Section 5.

When should the holder exercise?

The holder of the option is probably not delta hedging. It is unlikely that she is insensitive to the direction of the underlying asset. The initial assumption concerning the writer of the

option does not apply to the holder. Should the holder therefore act in the 'optimal' way as previously described?

Consider the very simplest possible scenario. You buy a call option because you believe that the underlying asset is going to rise significantly. If you are correct you will make a substantial return. If there are no dividends on the underlying then it is 'theoretically' never optimal to exercise before expiry.

Now suppose that the stock does indeed rise, but the economic situation makes you believe that a sudden fall is imminent. What should you do? The obvious solution is to sell the option and lock in your profit.

Selling the option may be quite straightforward if it is a liquid exchange-traded contract. But selling the option may not be possible, for example if the option is OTC. Many OTC contracts have features similar to early exercise. Indeed, any time that a contract allows the holder some freedom to make decisions about the future value or behavior of the contract we have exactly the same scenario concerning price-maximization. We shall return to this later, and consider some examples.

If you can't sell the option then you could start delta hedging and continue this until expiry. This might be relatively simple if the option is deep in the money. But if it is not, and if there are significant costs associated with buying and selling the underlying this may not be as easy as it would appear. Moreover, delta hedging requires access to a sophisticated dynamic model. This may not be available to the option holder especially if the contract is a complex OTC one.

In many situations, the only way of locking in the profit may be to exercise the option early. The "theory" says don't exercise, but if the stock does fall then you lose the profit. At this stage it is important to remember that the theory is not relevant to you.

Clearly, the writer and the holder of the option have different priorities, what is optimal to one is not necessarily optimal to the other. The holder of the option may simply have a gut feeling about the stock and decide to exercise. That is perfectly valid. Or she may have a stop-loss strategy in place.

She may even use a complex utility-maximization strategy.

It is highly unlikely that her exercise time will correspond to that calculated by the writer of the option.

We will examine optimal exercise from the holder's perspective in Sections 6 and 7.

How does the writer feel about exercise?

The writer has received a sum of money in exchange for the option. That sum of money was calculated assuming that the option holder exercises at a certain optimal time. This optimal exercise strategy gives the option its highest theoretical value. The writer receives this maximum amount even though the holder may exercise at any time.

Provided that his model is correct and that continuous delta hedging is possible, it is obvious that the writer can never lose. The worst that can happen to him is that the option is exercised at this theoretical optimal time. But this has already been priced into the premium he received.

On the other hand, if the holder exercises at some other time he can only benefit.

In Section 8 we examine the windfall profit made by the option writer, due to a utility-maximizing holder strategy rather than a price-maximizing one.

We will now consider some more scenarios in which the option holder may exercise non optimally, first in the world of exchange-traded options, and then in the OTC world.

2 Exchange-traded products

Various entities participate in the exchange-traded option markets. There are regulated trading shops, who are forced to maintain risk-free portfolios by regulators, using the market for the purpose of providing retail brokerage service, hedging their portfolio (including their long-dated structured portfolios), and a limited amount of speculative trading. There are unregulated entities, such as hedge funds and asset management companies, engaged in model trading based upon their proprietary speculation engines. There are also many personal investors enjoying punting with money from their own pocket.

Among all these market participants, only the regulated entities have a clear motivation for delta hedging. Others seek the best out of their invested capital while taking the risk of potential loss. Option markets are very appealing to those risk takers, because of the limited downside of option contracts and the leverage effect. In fact, a significant number of market participants is speculating on the market direction and volatility. To those risk takers, fair value of an option is merely the cover they pay when they enter the market expecting a significant upside. It is certainly not clear that the option holder is better off exercising her American option at the price-maximizing exercise time.

The price of an option depends upon the risk-neutral measure, which ignores the physical drift of the underlying asset price. The reason is that the existence of an option immediately allows one to construct a locally risk-free portfolio, and hence the risk-free rate is the only one that governs its price. As a result, the price-maximizing exercise time is also independent of the physical drift.

Can we assert that the optimal exercise time for the option holder is not affected by the market direction?

Consider the situation in which the American option you purchased when it was out-of-the-money becomes in-the-money, but now you feel that the market will move away from you. What would you do? One answer would be to close your position by selling the option. But who will purchase your option? There may be some who have the opposite opinion about the market direction, but they may prefer out-of-the-money or at-the-money options, simply because they will pay less for them. Hedgers will try to use the futures or forward market instead, for the same reason. Thus, selling your option may not be a feasible option, unless you can offer a substantial discount. (In-the-money options are not liquid. This does not invalidate the method of calculating the fair value, because risk-neutral valuation requires only the liquidity of the underlying asset not that of the options.) What about locking in your value by making your position delta neutral? Not a bad idea.

As every practitioner knows, delta hedging is far from being the simple risk-elimination strategy that one reads about in text books. In practice, the inability to hedge continuously and the presence of transaction costs on the underlying make delta hedging a less-than-perfect strategy.

But even if delta hedging could be done perfectly. There is still a simple situation in which writer and holder may disagree on the optimal exercise time. A different view on the volatility of the underlying asset is one obvious reason why investors may exercise their options at different times. And the same is true when investors are exposed to different interest rates.

In summary, it is nonsense to argue that a single exercise strategy is optimal for every investor. As an aside, there is empirical evidence (Overdahl and Martin, 1994, for example) that a substantial portion of all exercised American call options are exercised before their theoretical price-maximizing exercise boundary.

3 OTC products

Although a transaction of an American option in its pure form is rare in the OTC market, there are many structures that involve interactive decision making by each party during the lifetime of the contract. The results of embedded decisions are unknown *a priori*, and therefore, the fair valuation assumes that the outcome of decisions is such that the contract has the highest possible value. The reason for this is that the risk-neutral value of a contingent claim must not be smaller than the cost of maintaining the hedged portfolio, regardless of the result of the embedded decisions. If a party maintains a risk-free portfolio, he will be able to extract the fair value.

There follow several examples of contracts containing embedded decisions for which the fair valuation assumes that the contract writer is delta hedging and eliminates exposure to the contract holder's decisions by maximizing the price.

Equity-linked note

Consider the situation that an investment bank provides capital to a client and receives fixed interest on the notional amount and a series of call options on its client's equity as an upside participation. In addition, suppose that the note is cancelable: the client can pay back the notional and terminate the deal at any of the contract-specified dates. The benefit to the client in this deal is that she can lower the funding cost by sharing her returns when her business is successful. The bank will value the product under the assumption that the client's decision making is "economical", which is the worst case for the bank, and try to extract every penny of the option by maintaining a risk-free portfolio. On the other hand, hedging the option is neither in the client's best interest nor a feasible strategy because trading her own equity may be illegal. The client will make decisions based upon her best interest at any time. There is no reason why the client will choose early termination of the note at the price-maximizing boundary.

Knock-out instalment option

The knock-out instalment option in the term sheet shown in Figure 1 has a put payoff but expires worthless if the barrier is ever exceeded during the contract life. It differs from a vanilla up-and-out put in that it is not paid for with an up-front premium but in monthly instalments. The holder must decide each month whether it is worth paying the next instalment to keep the contract alive. This product is well suited as a credit hedge. The put payoff protects against bankruptcy as the share price would drop to zero in that event. The knock-out feature and the instalment plan makes the protection more affordable. If the holder's goal is to reduce her credit exposure, as is customary, then she would definitely not delta hedge. She will decide on whether to pay the next instalment based upon whether she is comfortable with her exposure or not.

In pricing this contract we put ourselves in the shoes of the delta-hedging writer. We then consider all possible combinations of payment of instalments and choose that which gives the option its highest value. Thus the pricing principle is exactly the same as for contracts with early exercise, the main difference being that there are now more decisions to be made; there is one decision to be made per month, instead of the single whether-to-exercise-or-not decision in an American option.

Preliminary and Indicative
For Discussion Purposes Only

USD/JPY KO Instalment-Premium Option

Notional Amount	USD 50,000,000
Option Type	133.25 (ATMS) USD Put/JPY Call with KO and Instalment Premium
Maturity	6 months from Trade Date
Knockout Mechanism	If, at any time from Trade Date to Maturity, the USD/JPY spot rate trades in the interbank market at or above JPY 140.00 per USD, the option will automatically be cancelled, with no further rights or obligations arising for the parties thereto.
Upfront Premium	JPY 1.50 per USD
Instalments	JPY 1.50 per USD, payable monthly from Trade Date (5 instalments)
Instalment Mechanism	As long as the instalments continue to be paid, the option will be kept alive, but the Counterparty has the right to cease paying the instalments and to thereby let the option be cancelled at any time.
Spot Reference	JPY 133.25 per USD

This indicative termsheet is neither an offer to buy or sell securities or an OTC derivative product which includes options, swaps, forwards and structured notes having similar features to OTC derivative transactions, nor a solicitation to buy or sell securities or an OTC derivative product. The proposal contained in the foregoing is not a complete description of the terms of a particular transaction and is subject to change without limitation.

Figure 1: Term sheet for a knock-out instalment option

Equity unwinder

An equity unwinder is a typical structure for a company which owns a substantial amount of shares of another company as a result of M&A, for example, but wants to dispose of them for cash. To avoid possible devaluation of the shares, the shareholder keeps the option of determining an unwinding schedule. In return, she shares her potential over-achievement with her cash provider. In a sophisticated variant, the cash provider shorts a put for credit protection, and decides the number of shares to be unwound within the contract-specified limits. The cash provider will value the product under the assumption that his client will unwind the shares in such a way that his hedge cost is maximal. On the other hand, the shareholder has no reason for hedging, which would only mean that she has to buy back a certain amount of shares that were supposed to be disposed. In fact, some contracts even specify that buy-back is a trigger event for insolvency, because it can be viewed as market manipulation. The shareholder's decision will heavily depend upon the market direction.

Convertible bond

A convertible bond (CB) is a coupon-paying bond which the holder can exchange for equity in a contract-specified ratio at contract-specified times before maturity. Because of this additional feature, a CB is not only a cheap funding vehicle, but also a strategic tool for M&A. When will the CB be converted? As in the previous case, the bond issuer may not be eligible to trade his own equity. In fact, he has no other choice but to enjoy the cheap funding for a while. The fair valuation assumes that the holder maintains a risk-free portfolio and converts the bond at the price-maximizing exercise boundary in order to extract the maximal risk-free compensation for the cheap coupons. However, a significant number of the convertible bond holders are speculators; speculating on direction, volatility or credit-worthiness.

Chooser range note

The vanilla range note has cashflows linked to the number of days that the reference rate (typically a LIBOR rate) lies within a specified band. In the Chooser Range Note (CRN), the band is not pre-specified in the contract but is chosen by the contract holder at the start of each period. In the example in the term sheet shown in Figure 2 there are four decisions to be made, one at the start of each period. And that decision is not of the simple binary type ("Do I exercise or not", "Do I pay the instalment or not") but is far more complex. At the start of each period the holder must choose a range, represented by, say, its mid-point. Thus there is a continuous and infinite amount of possibilities.

Preliminary and Indicative
For Discussion Purposes Only

GBP 2YR Chooser Range Accrual Note Linked to 6 month GBP LIBOR

The note pays a coupon based on the number of days that 6-month LIBOR sets within an 80bps range. The range is chosen by the buyer at the beginning of each coupon period.

Issue Date	24th March 2000
Maturity Date	24th March 2002
Issue Price	100%
Coupon	[6 month LIBOR + 1.00%] x N/D
N	Number of business days that 6 month LIBOR is within the RANGE
D	Number of business days in the OBSERVATION PERIOD
RANGE	Determined by the buyer two days prior to the beginning of each OBSERVATION PERIOD
OBSERVATION PERIOD	Period 1: 24th March 2000–24th September 2000 Period 2: 24th September 2000–24th March 2001 Period 3: 24th March 2001–24th September 2001 Period 4: 24th September 2001–24th March 2002

This indicative termsheet is neither an offer to buy or sell securities or an OTC derivative product which includes options, swaps, forwards and structured notes having similar features to OTC derivative transactions, nor a solicitation to buy or sell securities or an OTC derivative product. The proposal contained in the foregoing is not a complete description of the terms of a particular transaction and is subject to change without limitation.

Figure 2: Term sheet for a chooser range note

But again, this is not as complicated as it seems. The contract is priced from the hedger's perspective and the ranges are chosen so as to give the contract the highest possible value. We will not go into the details, just note that the problem must be solved in three dimensions, one for time, one for the interest rate model and one for the mid point of the range.

The hedging writer of the contract is exposed to risk-neutral interest rates, and the forward curve, the contract holder will choose ranges depending on her view on the direction of real rates. Since forward rates contain a component of "market price of risk" and since actual rates rarely show the same dramatic slope in rates and curvature as seen in the forward curve, then it is unlikely that the holder of the contract will choose the range that coincides with that giving the contract its highest value.

We will return to this example later.

Passport option

The final example is the passport option (also known as the perfect trader option, and other less polite expressions), perhaps the ultimate in contracts with embedded decisions. For here there

is a continuous spectrum of decisions to be made at each moment in time. The passport option is like insurance for your trading account. Invest in a stock, buying and selling according to your views (but only up to the limit specified in the term sheet; see Figure 3), and keep track of the amount of money that you make or lose. The amount that you accumulate is called the trading account. It will be positive if you have done well and negative if you have done badly. The passport option is a call option with zero strike on the money in this account. In other words, you keep the final profit but if it is negative then it is written off.

Preliminary and Indicative
For Discussion Purposes Only

USD/DEM 'Perfect Trader' Option

Notional Amount	USD 25,000,000+
Option Maturity	Three months from Trade Date
Allowed Position	Long or short up to Notional Amount
Transaction Frequency	Up to two times daily
Settlement Amount	Max(0,sum total in DEM of the gains + losses on each of the trades)
Upfront Premium	3.35% of Notional Amount

This indicative termsheet is neither an offer to buy or sell securities or an OTC derivative product which includes options, swaps, forwards and structured notes having similar features to OTC derivative transactions, nor a solicitation to buy or sell securities or an OTC derivative product. The proposal contained in the foregoing is not a complete description of the terms of a particular transaction and is subject to change without limitation.

Figure 3: Term sheet for a passport option

Clearly there are many decisions to be made. At each instant you must decide whether to buy or sell the underlying. To value this contract requires some knowledge of stochastic control. We put ourselves in the shoes of the delta hedger and assume that the holder makes decisions so as to give the contract its highest value. It is extremely unlikely that the holder will choose exactly the same strategy as deemed 'optimal' by the writer. See Ahn, Penaud, and Wilmott (1999) for details.

It is not too difficult to understand how this product came about. The product is popular among asset managers who engage in speculative trading.

4 Who wins and who loses?

Trading an option is not a two-person zero-sum game, because both the issuer and the holder can trade the underlying asset with other investors. If the issuer maintains a risk-free portfolio by trading the underlying asset while the holder leaves her position naked, the issuer's balance will be non-negative in the end while the holder's depends upon whether or not she guessed market direction correctly. If the holder exercises her option at her own optimal time, there is also a possibility that both issuer and holder can make a profit at the same time. Does this violate the conservation law? Not really, even if we consider an extreme case that everyone in the option market makes a profit. There would then have been an influx of capital from the spot market due to delta hedging.

Let us consider the potential cashflows in some detail. In Figure 4 we see the profit made by shareholders in a stock that has risen in value over some period, say a year. At the moment there are no options on this stock.

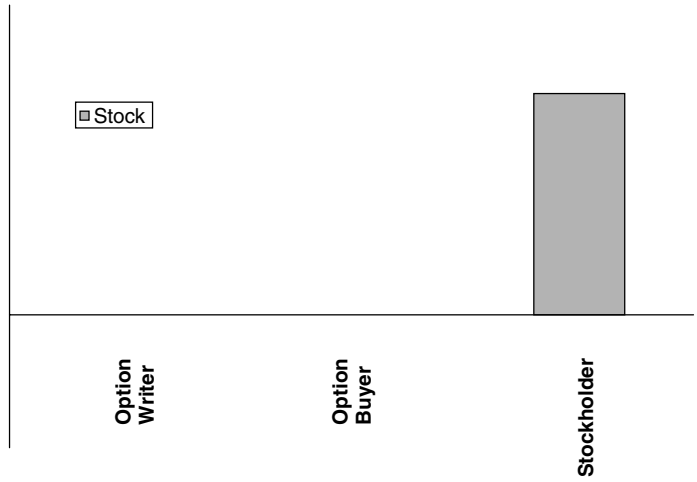


Figure 4: Profit made by shareholders in a rising market: No options traded

Now let us introduce an American option (or any contract that incorporates some choice/decisions for the holder) into this scenario, the expiry of the option coinciding with the horizon in the previous case, that is, one year. The stock will evolve dynamically exactly as in the non-option case.

What are the cashflows now?

1. If the writer sells the option for "fair value", according to the delta-hedging-and-price-maximizing method, *and* the holder exercises at the price-maximizing boundary then the writer makes no profit. The premium paid for the option is paid by the holder and is added to the profit made by the shareholders. The mechanism for the transfer of premium from holder to shareholders is simply the process of delta hedging.
2. On exercise or expiry the holder of the option may get a payoff. If the contract is a simple call and ends up in the money then the holder gets the difference between the share price and the strike price. This again does not have any impact on the writer of the option, but via delta hedging, it is taken from the profit of the shareholders.
3. Of course, the writer of the option is in practice going to add a mark-up, his profit margin. This is paid for by the option holder.
4. The final cashflow in this picture is concerned entirely with the non optimal, in the price-maximizing sense, exercise of the option. There is a windfall profit made by the option writer whenever the holder exercises 'non optimally.' (This profit could be thought of as part of the payoff, but in the diagram we have subtracted it from the shareholders.)

Figure 5 shows all of these cashflows. In the perfect Black–Scholes world the option writer profits only from mark-up and from the windfall. The option holder will gain or lose according

to the balance between the premium they pay and the payoff they receive. The market value of all the positions at expiry/exercise adds up to the value of the shareholders' profit in a world without options.

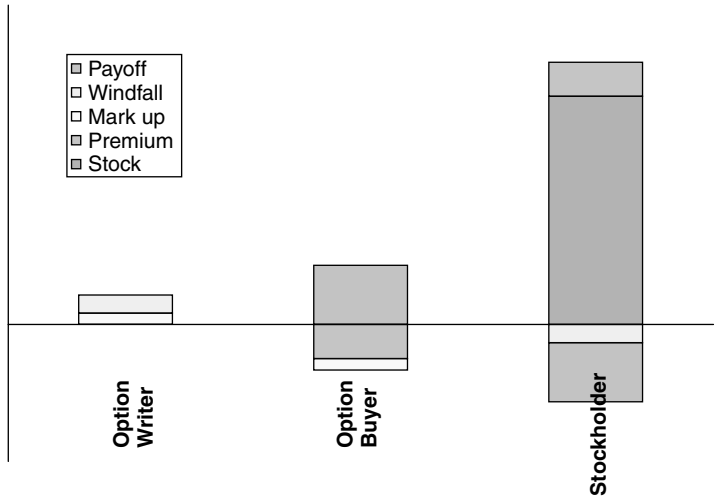


Figure 5: Profit made by shareholders in a rising market: American options traded

We are now going to embark on the mathematical journey. During this journey we ask that the reader hold the following thoughts:

- The writer is in the Black–Scholes world.
- The option holder can't or won't delta hedge.
- The option cannot be sold, think of it as OTC.

This means that the holder cannot or will not exit her position other than by exercising.

5 Optimal trading strategy: The classical formulation

First, we explain the classical valuation of American options. For a rigorous derivation see Myneni's article (1992), which contains a survey of the literature on the subject. In what follows, we assume that the underlying is tradable and its price satisfies the following stochastic differential equation (SDE):

$$dS(t) = M(t)S(t)\,dt + \sigma S(t)\,dW(t) \tag{1}$$

where W is a standard Wiener process, σ is volatility, and M is an adapted process. The only motivation for assuming a constant volatility is brevity. In fact, the following results can be

easily modified to accommodate other volatility structures. All we need is a suitable condition that (1) has a sensible solution.

As shown in Harrison and Pliska (1981), the complete-market assumption allows a trader to replicate the payoff of an arbitrary contingent claim by trading the underlying. Thus, we consider that an issuer of an option carries Δ units of the underlying. The value of his portfolio at any moment is

$$\Delta S(t) - V(t, S(t))$$

where V is the fair market price of the option he sold. The issuer wants to make sure that his portfolio earns at least as much as cash in his bank account:

$$\Delta\, dS(t) - dV(t, S(t)) \geq \Delta S(t) c\, dt - r V(t, S(t))\, dt \tag{2}$$

where c is the rate of cost for carrying the underlying and r the risk-free rate. For the time being, we assume that $V(t, S)$ is continuously differentiable with respect to the time variable and twice continuously differentiable with respect to spot, which guarantees:

$$dV(t, S(t)) = V_t(t, S(t))\, dt + V_S(t, S(t))\, dS(t) + \tfrac{1}{2}\sigma^2 S^2(t) V_{SS}(t, S(t))\, dt. \tag{3}$$

It is required for the issuer to pick $\Delta = V_S$ in order to fulfill (2) because the random fluctuation is of order $dt^{1/2}$ and is much larger than dt terms. Incorporating this choice and rearranging (2), we find an equivalent condition for the issuer:

$$\mathcal{L}V = V_t + cSV_S + \tfrac{1}{2}\sigma^2 S^2 V_{SS} - rV \leq 0. \tag{4}$$

At the same time, the value of the option must never fall below its immediate exercise value, because this would result in a simple arbitrage opportunity. This sets another restriction on the value function:

$$V \geq X(S) \tag{5}$$

where $X(S)$ denotes the exercise value. At each time t, the option holder may or may not exercise her option. If (5) holds strictly, then exercising the option at that instant is not the least favorable outcome for the issuer, because he can cash in the difference. In this case, (4) must vanish, because the issuer would have a "free lunch". Therefore we obtain the third condition:

$$(\mathcal{L}V) \cdot (V - X) = 0. \tag{6}$$

The inequalities (4),(5), and (6), together with terminal condition $V = X$, form a parabolic obstacle problem. We refer to Friedman (1988) for the existence and the uniqueness of the solution to such problems. Wilmott, Dewynne and Howison (1993) present a numerical method for solving the problem.

The curve that divides $V = X$ from $V > X$ to equal is often referred to as free boundary. As we explained, the free boundary for the above parabolic obstacle problem is the price-maximizing exercise boundary. Jaillet, Lamberton, and Lapeyre (1990) showed that the solution of the parabolic obstacle problem has a continuous gradient at the free boundary, and Van

Moerbeke (1976) showed that the free boundary is continuously differentiable. Thus, Itô's formula (3) is valid at least in a weak sense: see San Martin and Protter (1993) for details.

Figure 6 shows the fair value of an American put option *versus* the underlying. Where the fair value first touches the payoff is the price-maximizing free boundary.

In asset, time space we find a free boundary dividing the hold and exercise regions (see Figure 7).

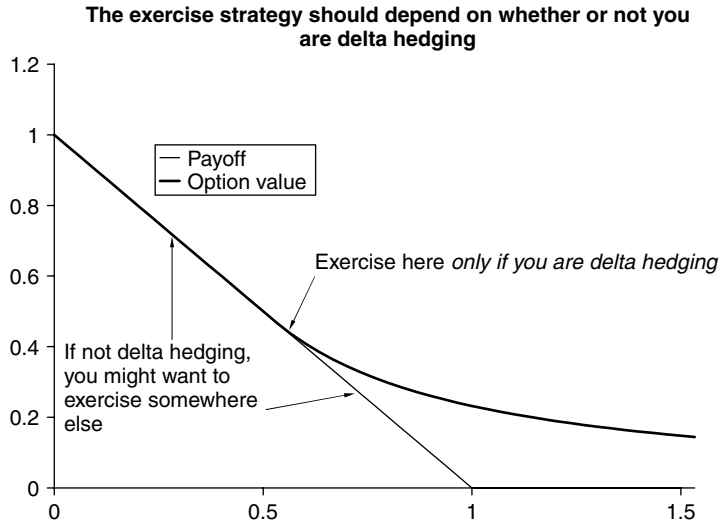


Figure 6: American put value vs. underlying

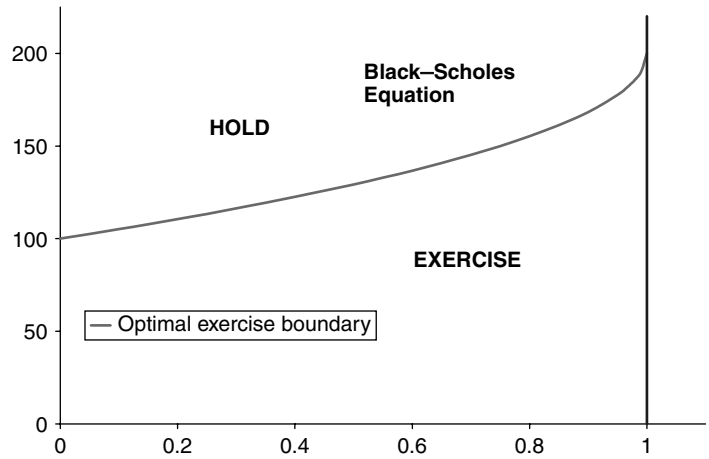


Figure 7: Underlying/time domain for an American option showing the free boundary dividing the exercise region from the hold region

6 Optimal exercise from the holder's perspective

Having dealt with the theoretically correct option-valuation problem we turn our attention to the problem as seen by the option holder. We present several possibilities for her exercise strategies. Remember that for the reasons explained above, should she wish to exit her position the only available possibility may be early exercise.

In order to provide concrete results we must first present a framework for a consistent and *quantitative* trading strategy.

As we mentioned earlier, feasible strategies for speculative traders are buying, selling, or exercising an option. The optimal strategy is predicated upon the trader's risk preference and both physical and risk-neutral dynamics of the spot. Choosing a risk preference is rather arbitrary, and in reality none of us knows how to postulate our own risk preference. However, one would ideally pick a systematic risk preference, because of the consistency of the decision and the automation of the decision-making process. We will assume that the trader's risk preference is governed by an expected value of a utility function u that is strictly increasing and twice continuously differentiable.

Assessing the dynamics of the spot, in practice, requires statistical analysis of the historical data, filtering the drift (as it is not observable), and calibration of the volatility structure. Here, we will keep the form (1) with:

$$dM(t) = b(t, M(t), S(t))\,dt + a(t, M(t), S(t))\,dZ(t) \tag{7}$$

where Z is a standard Brownian motion with $d[Z, W]_t = \rho\,dt$. If $\rho = 1$, $a = -\beta\sigma$, and:

$$b(t, M, S) = -\beta\left(M + \tfrac{1}{2}\sigma^2\right)$$

then the spot is essentially an exponential of an Ornstein–Uhlenbeck process. Another possible natural postulation is to make M itself mean-revert to a certain growth rate.

At time t, the investor, who longs an American option, faces the following optimal stopping problem:

$$U(t, S, M) = \sup_{\tau \geq t} E_t[u(Q(\tau, S(\tau)))] \tag{8}$$

where the discounted termination value Q is defined as:

$$Q(t, S) = e^{-\lambda t}(V(t, S)I_{B(S)} + X(S)I_{B^c(S)}).$$

As before, X is the exercise value and V is the market price of the option. In addition, λ is the accrual rate of the trader's capital, and B is the liquidity zone that depends upon the moneyness of the contract. For example, the boundary of B may simply be the same as the strike, meaning that when the contract becomes in-the-money, it is not traded. Finally, the conditional expectation is under the physical measure.

The optimal stopping problem (8) is not much different from the risk-neutral valuation of an American option, mathematically, and the maximum utility value function U is a solution of a parabolic obstacle problem. First:

$$U \geq u(Q). \tag{9}$$

This is because the maximum expected utility is not smaller than the utility of immediate exercise or liquidation, which is a special case of many feasible stopping times. Next, the following inequality holds for each $t < T - \delta$:

$$U(t, S, M) \geq E_t[U(t + \delta, S(t + \delta), M(t + \delta))]. \tag{10}$$

To see this, note that the right-hand side of (10) is the expected utility when the option holder pursues optimal stopping only after δ elapses. In other words, the trader is dormant until time $t + \delta$ and she tries to find an optimal stopping time from then on. Thus, the value of this expected utility shouldn't exceed the maximum expected utility, which is on the left-hand side of (10).

The implication of this in an infinitesimal time is the following:

$$\mathcal{L}_p U = U_t + bU_M + MSU_S + \tfrac{1}{2}a^2 U_{MM} + \rho a\sigma SU_{SM} + \tfrac{1}{2}\sigma^2 S^2 U_{SS} \leq 0 \tag{11}$$

which is obtained from Dynkin's formula. If $\mathcal{L}_p U < 0$, then the maximum expected utility is expected to fall in an infinitesimal time, and hence the optimal strategy is to exercise or sell the option immediately. Therefore:

$$(\mathcal{L}_p U)(U - u(Q)) = 0. \tag{12}$$

The set of variational inequalities (9), (11), and (12), together with terminal data $U(t, S, M) = u(X(S))$ characterize the maximum expected utility. The optimal stopping time is the first time the spot and its drift hit the free boundary of the inequalities. If the spot is in the liquid zone, the stopping means selling, exercising otherwise.

Expected utility is a consistent risk preference that separates different outcomes. When it comes to seeing what this means in dollar terms, however, it is not that transparent. In this case, it is easier to use the certainty equivalence, which is the cash amount that is indifferent from the random payoff of the optimal strategy:

$$H(t, S, M) = e^{\lambda t} u^{-1}(U(t, S, M)).$$

It can be shown that, by chain rule, the maximum utility certainty equivalent satisfies the following non-linear variational inequalities:

$$H(t, S, M) \geq V(t, S) I_{B(S)} + X(S) I_{B^c(S)} \tag{13}$$

$$\begin{aligned} \mathcal{N}_p H = {} & H_t + bH_M + MSH_S + \frac{1}{2}a^2 H_{MM} + \rho a\sigma SH_{SM} + \frac{1}{2}\sigma^2 S^2 H_{SS} \\ & - \lambda H + \frac{1}{2}\frac{u''}{u'} e^{-\lambda t}((aH_M)^2 + \rho a\sigma SH_M H_S + (\sigma SH_S)^2) \leq 0 \end{aligned} \tag{14}$$

$$(\mathcal{N}_p H(t, S, M))(H(t, S, M) - V(t, S) I_{B(S)} - X(S) I_{B^c(S)}) = 0. \tag{15}$$

Therefore the optimal strategy depends upon the physical drift and Pratt's measure of absolute risk aversion $-u''/u'$, and is different from the price-maximizing boundary. The distortion in discount is caused by the non-linearity of the utility function.

Figure 8 is for illustrative purposes only (and in just S, t space), but shows a possible utility-maximizing boundary.

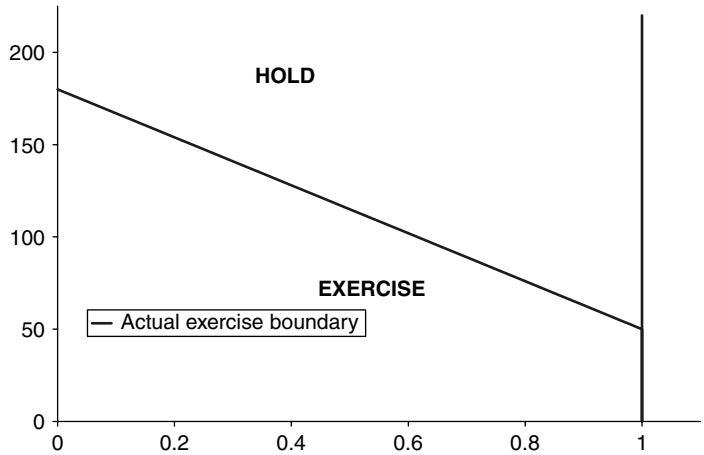


Figure 8: An example of the possible exercise boundary adopted by the option holder

7 The exercise boundary for the holder: examples

In the previous section, we have seen that the exercise boundary of the utility-maximizing strategy can be quite different from the classical exercise boundary. There are several factors that determine the behavior of the utility-maximizing exercise boundary. In fact, we will show that the factors are those we've already discussed in the introduction. Throughout this section, we will assume that the drift M is simply a constant, say μ, and ignore the scenario of selling the option. In this way, we can present our ideas without too many technical details.

Theorem *The utility-maximizing exercise time for an American option has the following properties:*

(i) If the absolute risk aversion is sufficiently large, then there is a positive probability of early exercise for both calls and puts.
(ii) The exercise time for a call option is non-decreasing in drift.
(iii) The exercise time for a put option is non-increasing in drift.

Proof Note that the exercise region coincides with the space-time domain of $\mathcal{N}H < 0$. If the absolute risk aversion $-u''/u'$ tends to infinity uniformly in its argument, then $\{(t, s) : \mathcal{N}H < 0, 0 \leq t \leq T, s > 0\}$ is a set of a positive measure. Since the support of a non-degenerate geometric Brownian motion (i.e., $\sigma^2 > 0$) occupies the entire positive plane, the utility-maximizing exercise time can be less than the maturity with a positive probability. This proves (i). When the option is a call, H_S is positive. Thus $\mathcal{N}H$ becomes more negative when the drift becomes smaller. If the option is a put, H_S is negative, and hence $\mathcal{N}H$ becomes more negative when the drift becomes larger. Therefore we have (ii) and (iii).

Our next task is to locate the boundary when the time to maturity is arbitrarily close to zero. Note that the certainty equivalence H tends to X as $t \to T$ and the utility-maximizing exercise boundary (as a function of time) is continuously differentiable. Thus when t is near T, the utility-maximizing exercise boundary is close to the boundary of $\mathcal{N}X < 0$. This is *the boundary at maturity*. If $X(S) = \max(S - K, 0)$ (i.e., a call option), then the boundary is above the strike K for each t and hence the boundary at maturity is:

$$\partial \left[S > K : \frac{1}{2}\sigma^2 S^2 \frac{u''}{u'}(e^{-\lambda T}(S - K))e^{-\lambda T} + (\mu - \lambda)S + \lambda K < 0 \right]. \tag{16}$$

Here, the symbol ∂ is used for indicating the boundary of a set. Similarly, if $X(S) = \max(K - S, 0)$ (i.e., a put option), the boundary at maturity is:

$$\partial \left[S < K : \frac{1}{2}\sigma^2 S^2 \frac{u''}{u'}(e^{-\lambda T}(K - S))e^{-\lambda T} - (\mu - \lambda)S - \lambda K < 0 \right]. \tag{17}$$

Sometimes (16) and (17) may contain more than a point. In such a case, the free boundary bifurcates.

In the remainder of this section, we provide an explicit expression for the boundary at maturity when the option holder's utility belongs to one of the following categories: the CARA, the HARA, and expected return (i.e., linear utility).

Constant absolute risk aversion, CARA

CARA is when the absolute risk aversion is a constant regardless of the wealth of the investor. That is, $-u''/u' \equiv \gamma$ for a positive constant γ. Up to a constant, the utility is of the form $u(\omega) = -\alpha e^{-\gamma\omega}$ for a positive constant α.

First we consider a call option. Expression (16) becomes:

$$\max\left(K, \frac{1}{\gamma\sigma^2}(\mu - \lambda + \sqrt{(\mu - \lambda)^2 + 2\gamma\sigma^2 K\lambda e^{-\lambda T}}e^{\lambda T}\right). \tag{18}$$

Note that (18) tends to infinity as γ tends to zero. Hence as the risk aversion of the option holder vanishes, the utility-maximizing exercise time tends to the maturity which coincides with the classical exercise time for call provided that the carrying cost is the same as the risk-free rate. Next we consider a put option. The inequality in (17) is:

$$-\tfrac{1}{2}\sigma^2\gamma e^{-\lambda T}S^2 - (\mu - \lambda)S - \lambda K < 0. \tag{19}$$

If the physical drift exceeds the holder's time value of cash ($\mu \geq \lambda$), (19) is true for all positive S. Thus the boundary at maturity is K. Suppose that $\mu < \lambda$. The quadratic inequality (19) is always satisfied if:

$$D = (\mu - \lambda)^2 - 2\gamma\sigma^2 K\lambda e^{-\lambda T} < 0.$$

In this case, the boundary at maturity is also K. Finally, if $D \geq 0$ as well as $\mu < \lambda$, the boundary at maturity is:

$$\min\left(K, \frac{1}{\gamma\sigma^2}\left(\lambda - \mu + \sqrt{(\lambda - \mu)^2 - 2\gamma\sigma^2 K\lambda e^{-\lambda T}}e^{\lambda T}\right)\right).$$

Hyperbolic absolute risk aversion, HARA

Merton (1990) provides a complete description of this family of utility functions. The hyperbolic absolute risk aversion means $-u''/u' = \gamma/(\omega + \alpha)$ for a positive constant γ. This utility applies in the case when the wealth of the investor is bounded below: $\omega + \alpha > 0$. Thus the richer the investor is, the less risk averse. Up to a constant shift:

$$U(\omega) = \begin{cases} \dfrac{1}{\beta^\gamma}\dfrac{(\omega+\alpha)^{1-\gamma}}{1-\gamma}, & \text{if } \gamma \neq 1 \\ \dfrac{1}{\beta}\log(\omega+\alpha), & \text{else} \end{cases}$$

where $\beta > 0$. The parameter α is assumed positive as the option payoff could be zero. Simple algebra reduces the inequalities in (16) and (17) to quadratic inequalities. For example, (16) is equivalent to:

$$\partial[S > K : AS^2 + BS + C < 0]$$

where $A = (\mu - \lambda - \sigma^2\gamma/2)e^{-\lambda T}$, $B = (\mu - \lambda)(\alpha - e^{-\lambda T}K) + \lambda e^{-\lambda T}K$, and $C = \lambda K(\alpha - e^{-\lambda T}K)$. The continuation region and the exercise boundary depend upon the choice of parameters.

An unusual case is when the parameters satisfy the following:

$$\lambda + \frac{1}{2}\sigma^2\gamma < \mu < \frac{1}{2}\sigma^2\gamma\frac{e^{-\lambda T}K}{\alpha}.$$

In this case, the continuation region near maturity is separated by the exercise region:

$$\left[S : K < S < \frac{-B + \sqrt{B^2 - 4AC}}{2A}\right].$$

If the physical drift is sufficiently large, the option is very valuable to the holder when the option is very in-the-money. If not, the curvature reduces the holder's utility. Also note that there is no exercise boundary if:

$$\mu > \lambda + \tfrac{1}{2}\sigma^2\gamma \qquad \text{and} \qquad \alpha > e^{-\lambda T}K.$$

This is the case when the physical drift is large while the risk aversion is not.

Expected return

This is a special case of $u(\omega) = \alpha\omega + \beta$ for a positive constant α. As u'' vanishes in this case, our analysis of the boundary at maturity becomes straightforward. When the option is a call, the inequality in (16) becomes $(\mu - \lambda)S + \lambda K < 0$. This is never satisfied if $\mu \geq \lambda$. In this case, the utility-maximizing exercise time is the maturity. If $\mu < \lambda$, on the other hand, the boundary at maturity is:

$$\max\left(K, \frac{\lambda}{\lambda - \mu}K\right).$$

Next we consider a put option. If $\mu \geq \lambda$, then the inequality in (17) is always satisfied, and the boundary at maturity is K. If $\mu < r$, the boundary at maturity becomes:

$$\min\left(K, \frac{\lambda}{\lambda - \mu}K\right).$$

8 The hedger's windfall profit

We have observed that the option holder's exercise time can differ from the price-maximizing exercise time for various reasons. When this happens, the issuer gains from the difference. In this section, we examine the profit from selling American options to utility-maximizing investors. To avoid the complication of modeling investors with different risk preference, we will simply exclude the case when the ownership of the option changes before the expiration.

The issuer charges $V(0, S(0))$ at time 0, hedges his short position until the option is exercised or expired. The discounted potential liability of the issuer is $e^{-r\tau}X(S(\tau))$ where τ is the termination time, either early exercise or expiration. The present value of the issuer's profit becomes:

$$V(0, S(0)) + \int_0^\tau e^{-rt}\Delta(dS(t) - cS(t)\,dt) - e^{-r\tau}X(S(\tau)). \tag{20}$$

The second term in the equation (20) is the result of delta hedging with the cost of carry.

First we add and subtract $e^{-r\tau}V(\tau, S(\tau))$ from the profit (20). Applying Itô's formula to V yields:

$$\begin{aligned} V(0, S(0)) &+ \int_0^\tau e^{-rt}\Delta(dS(t) - cS(t)\,dt) - e^{-r\tau}V(\tau, S(\tau)) \\ &= -\int_0^\tau dt e^{-rt}\mathcal{L}V \end{aligned}$$

where $\mathcal{L}$ is as defined as in (4). Thus, we may rewrite the profit (20) as:

$$-\int_0^\tau dt e^{-rt}\mathcal{L}V + e^{-r\tau}(V(\tau, S(\tau)) - X(S(\tau))).$$

We define the expected profit at time t as:

$$\Psi(t, S, M) = E_t\left[-\int_0^\tau dt e^{-rt}\mathcal{L}V + e^{-r\tau}(V(\tau, S(\tau)) - X(S(\tau)))\right].$$

We will show that Ψ satisfies a diffusion equation with a moving boundary which is known *a priori*. Recall that we ignore the ownership changes. Thus, if H is the maximum expected certainty equivalence of the option holder, then its free boundary gives the optimal exercise time τ. Let $\mathcal{H}$ and $\mathcal{V}$ be the domains defined by $H > X$ and $V > X$, respectively. These are the regions of continuation for the utility maximization and the price maximization. We also

define $\mathcal{G} = \mathcal{H} \backslash \mathcal{V}$. Since $\mathcal{L}V$ vanishes on $\mathcal{V}$, the expected profit Ψ satisfies:

$$\mathcal{L}_p \Psi - e^{-rt} \mathcal{L} V I_{\mathcal{G}} = 0 \tag{21}$$

subject to $\Psi(T, S, M) = 0$ and $\Psi = e^{-rt}(V - X)$ on $\partial\mathcal{H}$, the utility-maximizing exercise boundary. In the above, $\mathcal{L}_p$ is defined as in (11). The indicator $I_{\mathcal{G}}$ is one if (t, s) belongs to $\mathcal{G}$ and zero otherwise. If the option is a call and the cost of carry is the same as the risk-free rate, the source term of (21) vanishes because $\mathcal{G}$ is empty. If the option is a put, then $V = X$ on the complement of $\mathcal{V}$, and therefore:

$$e^{-rt} \mathcal{L} V I_{\mathcal{G}} = -r e^{-rt} K I_{\mathcal{G}}$$

where K is the strike price. Here we have used the fact that the price-maximizing exercise boundary is not above the strike when the option is a put.

Figure 9 combines both the price-maximizing and utility-maximizing boundaries on the same graph, and shows the region in which the writer's windfall profit is the sudden excess and the gradually incremented excess.

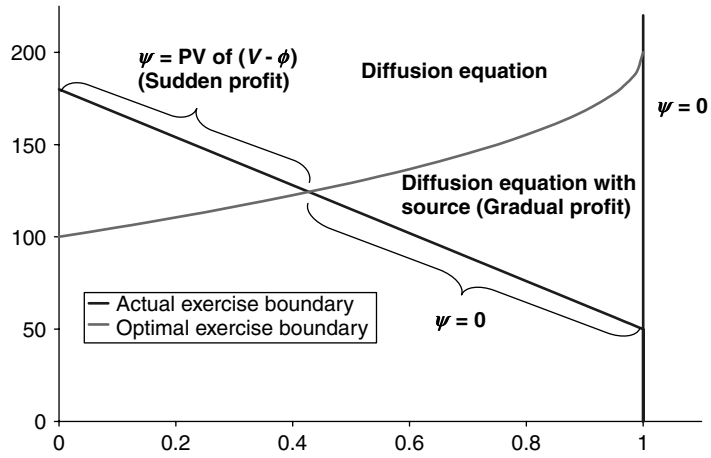


Figure 9: Underlying/time domain showing the two boundaries (price optimal and utility optimal), and the relevant governing equations/boundary conditions

Example 1: Maximization of the expected profit

Figure 10 shows the expected profit from selling an at-the-money American put to an investor who maximizes expected profit. In this case, the option holder's criterion in choosing the exercise time is free from risk aversion, and hence the outcome can be considered as the marginal effect of the physical drift to the issuer's expected gain. The initial asset price is 50, the asset volatility is 20% per annum, the maturity of the option is six months, and the risk-free rate, carrying cost rate, and the option holder's discount rate are set to 8% per annum. When the physical drift coincides with this rate, the holder's exercise boundary is inside the price-maximizing exercise boundary. In this case, $\mathcal{G}$ is empty and the only source of the issuer's

profit is the difference between the value of the option and the exercise value (i.e., the value of P on the moving boundary $\partial\mathcal{H}$). If the physical drift is less than the rate, then the holder's exercise boundary is outside of the price-maximizing exercise boundary, and hence the issuer's profit grows with the occupation time of the asset price in between the two boundaries. This explains the asymmetry in the picture.

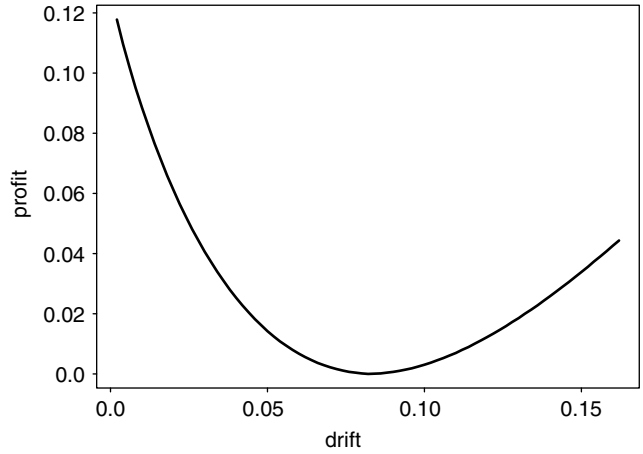


Figure 10: Expected windfall profit vs. growth of the underlying

Example 2: Maximization of expected CARA utility

Figure 11 is the issuer's expected profit as a function of the absolute risk aversion. The option holder's exercise time maximizes the expected CARA utility, while the physical drift coincides with the rate (8%). Thus, the outcome is the marginal effect of the absolute risk aversion to the issuer's expected profit. Again, the option is an at-the-money American put, and parameters are set as before. If the absolute risk aversion vanishes and the physical drift coincides with the rate, then the utility-maximizing exercise boundary coincides with the price-maximizing one, and hence the expected profit vanishes.

Example 3: The chooser range note

We end with a few words on the Chooser Range Note whose term sheet we saw earlier.

This contract requires the holder to make four decisions during its life. Each of these four decisions involves choosing the mid point of an interest rate range, a continuous spectrum of possibilities.

In Figure 12 we see the forward rate curve as it might be at the start of the contract's life. The shape of this curve is more often than not upward sloping, representing adjustment for the price of risk. One expects a higher return for holding something for a longer term.

The figure also shows a possible evolution of short-term interest rates. It is this path which determines, in part, the final payoff. Notice how the path of rates does not follow the forward curve. Obviously it is stochastic, but it does not usually exhibit the rapid growth at the short end.

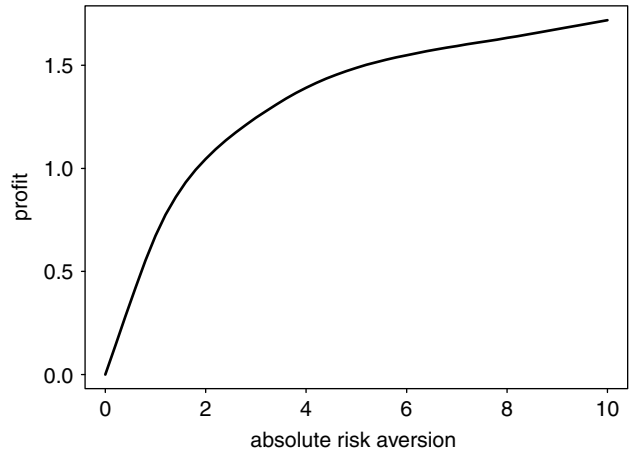


Figure 11: Expected windfall profit vs. risk aversion (CARA utility)

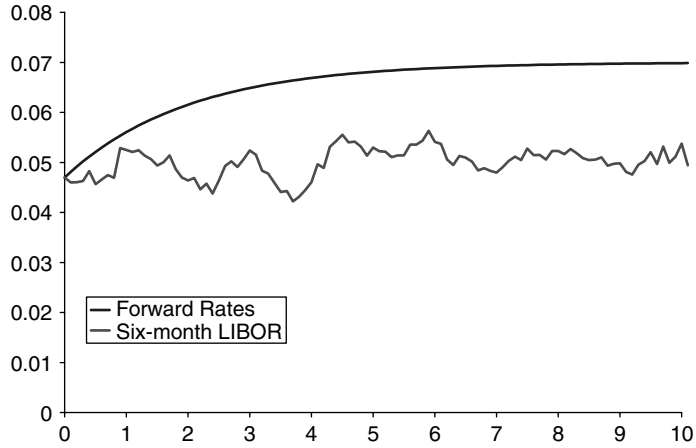


Figure 12: Typical forward rate curve at start of contract's life, and typical evolution of actual short-term interest rates over its life

Figure 13 shows a plausible choice of price-maximizing ranges. These will naturally be dependent upon the forward curve (the figure is schematic only. The actual ranges 'chosen' by the writer when maximizing the price will depend on the volatility of interest rates as well. But this is not the place to go into detail about the pricing of such contracts.)

Figure 14 shows the ranges as chosen by the holder of the contract. She makes a decision about each range at the start of each new period. Of course, her choice will be closely related to where the short-term rate is at that time, with some allowance for her view.

Clearly there is great scope for a significant difference between the price-maximizing choice and the final choices made by the holder. Our concept applies equally well to this case as to

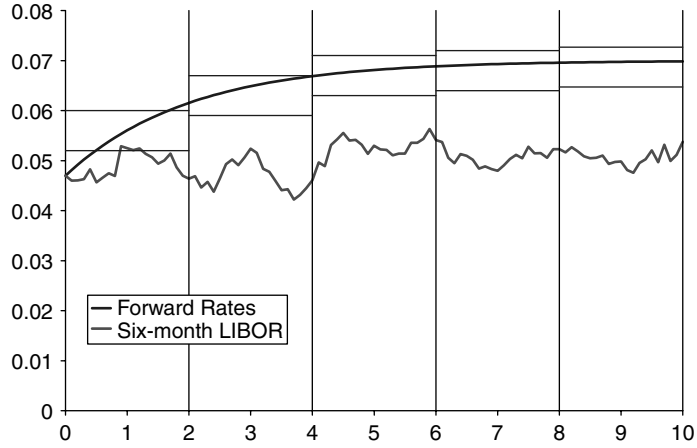


Figure 13: The price-maximizing ranges will depend on the risk-neutral, forward rate curve (schematic only; the choice will also depend on the volatility of the curve)

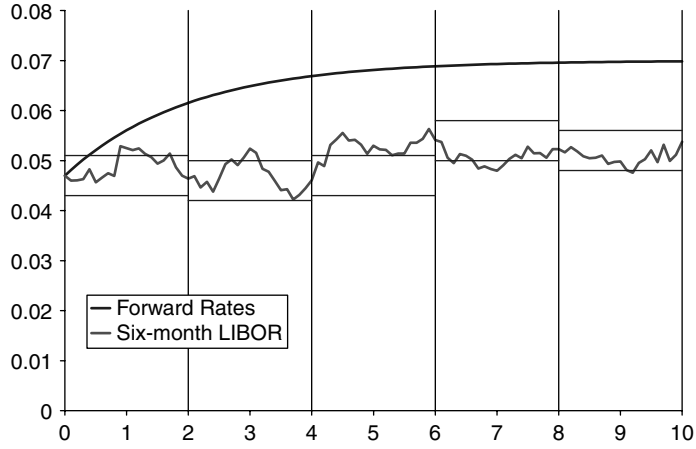


Figure 14: The ranges chosen by the holder are more likely to represent the best guess at the evolution of actual rates

the exercise of American options; the writer of the option can expect a windfall profit which depends on the difference between the holder's strategy and the price-maximizing strategy.

Conclusion

The theory of optimal stopping has been applied to the valuation of American options. People are prone to misuse the terminologies of the theory of optimal stopping when they talk about American options. For example, the price-maximizing exercise boundary has been referred to

as the optimal exercise boundary, while it is optimal to neither the issuer nor the option holder. This causes confusion to students, practitioners, and even academic researchers in the field.

In this article, we have presented various reasons why American options may be exercised on boundaries other than the one that is assumed in the fair pricing. The hedger can gain from the discrepancy between the actual exercise time and the price-maximizing exercise time. The concept applies to many OTC structure products as well: users of these products are forced to use strategies other than maintaining a risk-free portfolio, or voluntarily take risk, pursuing their own rational strategies. Financial institutions, who provide these custom-made products, can capitalize on the difference between the classical theory and the rational practice.

A relevant issue, which we haven't discussed here, is whether the issuer should report his risk based upon the worst-case scenario. Mortgage backed securities with prepayment are cases in point. Should a bank report prepayment risk based upon the scenario that every mortgage owner behaves optimally? Or should a power marketer report attrition risk based upon the scenario that every client performs optimal gaming? When was the last time you switched your utility provider or long-distance communication provider?

FOOTNOTE & REFERENCES

1. This is definitely not political correctness. In our examples the holder of the option is female and the writer male, to aid our descriptions.

■ Ahn, H., Penaud, A. and Wilmott, P. (2000) Various passport options and their valuation. *Applied Mathematical Finance* 6, 275–292.
■ Friedman, A. (1988) *Variational Principles and Free Boundary Problems*. Robert E. Krieger Publishing, New York.
■ Harrison, J. M. and Pliska, S. R. (1981) Martingales and stochastic integrals in the theory of continuous trading. *Stoch Proc Appl*, 11, 215–260.
■ Jaillet, P., Lamberton, D. and Lapeyre, B. (1990) Variational inequalities and the pricing of American options. *Acta Appl Math*, 21, 263–289.
■ Merton, R. C. (1990) *Continuous-Time Finance*. Blackwell, Oxford.
■ Myneni, R. (1992) The pricing of the American option. *Annals Appl. Probab.* 2(1), 1–23.
■ Overdahl, J. A. and Martin, P. G. (1994) The exercise of equity options: theory and empirical tests. *J. Derivatives, Fall*, 38–51.
■ San Martin, J. and Protter, P. (1993) General change of variables formulas for semi-martingales in one and finite dimensions. *Probab. Theory Rel. Fields* 97, 363–381.
■ Van Moerbeke, P. (1976) On optimal stopping and free boundary problems. *Arch. Rational Mech. Anal.* 60, 101–148.
■ Wilmott, P., Dewynne, J. N. and Howison, S. D. (1993) *Option Pricing: Mathematical Models and Computation*. Oxford Financial Press, Oxford.

17
Phi-alpha Optimal Portfolios and Extreme Risk Management

R. Douglas Martin,*† Svetlozar (Zari) Rachev,†# and Frederic Sibouletˆ**

Wilmott magazine, November 2003

> When anyone asks me how I can describe my experience of nearly forty years at sea, I merely say uneventful. Of course there have been winter gales and storms and fog and the like, but in all my experience, I have never been in an accident of any sort worth speaking about. I have seen but one vessel in distress in all my years at sea (...) I never saw a wreck and have never been wrecked, nor was I ever in any predicament that threatened to end in disaster of any sort.
>
> E. J. Smith, Captain, 1907, RMS Titanic

High market volatility demands new solutions

Paul Wilmott likes to recount the ritual by which he questions his students on the likelihood of Black Monday 1987. Under the commonly accepted Gaussian risk factor distribution assumption, they calculate that there should be no such event in the entire existence of the universe and beyond!

The last two decades have witnessed a considerable increase in the volatility of financial markets – dramatically so in the last few years. Extreme events are the corollary of that increased volatility.

Contact addresses: * University of Washington, Department of Statistics, Seattle, WA, USA.
** University of California, Applied Probabilities, Santa Barbara, CA, USA.
† Fin Analytica, 225 East 95th Street #33k, New York, NY 10128, USA.
E-mail: frederic.siboulet@finanalytica.com Telephone: 001 646 244 4462 www.finanalytica.com
University of Karlsruhe, Chair of Statistics, Germany.

Legacy risk systems have done a reasonable job at managing *ordinary* financial events. However up to now, very few institutions or vendors have demonstrated the systematic ability to deal with the *unusual* event, the one that should *almost never* happen. Therefore, one can reasonably question the soundness of some of the current risk management practices and tools used in Wall Street as far as extreme risk is concerned.

The two principal approaches to modeling asset returns are based either on Historical or on Normal (Gaussian) distribution. Neither approach adequately captures unusual asset price and return behaviors. The *Historical* model is bounded by the scope of the available observations and the *Normal* model inherently cannot produce atypical returns. The financial industry is beleaguered with both under-optimized portfolios with often-shabby ex-post risk-adjusted returns, as well as deceptive aggregate risk indicators (e.g. VaR) that lead to substantial unexpected losses.

The inadequacy of the Normal distribution is well recognized by the risk management community. Yet up to now, no consistent and comprehensive alternative probability models had adequately addressed unusual returns. To quote one major vendor:

> It has often been argued that the true distributions returns (even after standardizing by the volatility) imply a larger probability of extreme returns than that implied from the Normal distribution. Although we could try to specify a distribution that fits returns better, it would be a daunting task, especially if we consider that the new distribution would have to provide a good fit across all asset classes. (Technical Manual, RMG, 2001).

In response to the challenge, we use Stable risk-factor distributions and generalized risk-factor dependencies, thereby creating a *paradigm shift* to consistent and uniform use of the most viable class of non-Normal probability models in finance. This approach leads to distinctly improved financial risk management and portfolio optimization solutions for highly volatile markets with extreme events.

The stable distribution framework

Stable distributions

In spite of wide-spread awareness that most risk factor distributions are heavy-tailed, to date, risk management systems have essentially relied either on historical, or on univariate and multivariate Normal (or Gaussian) distributions for Monte Carlo scenario generation. Unfortunately, historical scenarios only capture conditions actually observed in the past, and in effect use empirical probabilities that are zero outside the range of the observed data, a clearly undesirable feature. On the other hand Gaussian Monte Carlo scenarios have probability densities that converge to zero too quickly (exponentially fast) to accurately model real-world risk factor distributions that generate extreme losses. When such large returns occur separately from the bulk of the data they are often called outliers.

Figure 1 shows quantile–quantile (qq)-plots of daily returns versus the best-fit Normal distribution of nine randomly selected Micro-cap stocks for the two-year period 2000–2001. If the returns were Normally distributed, the quantile points in the qq-plots would all fall close to a straight line. Instead they all deviate significantly from a straight line (particularly in the tails), reflecting a higher probability of occurrence of extreme values than predicted by the Normal distribution, and showing several outliers.

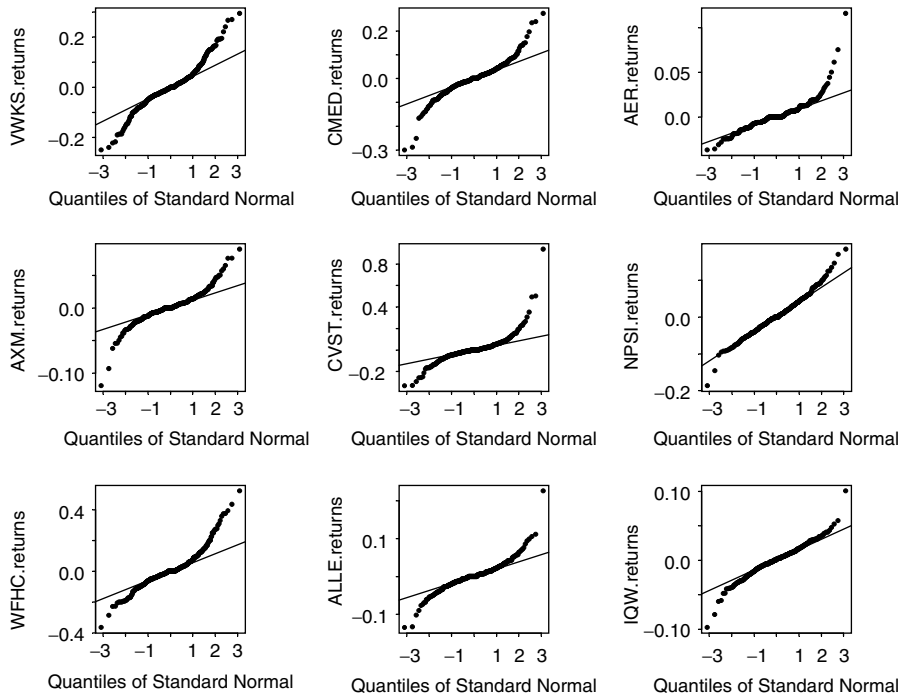


Figure 1

Such behavior occurs in many asset and risk factor classes, including well-known indices such as the S&P 500, and corporate bond prices. The latter are well known to have quite non-Gaussian distributions that have substantial negative skews to reflect down-grading and default events. For such returns, non-Normal distribution models are required to accurately model the tail behavior and compute probabilities of extreme returns.

Various non-Normal distributions have been proposed for modeling extreme events, including:

- Mixtures of two or more Normal distributions.
- t-distributions, hyperbolic distributions, and other scale mixtures of normal distributions.
- Gamma distributions.
- Extreme Value distributions.
- Stable non-Gaussian distributions (also known as Lévy-Stable and Pareto-Stable distributions).

Among the above, only Stable distributions have attractive enough mathematical properties to be a viable alternative to Normal distributions in trading, optimization and risk management systems. A major drawback of all alternative models is their lack of stability. Benoit Mandelbrot (1963) demonstrated that the stability property is highly desirable for asset returns. These advantages are particularly evident in the context of portfolio analysis and risk management.

An attractive feature of Stable models, not shared by other distribution models, is that they allow generation of Gaussian-based financial theories and, thus allow construction of a coherent

and general framework for financial modeling. These generalizations are possible only because of specific probabilistic properties that are unique to (Gaussian and non-Gaussian) Stable laws, namely; the Stability property, the Central Limit Theorem, and the Invariance Principle for Stable processes.

Benoit Mandelbrot (1963), then Eugene Fama (1963), provided seminal evidence that Stable distributions are good models for capturing the heavy-tailed (leptokurtic) returns of securities. Many follow-on studies came to the same conclusion, and the overall Stable distributions theory for finance is provided in the definitive work of Rachev and Mittnik (2000).

But in spite the convincing evidence, Stable distributions have seen virtually no use in capital markets. There have been several barriers to the application of stable models, both conceptual and technical:

- Except for three special cases, described below, Stable distributions have no closed form expressions for their probability densities.
- Except for Normal distributions, which are a limiting case of Stable distributions (with $\alpha = 2$ and $\beta = 0$), Stable distributions have infinite variance and only a mean value for $\alpha > 1$.
- Without a general expression for stable probability densities, one cannot directly implement maximum likelihood methods for fitting these densities, even in the case of a single (univariate) set of returns.

The availability of practical techniques for fitting univariate and multivariate stable distributions to asset and risk factor returns has been *the* barrier to the progress of Stable distributions in finance. Only the recent development of advanced numerical methods has removed this obstacle. These patented methods form the foundation of the ***Cognity***™ market & credit risk management and portfolio optimization solution (see further comments in the concluding section).

Univariate Stable distributions A Stable distribution for a random risk factor X is defined by its characteristic function:

$$F(t) = E(e^{itX}) = \int e^{itX} f_{\mu,\sigma}(x)\, dx,$$

where:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

is any probability density function in a location-scale family for X:

$$\log F(t) = \left\{ \begin{array}{l} -\sigma^{\alpha}|t|^{\alpha}\left(1 - i\beta \operatorname{sgn}(t) \tan\left(\frac{\pi\alpha}{2}\right)\right) + i\mu t, \quad \alpha \neq 1 \\ -\sigma|t|\left(1 - i\beta\frac{2}{\pi}\operatorname{sgn}(t)\log|t|\right) + i\mu t, \quad \alpha = 1 \end{array} \right\}$$

A stable distribution is therefore determined by the four key parameters:

1. α determines density's kurtosis with $0 < \alpha \leq 2$ (e.g. tail weight).
2. β determines density's skewness with $-1 \leq \beta \leq 1$.

3. σ is a scale parameter (in the Gaussian case, $\alpha = 2$ and $2\sigma^2$ is the variance).
4. μ is a location parameter (μ is the mean if $1 < \alpha \leq 2$).

Stable distributions for risk factors allow for skewed distributions when $\beta \neq 0$ and fat tails relative to the Gaussian distribution when $\alpha < 2$. Figure 2 shows the effect of α on tail thickness of the density as well as peakedness at the origin relative to the Normal distribution (collectively the "kurtosis" of the density), for the case of $\beta = 0$, $\mu = 0$, and $\sigma = 1$. As the values of α decrease the distribution exhibits fatter tails and more peakedness at the origin.

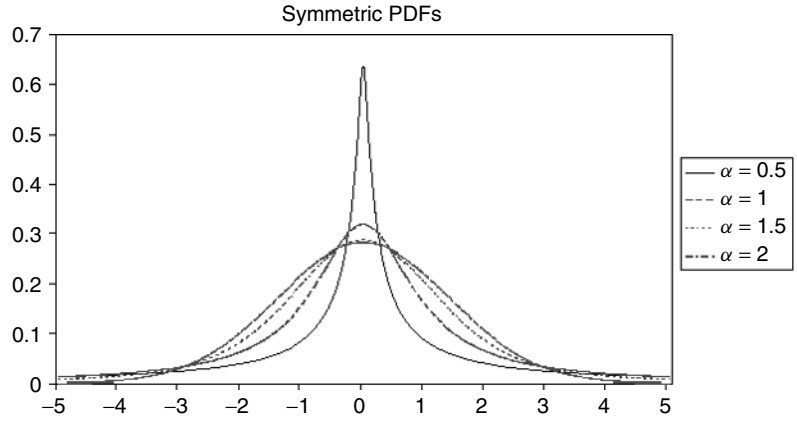


Figure 2

The case of $\alpha = 2$ and $\beta = 0$ and with the reparameterization in scale, $\tilde{\sigma} = \sqrt{2}\sigma$, yields the Gaussian distribution, whose density is given by:

$$f_{\mu\tilde{\sigma}}(x) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} e^{-\frac{(x-\mu)^2}{2\tilde{\sigma}^2}}.$$

The case $\alpha = 1$ and $\beta = 0$ yields the Cauchy distribution with much fatter tails than the Gaussian, and is given by:

$$f_{\mu,\sigma}(x) = \frac{1}{\pi \cdot \sigma} \left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)^{-1}$$

Figure 3 illustrates the influence of β on the skewness of the density for $\alpha = 1.5$, $\mu = 0$ and $\sigma = 1$. Increasing (decreasing) values of β result in skewness to the right (left).

Fitting Stable and Normal distributions: DJIA example. Aside from the Gaussian, Cauchy, and one other special case of stable distribution for a positive random variable with $\alpha = 0.5$, there is no closed form expression for the probability density of a Stable random variable.

Thus one is not able to directly estimate the parameters of a Stable distribution by the method of maximum likelihood. To estimate the four parameters of the stable laws, the ***Cognity***™

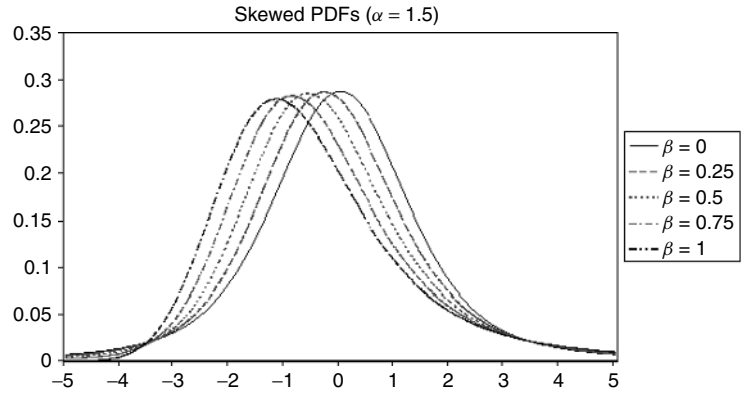


Figure 3

solution uses a special patent-pending version of the FFT (Fast Fourier Transform) approach to numerically calculate the densities with high accuracy, and then applies MLE (Maximum Likelihood Estimation) to estimate the parameters.

The results from applying the ***Cognity***^{TM} Stable distribution modeling to the DJIA daily returns from 1 January 1990 to 14 February 2003 is displayed in Figure 4. In both cases a GARCH model has been used to account for the clustering of volatility.

Figure 4 shows the left-hand tail detail of the resulting stable density, along with that of a Normal density fitted using the sample mean and sample standard deviation, and that of a non-parametric kernel density estimate (labeled "Empirical" in the plot legend). The parameter estimates are:

- Stable parameters $\hat{\alpha} = 1.699$, $\hat{\beta} = -0.120$, $\hat{\mu} = 0.0002$, and $\hat{\sigma} = 0.006$.
- Normal density parameter estimates $\hat{\mu} = 0.0003$, and $\hat{\sigma} = 0.010$.

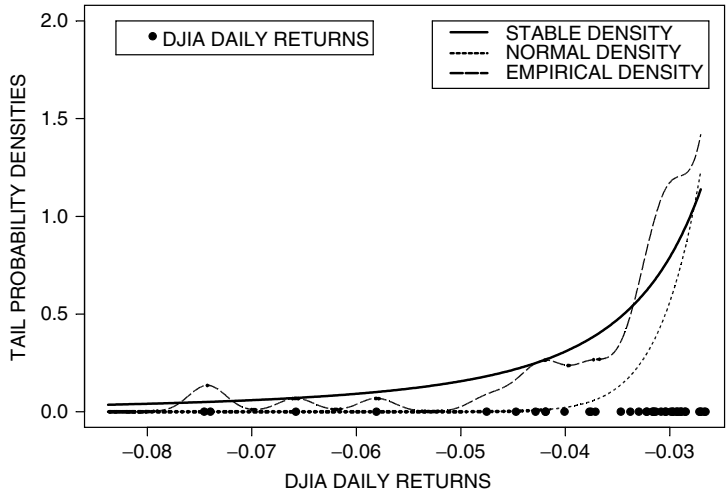


Figure 4

Note that the Stable density tail behavior is reasonably consistent with the Empirical nonparametric density estimate, indicating the existence of some extreme returns. At the same time it is clear from the figure that the tail of the Normal density is much too thin, and will provide inaccurate estimates of tail probabilities for the DJIA returns. Table 1 shows just how bad the Normal tail probabilities are for several negative returns values.

TABLE 1: PROBABILITY (DJIA RETURN < X)

x	-0.04	-0.05	-0.06	-0.07
Stable Fit	0.0066	0.0043	0.0031	0.0023
Normal Fit	0.000056	0.0000007	3.68E-09	7.86E-12

A daily return smaller than -0.04 with the Stable distribution occurs with probability 0.0066, or roughly seven times every four years, whereas such a return with the Normal fit occurs on the order of once every four years.

Similarly, a return smaller than -0.05 with the Stable occurs about once per year and with the Normal fit about once every forty years. Clearly the Normal distribution fit is an exceedingly optimistic predictor of DJIA tail return values.

Figure 5 displays the central portion of the fitted densities as well as the tails, and shows that the Normal fit is not nearly peaked enough near the origin as compared with the empirical density estimate (even though the GARCH model was applied), while the stable distribution matches the empirical estimate quite well in the center as well as in the tails.

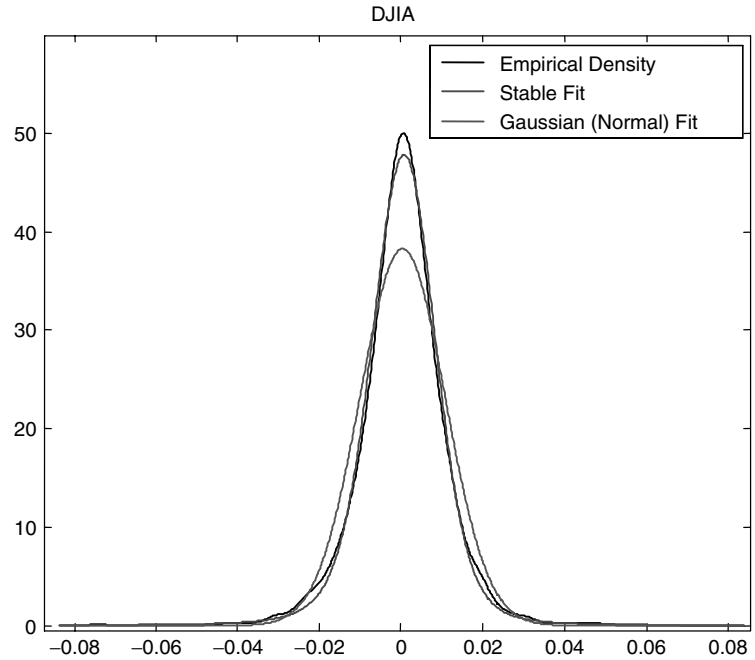


Figure 5

Fitting Stable distributions: micro-caps example. Noting that micro-cap stock returns are consistently strongly non-normal (see sample of normal qq-plots at the beginning of this section), we fit stable distributions to a random sample of 182 micro-cap daily returns for the two-year period 2000–2001. The results of the 95% confidence interval for the estimation of the tail weight parameter alpha are displayed in the boxplot in Figure 6.

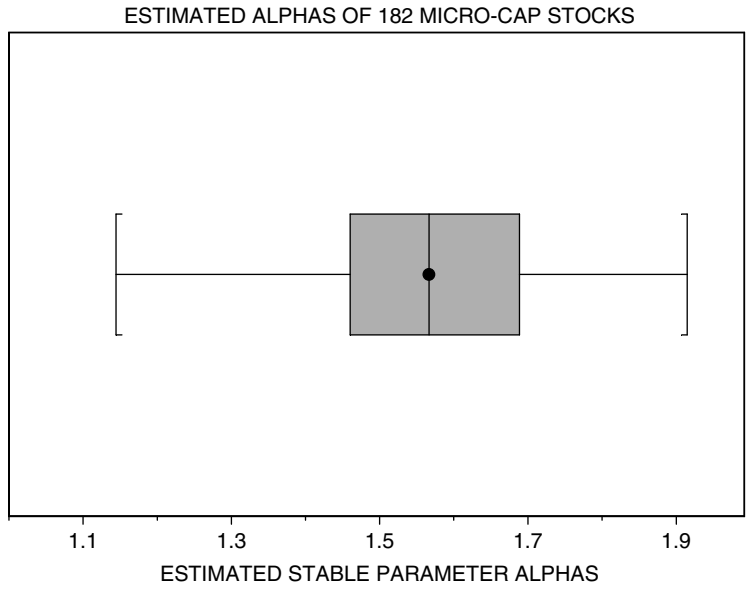


Figure 6

The median of the estimated alphas is 1.57, and the upper and lower quartiles are 1.69 and 1.46 respectively. Somewhat surprisingly, the distribution of the estimated tail weight parameter alpha turns out to be quite Normal.

Multivariate Stable distribution modeling Multivariate Stable distribution modeling involves univariate Stable distributions for each risk factor, each with its own parameter estimates $\hat{\alpha}_i, \hat{\beta}_i, \hat{\mu}_i, \hat{\sigma}_i, i = 1, 2, \cdots, K$, where K is the number of risk factors, along with a dependency structure.

One way to produce the dependency structure is through a subordinated process approach as follows. First compute a robust mean vector and covariance matrix estimate of the risk factors by trimming a small percentage of the observations on a coordinate-wise basis (to get rid of the outliers, and have a good covariance estimate for the central bulk of the data). Next you generate multivariate normal scenarios with this mean vector and covariance matrix. Then you multiply each random variable component of the scenarios by a Stable subordinator which is a strictly positive Stable random variable with index $\hat{\alpha}_i/2$, $i = 1, 2, \cdots, K$. The vector of subordinators is usually independent of the normal scenario vectors, but it can also be dependent. See, for example, Rachev and Mittnik (2000), and Rachev, Schwartz and Khindanova (2003).

Another very promising approach to building the cross-sectional dependence model is through the use of copulas, an approach that is quite attractive because it allows for modeling

higher correlations during extreme market movements, thereby accurately reflecting lower portfolio diversification at such times. The next section briefly discussion copulas.

Copula multivariate dependence models

Why copulas? Classical correlations and covariances are quite limited measures of dependence, and are only adequate in the case of multivariate Gaussian distributions. A key failure of correlations is that, for non-Gaussian distributions, zero correlation does not imply independence, a phenomenon that arises in the context of time-varying volatilities represented by ARCH and GARH models. The reason we use copulas is that we need more general models of dependence, ones which:

- Are not tied to the elliptical character of the multivariate normal distribution.
- Have multivariate contours and corresponding data behavior that reflect the local variation in dependence that is related to the level of returns, in particular, those shapes that correspond to higher dependence for extreme values of two or more of the returns.

What are copulas? A copula may be defined as a multivariate cumulative distribution function with uniform marginal distributions:

$$C(u_1, u_2, \cdots, u_n), \;\; u_i \in [0, 1] \text{ for } i = 1, 2, \cdots, n$$

where:

$$C(u_i) = u_i \text{ for } i = 1, 2, \cdots, n.$$

It is known that for any multivariate cumulative distribution function:

$$F(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots X_n \leq x_n)$$

there exists a copula C such that:

$$F(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n))$$

where the $F_i(x_i)$ are the marginal distributions of $F(x_1, x_2, \cdots, x_n)$, and conversely for any copula C the right-hand-side of the above equation defines a multivariate distribution function $F(x_1, x_2, \cdots, x_n)$. See, for example, Bradley and Taqqu (2001) and Embrechts *et al* (2003).

The main idea behind the use of copulas is that one can first specify the marginal distributions in whatever way makes sense, e.g. fitting marginal distribution models to risk factor data, and then specify a copula C to capture the multivariate dependency structure in the best suited manner.

There are many classes of copula, particularly for the special case of bivariate distributions. For more than two risk factors beside the traditional Gaussian copula, the t-copula is very tractable for implementation and provides a possibility to model dependencies of extreme events. It is defined as:

$$C_{\nu,c}(u_1, u_2, \cdots, u_n) = \frac{\Gamma((\nu + n)/2)}{\Gamma(\nu/2)\sqrt{|\mathbf{c}|(\nu\pi)^n}} \int_{-\infty}^{t_\nu^{-1}(u_1)} \cdots \int_{-\infty}^{t_\nu^{-1}(u_n)} \left(1 + \frac{\mathbf{s}\mathbf{c}^{-1}\mathbf{s}}{\nu}\right) d\mathbf{s}$$

where $\mathbf{c}$ is a correlation matrix.

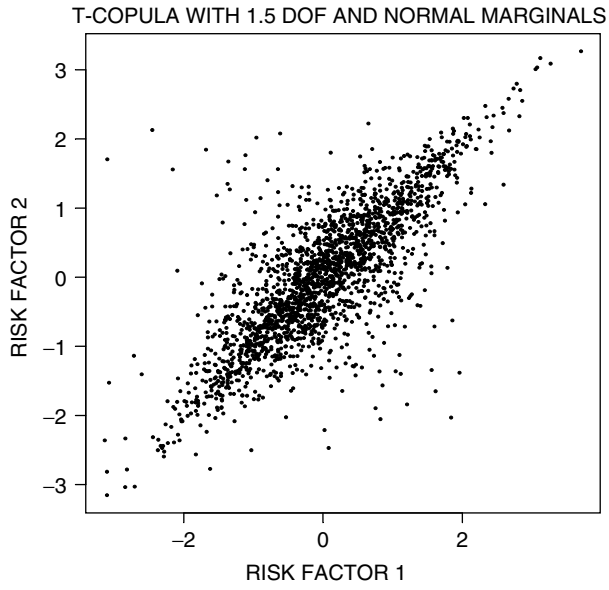


Figure 7

A sample of 2000 bivariate simulated risk factors generated by a t-copula with 1.5 degrees of freedom and Normal marginal distributions is displayed in Figure 7.

The example illustrates that these two risk factors are somewhat uncorrelated for small to moderately large returns, but are highly correlated for the infrequent occurrence of very large returns. This can be seen by noting that the density contours of points in the scatter plot are somewhat eliptical near the origin, but are nowhere close to eliptical for more extreme events. This situation is in contrast to a Gaussian linear dependency relationship where the density contours are expected to be eliptical.

Volatility modeling and Stable vs. Normal VaR

It is well known that risk factors returns exhibit volatility clustering, and that even after adjusting for such clustering the returns will still be non-normal and contain extreme values. There may also be some serial dependency effects to account for. In order to adequately model these collective behaviors we recommend using ARIMA models with an ARCH/GARCH "time-varying" volatility input, where the latter has Stable (non-Gaussian) innovations. This approach is more flexible and accurate than the commonly used simple EWMA (exponentially weighted moving average) volatility model, and provides accurate time-varying estimates of VaR and Expected Tail Loss (ETL) risk measures. See the next section for discussion of ETL vs. VaR that emphasizes the advantages of ETL. However, we stress that those who must use VaR to satisfy regulatory requirements will get much better results with Stable VaR than with Normal VaR, as the following example vividly shows.

Consider the following portfolio of Brady bonds:

- Brazil C 04/14.
- Brazil EIB 04/06.

- Venezuela DCB Floater 12/07.
- Samsung KRW Ord Shares.
- Thai Farmers Bank THB.

We have run Normal, Historical and Stable 99% (1% tail probability) VaR calculations for one-year of daily data from January 9, 2001 to January 9, 2002. We used a moving window with 250 historical observations for the Normal VaR model, 500 for the historical VaR model and 700 for the Stable VaR model. For each of these cases we used a GARCH(1,1) model for volatility clustering of the risk factors, with Stable innovations. We back-tested these VaR calculations by using the VaR values as one-step ahead predictors, and got the results shown in Figure 8.

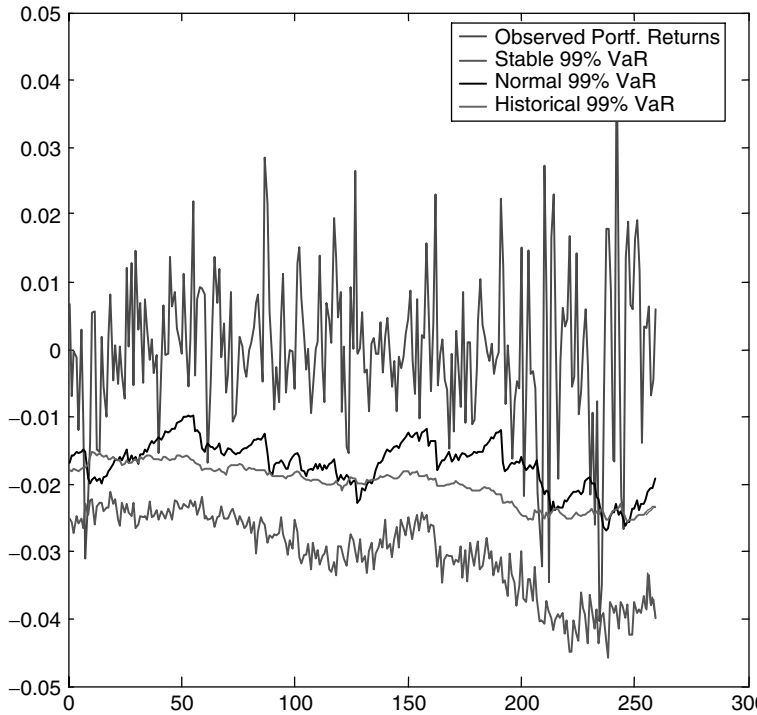


Figure 8

The figure shows: the returns of the Brady bond portfolio (top curve); the Normal + EWMA (à la RiskMetrics) VaR (curve with jumpy behavior, just below the returns); the Historical VaR (the smoother curve mostly below but sometimes crossing the Normal + EWMA VaR); the Stable + GARCH VaR (the bottom curve). The results with regard to exceedances of the 99% VaR, and keeping in mind Basel II guidelines, may be summarized as follows:

- Normal 99% VaR produced 12 exceedances.
- Historical 99% VaR produced 9 exceedances.
- Stable 99% VaR produced 1 exceedence.

Clearly Stable (+GARCH) 99% VaR produces much better results with regard to Basel II compliance. This comes at the price of higher initial capital reserves, but results in a much safer level of capital reserves and a very clean bill of health with regard to compliance. Note that some organizations may be fined and will have to increase their capital reserves by up to 33%, which at some times for some portfolios will result in larger capital reserves than when using the Stable VaR, this in addition to being viewed as having inadequate risk measures. This unfortunate situation is much less likely to happen to the organization using Stable VaR.

ETL is the next generation risk measure

Why not value-at-risk (VaR)?

There is no doubt that VaR's popularity is in large part due to its simplicity and its ease of calculation for 1 to 5% confidence levels. However, there is a price to be paid for the simplicity of VaR in the form of several limitations:

- VaR does not give any indication of the risk beyond the quantile!
- VaR portfolio optimization is a non-convex, non-smooth problem with multiple local minima that can result in portfolio composition discontinuities. Furthermore it requires complex calculation techniques such as integer programming.
- VaR is not sub-additive; i.e. the VaR of the aggregated portfolio can be larger than the sum of the VaR's of the sub-portfolios.
- Finally, and most importantly, VaR can be a very misleading risk indicator: examples exist where an investor, unintentionally or not, decreases portfolio VaR while simultaneously increasing the expected losses beyond the VaR, i.e., by increasing the "tail risk" of the portfolio (see the discussion of ETL below).

In addition to these intrinsic limitations, specific VaR implementations are fraught with further flaws:

- Historical VaR limits the range of the scenarios to data values that have actually been observed, while Normal Monte Carlo tends to seriously underestimate the probability of extreme returns. In either case, the probability functions beyond the sample range are either zero or excessively close to zero.
- Lacking the ability to accurately model extreme returns, practitioners are forced to use stress testing as a palliative to the limitations of traditional VaR models. In doing so, they use a large degree of subjectivity in the design of the stress test and in the selection of past data to use in making a risk assessment.
- The traditional modeling of risk factor dependences cannot account for intraday volatility patterns, long-term volatility patterns, or more importantly unusual extreme volatility. In stressed markets, the simple linear diversification assumptions fail, and atypical short-term concentration patterns that bind all the assets in a bearish spiral emerge.

Yamai and Yoshiba (2002) note in their concluding remarks:

The widespread use of VaR for risk management could lead to market instability. Basak and Shapiro (2001) show that when investors use VaR for their risk management, their

optimizing behavior may result in market positions that are subject to extreme loss because VaR provides misleading information regarding the distribution's tail.

ETL and Stable vs. Normal distributions

Expected Tail Loss (ETL) is simply the average (or expected value) loss for losses larger than VaR. ETL is also known as *Conditional Value-at-Risk* (CVaR), or *Expected Shortfall* (ES).

Usual (1 to 5%) Normal ETL is close to Normal VaR (See VaR by Jorion, 2001 p. 98):

- For CI = 5%, VaR = 1.645 and ETL = 2.062.
- For CI = 1%, VaR = 2.336 and ETL = 2.667.

By failing to capture kurtosis, Normal distributions underestimate ETL. The ubiquitous Normal assumption makes ETL difficult to interpret, in spite of ETL's remarkable properties (see below). Unlike Normal distributions, Stable distributions capture leptokurtic tails ("fat tails"). Unlike Normal ETL, Stable ETL provides reliable values. Further, when Stable distributions are used, ETL is generally substantially different from VaR.

Figure 9 shows the time series of daily returns for the stock OXM from January 2000 to December 2001. Observe the occurrences of extreme values.

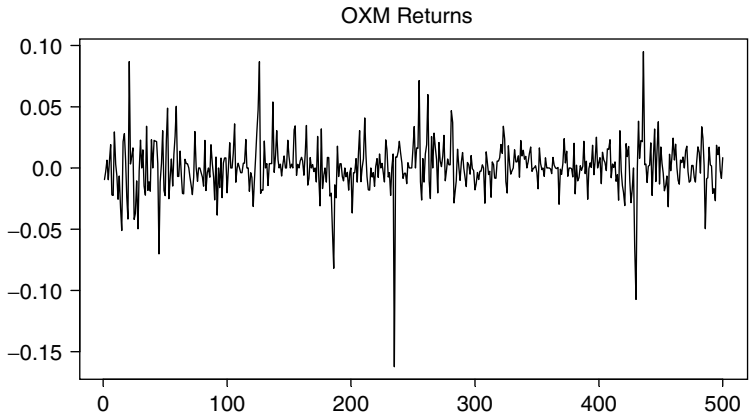


Figure 9

While this series also displays obvious volatility clustering, that deserves to be modeled as described in the next section, we shall ignore this aspect for the moment. Rather, here we provide a compelling example of the difference between ETL and VaR based on a well-fitting stable distribution, as compared with a poor fitting Normal distribution.

Figure 10 shows a histogram of the OXM returns with a Normal density fitted using the sample mean and sample standard deviation, and a Stable density fitted using maximum-likelihood estimates of the Stable distribution parameters. The Stable density is shown by the solid line and the normal density is shown by the dashed line. The former is obviously a better fit than the latter, when using the histogram of the data values as a reference. The estimated Stable tail thickness index is $\hat{\alpha} = 1.62$. The 1% VaR values for the Normal and Stable fitted densities

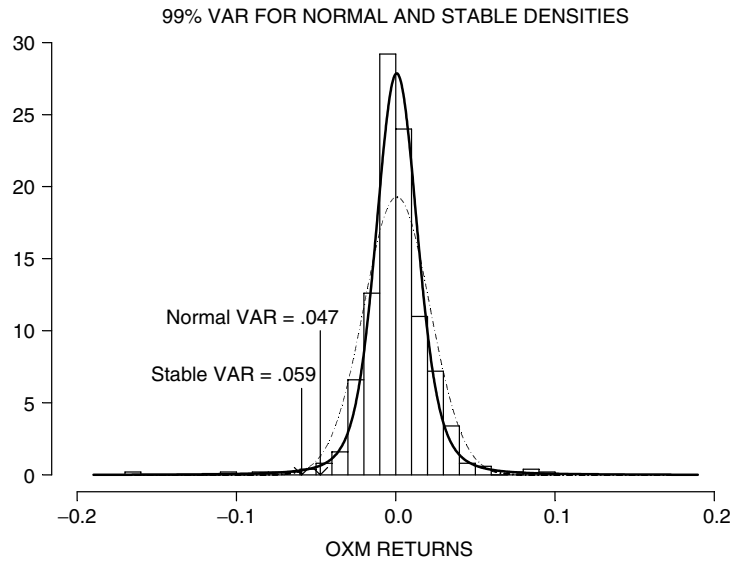


Figure 10

are 0.047 and 0.059 respectively, a ratio of 1.26 which reflects the heavier-tailed nature of the Stable fit.

Figure 11 displays the same histogram and fitted densities with 1% ETL values instead of the 1% VaR values. The 1% ETL values for the Normal and Stable fitted densities are 0.054 and 0.147, respectively, a ratio of a little over three-to-one. This larger ratio is due to the Stable density's heavy tail contribution to ETL relative to the Normal density fit.

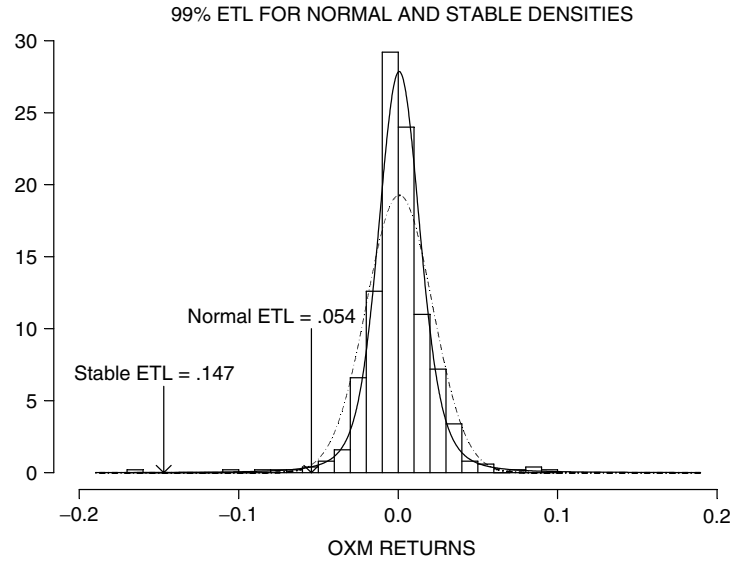


Figure 11

Unlike VaR, ETL has a number of attractive properties:

- ETL gives an informed view of losses beyond VaR.
- ETL is a convex, smooth function of portfolio weights, and is therefore attractive to optimize portfolios (see Uryasev and Rockafellar, 2000). This point is vividly illustrated in the subsection below on ETL and Portfolio Optimization.
- ETL is sub-additive and satisfies a complete set of coherent risk measure properties (see Artzner *et al*, 1999).
- ETL is a form of expected loss (i.e. a conditional expected loss) and is a very convenient form for use in scenario-based portfolio optimization. It is also quite a natural risk-adjustment to expected return (see STARR, or Stable Tail Adjusted Return Ratio).

The limitations of current Normal risk factor models and the absence of regulator blessing have held back the widespread use of ETL, in spite of its highly attractive properties.

For portfolio optimization, we recommend the use of Stable ETL, and limiting the use of Historical, Normal or Stable VaR to regulatory reporting purposes. At the same time, organizations should consider the advantages of Stable ETL for risk assessment purposes and non-regulatory reporting purposes.

The following quotation is relevant:

> Expected Tail Loss gives an indication of extreme losses, should they occur. Although it has not become a standard in the financial industry, expected tail loss is likely to play a major role, as it currently does in the insurance industry. (Embrechts *et al*, 1997).

Portfolio optimization and ETL vs. VaR

To the surprise of many, portfolio optimization with ETL turns out to be a smooth, convex problem with a unique solution (Rockafellar and Uryasev, 2000). These properties are in sharp contrast to the non-convex, rough VaR optimization problem.

The contrast between VAR and ETL portfolio optimization surfaces is illustrated in Figure 12(a) and (b) for a simple two-asset portfolio. The horizontal axes show one of the portfolio weights (from 0 to 100%) and the vertical axes display portfolio VAR and ETL respectively. The data consist of 200 simulated uncorrelated returns.

The VAR objective function is quite rough with respect to varying the portfolio weight(s), while that of the ETL objective function is smooth and convex. One can see that optimizing with ETL is a much more tractable problem than optimizing with VaR.

Rockafellar and Uryasev (2000), show that the ETL Optimal Portfolio (ETLOP) weight vector can be obtained based on historical (or scenario) returns data by minimizing a relatively simple convex function (Rockafellar and Uryasev used the term CVaR, whereas we use the less confusing synonym ETL). Assuming p assets with single period returns $\mathbf{r}_i = (r_{i1}, r_{i2}, \cdots, r_{ip})$, for period i, and a portfolio weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_p)$, the function to be minimized is:

$$F(\mathbf{w}, \gamma) = \gamma + \frac{1}{\varepsilon \cdot n} \sum_{i=1}^{n} [\mathbf{w}'\mathbf{r}_i - \gamma]^+$$

where $[x]^+$ denotes the positive part of x. This function is to be minimized jointly with respect to $\mathbf{w}$ and γ, where ε is the tail probability for which the expected tail loss is computed. Typically

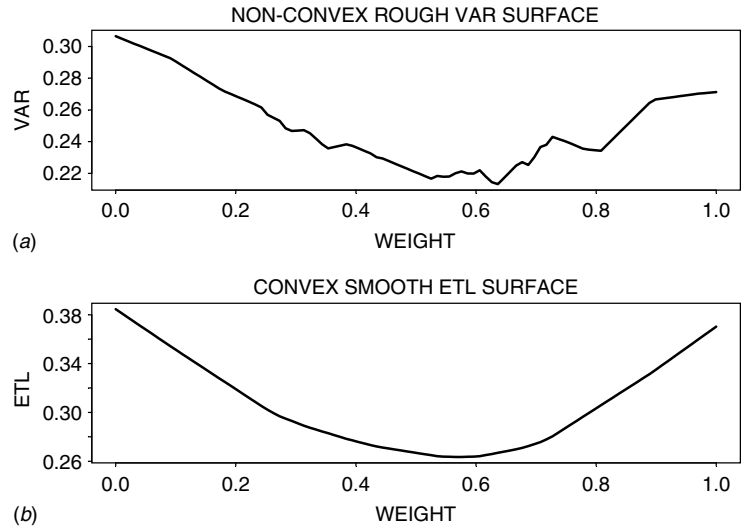


Figure 12

$\varepsilon = .05$ or .01, but larger values may be useful, as we discuss in the section on Choice of tail permeability. The authors further show that this optimization problem can be cast as a LP (linear programming) problem, solvable using any high-quality LP software.

Cognity™ combines this approach along with multivariate Stable scenario generation. The stable scenarios provide accurate and well-behaved estimates of ETL for the optimization problem.

Stable ETL leads to higher risk adjusted returns

ETLOP (Expected Tail Loss Optimal Portfolio) techniques, combined with multivariate Stable distribution modeling, can lead to significant improvements in risk adjusted return as compared to not only Normal VAROP methods but also compared to Normal ETL optimization. In practice, a VAR Optimal Portfolio (VAROP) is difficult to compute accurately with more than two or three assets.

Figures 13 and 14 are supplied to illustrate the claim that Stable ETLOP produces consistently better risk-adjusted returns. These figures show the risk adjusted return MU/VAR (mean return divided by VAR) and MU/ETL (mean return divided by ETL) for 1% VAROP and ETLOP, and using a multi-period fixed-mix optimization in all cases.

In this simple example, the portfolio to be optimized consists of two assets, cash and the S&P 500. The example is based on monthly data from February 1965 to December 1999. Since we assume full investment, the VAROP depends only on a single portfolio weight and the optimal weight(s) is found by a simple grid search on the interval 0 to 1. The use of a grid search technique, overcomes the problems with non-convex and non-smooth VAR optimization. In this example the optimizer is maximizing $MU - c \cdot VAR$ and $MU - c \cdot ETL$, where c is the risk aversion (parameter), and with VAR or ETL as the penalty function.

Figure 13 shows that even using VAROP, one gets a significant relative gain in risk-adjusted return using Stable scenarios when compared to Normal scenarios, and with the relative gain increasing with increasing risk aversion. The reason for the latter behavior is that with Stable

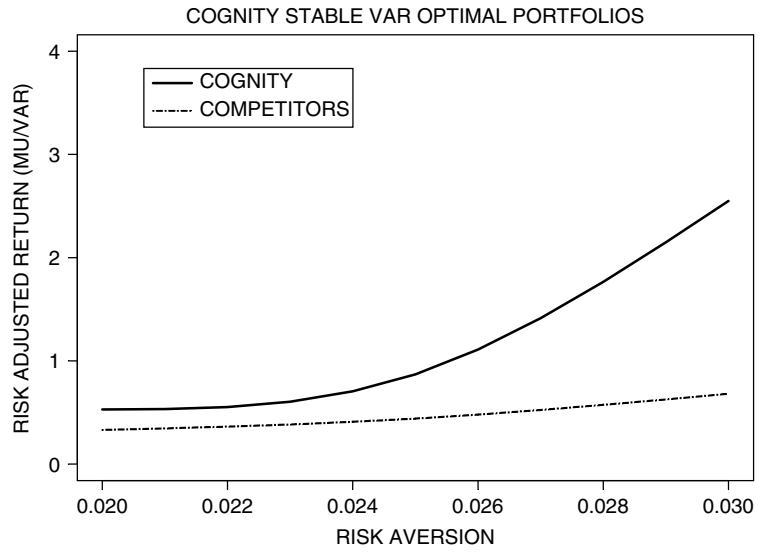


Figure 13

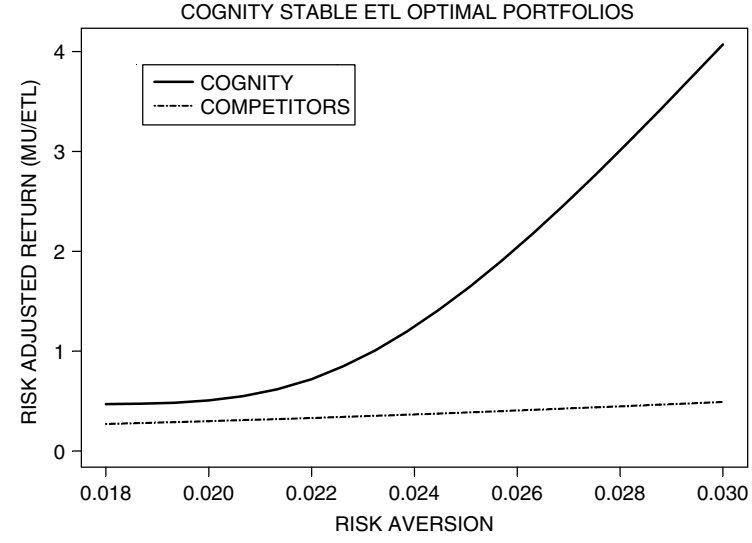


Figure 14

distributions the optimization pays more attention to the S&P returns distribution tails, and allocates less investment to the S&P under stable distributions than under Normal distributions as risk aversion increases.

Figure 14 below for the risk-adjusted return for the ETLOP has the same vertical axis range as the previous plot for VAROP. The figure below shows that the use of ETL results in much greater gain under the Stable distribution relative to the Normal than in the case of VAROP.

At every level of risk aversion, the investment in the S&P 500 is even less in ETLOP than in the case of the VAROP. This behavior is to be expected because the ETL approach pays attention to the losses beyond VAR (the expected value of the extreme loss), and which in the Stable case are much greater than in the Normal case.

The ϕ_α paradigm and ϕ_α optimal portfolios

The phi-alpha (ϕ_α) paradigm

Our approach uses multi-dimensional risk factor models based on multivariate Stable process models for risk management and constructing optimal portfolios, and stresses the use of Stable ETL as the risk measure of choice. These Stable distribution models incorporate generalized dependence structure with copulas, and include time varying volatilities based on GARCH models with Stable innovations. Collectively these modeling foundations form the basis of a new, powerful overall basis for investment decisions that we call the *Phi-Alpha* (ϕ_α) *Paradigm*.

Currently the ϕ_α Paradigm has the following basic components: ϕ_α scenario engines, ϕ_α integrated market risk and credit risk (with integrated operational risk under development), ϕ_α optimal portfolios and efficient frontiers, and ϕ_α derivative pricing. Going forward, additional classes of ϕ_α investment decision models will be developed, such as ϕ_α betas, ϕ_α factor models, and ϕ_α asset liability models. The rich structure of these models will encompass the heavy-tailed distributions of the asset returns, stochastic trends, heteroscedasticity, short- and long-range dependence, and more. We use the term "ϕ_α model" to describe any such model in order to keep in mind the importance of the Stable tail-thickness parameter α in financial investment decisions.

It is essential to keep in mind the following ϕ_α fundamental principles concerning risk factors:

(P1) Risk factor returns have Stable distributions where each risk factor i typically has a different Stable tail-index α_i.

(P2) Risk factor returns are associated through models that describe the dependence between the individual factors more accurately than classical correlations. Often these will be copula models.

(P3) Risk factor modeling typically includes a ϕ_α-econometric model in the form of multivariate ARIMA-GARCH processes with residuals driven by fractional Stable innovations. The ϕ_α econometric model captures clustering and long-range dependence of the volatility.

Phi-alpha (ϕ_α) optimal portfolios

A Phi-Alpha (ϕ_α) optimal portfolio is one that minimizes portfolio expected tail loss (ETL) subject to a constraint of achieving expected portfolio returns at least as large as an investor defined level, where both quantities are evaluated in ϕ_α. Alternatively, a ϕ_α optimal portfolio solves the dual problem of maximizing portfolio expected return subject to a constraint that portfolio expected tail loss (ETL) is not greater than an investor defined level, where again both quantities are evaluated in ϕ_α. In order to define the above ETL precisely we use the following quantities:

R_p:	the random return of portfolio p
SER_p:	the ϕ_α expected return of portfolio p

$L_p = -R_p + SER_p$:	the loss of portfolio p relative to its ϕ_α expected return
ε:	a tail probability of the ϕ_α distribution L_p
$SVaR_p(\varepsilon)$:	the ϕ_α Value-at-Risk for portfolio p

The latter is defined by the equation:

$$\mathbf{Pr}[L_p > SVaR_p(\varepsilon)] = \varepsilon$$

where the probability is calculated in ϕ_α, that is $SVaR_p(\varepsilon)$ is the ε-quantile of the ϕ_α distribution of L_p. In the value-at-risk literature $(1 - \varepsilon) \times 100\%$ is called the confidence level. Here we prefer to use the simpler, unambiguous term *tail probability*. Now we define the ϕ_α expected tail loss as:

$$SETL_p(\varepsilon) = E[L_p | L_p > SVaR_p(\varepsilon)]$$

where the conditional expectation is also computed in ϕ_α. Note that the ϕ_α expected value of L_p is zero. We use the "S" in SER_p, $SVaR_p(\varepsilon)$ and $SETL_p(\varepsilon)$ as a reminder that Stable distributions are a key aspect of the ϕ_α (but not the only aspect!).

Proponents of (Gaussian) VaR typically use tail probabilities of .01 or .05. When using $SETL_p(\varepsilon)$ risk managers may wish to use other tail probabilities such as 0.1, 0.15, 0.20, 0.25, or 0.5. We note that use of different tail probabilities is similar in spirit to using different utility functions. We return to discuss this point further below.

The following assumptions are in force for the ϕ_α investor:

(A1) The universe of assets is Q (the set of mandate admissible portfolios).
(A2) The investor may borrow or deposit at the risk-free rate r_f without restriction.
(A3) The portfolio is optimized under a set of asset allocation constraints λ.
(A4) The investor seeks an expected return of at least μ (alternatively an ETL risk of at most η).

To simplify the notation we shall let A3 be implicit in the following discussion. At times we shall also suppress the ε when its value is taken as fixed and understood.

The ϕ_α investor's optimal portfolio is:

$$\omega_\alpha(\mu|\varepsilon) = \arg\min_{q \in Q} SETL_q(\varepsilon)$$

subject to:

$$SER_q \geq \mu.$$

In other words the ϕ_α optimum portfolio ω_α minimizes the ϕ_α expected tail loss (SETL) among all portfolios with ϕ_α mean return (SER) at least μ, for fixed tail probability ε asset allocation constraints λ. Alternatively, the ϕ_α optimum portfolio ω_α solves the dual problem:

$$\omega_\alpha(\eta|\varepsilon) = \arg\max_{q \in Q} SER_q$$

subject to:

$$SETL_q(\varepsilon) \leq \eta.$$

The ϕ_α *efficient frontier* is given by $\omega_\alpha(\mu|\varepsilon)$ as a function of μ for fixed ε, as indicated in Figure 15. If the portfolio includes cash account with risk free rate r_f, then ϕ_α *efficient frontier* will be the ϕ_α capital market line (CML_α) that connects the risk-free rate on the vertical axis with the ϕ_α tangency portfolio (T_α), as indicated in the figure.

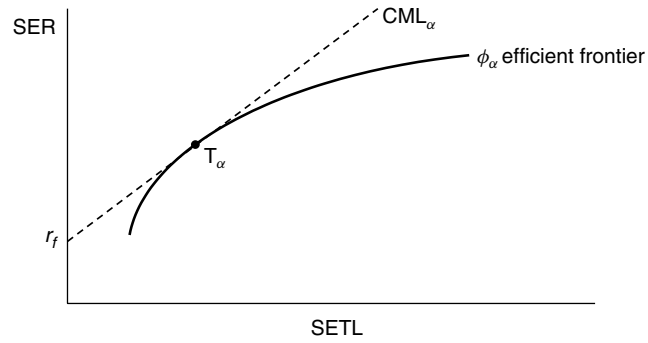


Figure 15

We now have a ϕ_α separation principal analogous to the classical separation principal: The tangency portfolio T_α can be computed without reference to the risk-return preferences of any investor. Then an investor chooses a portfolio along the ϕ_α capital market line CML_α according to his/her risk-return preference.

We note that it is convenient to think of ω_α in two alternative ways: (1) the ϕ_α optimal portfolio, or (2) the vector of ϕ_α optimal portfolio weights.

Keep in mind that in practice when a finite sample of returns one ends up with a ϕ_α efficient frontier, tangency portfolio and capital market line that are estimates of true values for these quantities. Under regularity conditions these estimates will converge to true values as the sample size increases to infinity.

Markowitz mean-variance portfolios are sub-optimal

While the ϕ_α investor has a ϕ_α optimal portfolio described above, let's assume that the mean-variance investor is not aware of the ϕ_α paradigm and constructs a mean-variance optimal portfolio. We assume that the mean-variance investor operates under the same assumptions A1–A4 as the ϕ_α investor. Let ER_q be the expected return and σ_q the standard deviation of the returns of a portfolio q. The mean-variance investor's optimal portfolio is:

$$\omega_2(\mu) = \arg\min_{q \in Q} \sigma_q$$

subject to:

$$ER_q \geq \mu.$$

The mean-variance optimal portfolio can also be constructed by solving the obvious dual optimization problem of maximizing the expected return for a constrained risk level. One knows that, in the mean-variance paradigm, contrary to the ϕ_α paradigm, the mean-variance optimal portfolio is independent of any *ETL* tail probability specification.

The subscript 2 is used in ω_2 as a reminder that when $\alpha = 2$ you have the limiting Gaussian distribution member of the Stable distribution family, and in that case the mean-variance portfolio is optimal. Alternatively you can think of the subscript 2 as a reminder that the mean-variance optimal portfolio is a second-order optimal portfolio, i.e., an optimal portfolio based on only first and second moments.

Note that the mean-variance investor ends up with a different portfolio, i.e., a different set of portfolio weights with different risk versus return characteristics, than the ϕ_α investor.

The performance of the mean-variance portfolio, like that of the ϕ_α portfolio, is evaluated under the ϕ_α distributional model, i.e., its expected return and expected tail loss are computed under the ϕ_α distributional model. If in fact the distribution of the returns were exactly multivariate Gaussian (which they never are) then the ϕ_α investor and the mean-variance investor would end up with one and the same optimal portfolio. However, when the returns are non-Gaussian ϕ_α returns, the mean-variance portfolio is sub-optimal. This is because the ϕ_α investor constructs his/her optimal portfolio using the ϕ_α distribution model, whereas the mean-variance investor does not. Thus the mean-variance investor's frontier lies below and to the right of the ϕ_α efficient frontier, as shown in Figure 16, along with the mean-variance tangency portfolio T_2 and mean-variance capital market line CML_2.

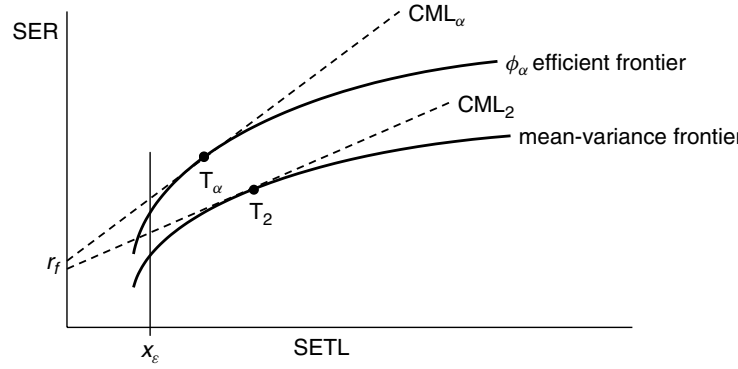


Figure 16

As an example of the performance improvement achievable with the ϕ_α optimal portfolio approach, we computed the ϕ_α efficient frontier and the mean-variance frontier for a portfolio of 47 micro-cap stocks with the smallest alphas from the random selection of 182 micro-caps described above. The results are displayed in Figure 17. The results are based on 3,000 scenarios from the fitted ϕ_α multivariate distribution model based on two years of daily data during years 2000 and 2001. We note that, as is generally the case, each of the 47 stock returns has its own estimate Stable tail index $\hat{\alpha}_i, i = 1, 2, \ldots, 47$.

Here we have plotted values of $TailRisk = \varepsilon \cdot SETL(\varepsilon)$, for $\varepsilon = .01$, as a natural decision theoretic risk measure, rather than $SETL(\varepsilon)$ itself. We note that over a large range of tail risk the ϕ_α efficient frontier dominates the mean-variance frontier by 14–20 bp daily!

We note that the 47 micro-caps with the smallest alphas used for this example have quite heavy tails as indicated by the boxplot of their estimated alphas, shown in Figure 18.

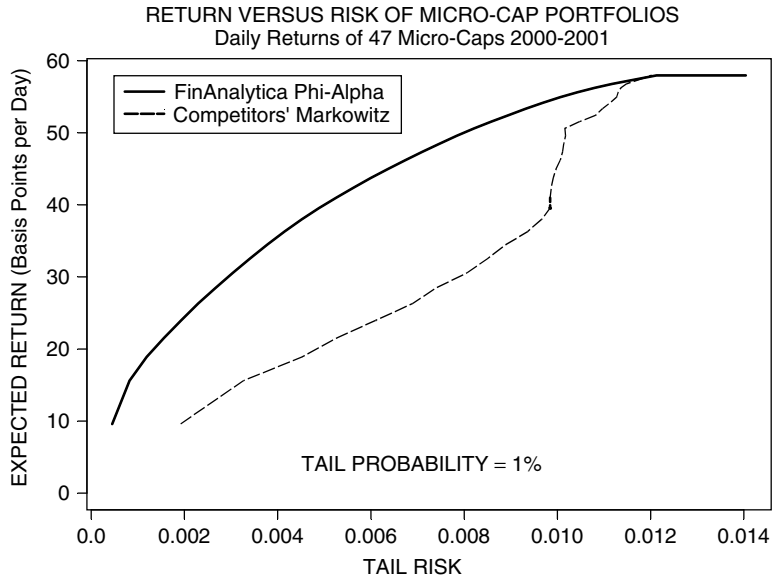


Figure 17

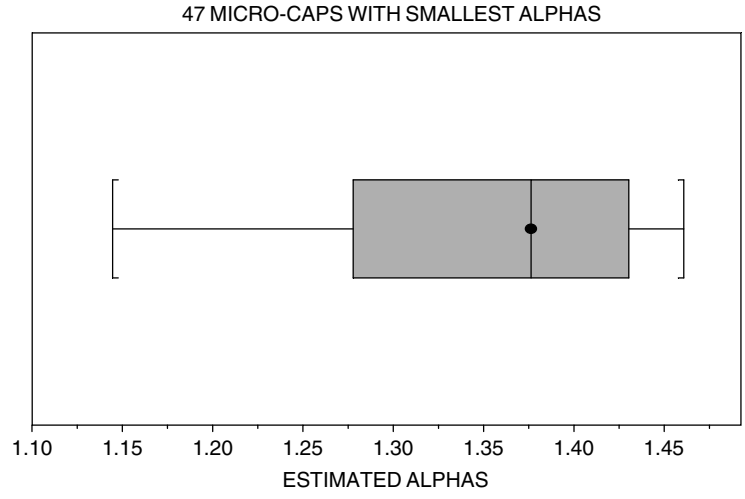


Figure 18

Here the median of the estimated alphas is 1.38, while the upper and lower quartiles are 1.43 and 1.28 respectively. Evidently there is a fair amount of information in the non-Gaussian tails of such micro-caps that can be exploited by the ϕ_α approach.

We also note that the gap between the ϕ_α efficient frontier and the Markowitz mean-variance frontier will decrease as the Stable tail index values α_i get closer to 2, i.e., as the multivariate distribution gets closer to a multivariate Normal distribution. This will be the case, for example, when moving from micro-cap and small-cap stocks to mid-cap and large cap stocks.

The Excess Profit Pi-Alpha

Let's first construct the traditional Markowitz mean-variance optimal portfolio for a selected risk level, as measured by the variance. Let $x_\varepsilon = SETL(\varepsilon)$ be the value of the Stable expected tail loss of that mean-variance portfolio. Let's then construct the Phi-Alpha optimal portfolio with the identical Stable expected tail loss value x_ε (assuming it exist). By construction, both portfolios have the same Phi-Alpha risk level. Also, let $SER_{\omega_\alpha}(x_\varepsilon)$ be the ϕ_α expected return of that ϕ_α optimal portfolio, and let $SER_{\omega_\alpha}(x_\varepsilon)$ be the ϕ_α expected return of that Markowitz mean-variance optimal portfolio. We define the excess profit (Pi-Alpha) of the ϕ_α portfolio in basis points as:

$$\pi_\alpha = \pi_\alpha(\varepsilon, x_\varepsilon) = [SER_{\omega_\alpha}(x_\varepsilon) - SER_{\omega_2}(x_\varepsilon)] * 100.$$

In the case of portfolios with no cash component the values of Stable expected returns above are the ordinate values where the vertical line at x_ε intersects the two frontiers (see the figure above). When cash is present these expected returns are the values where the vertical line at x_ε intersects the capital market lines CML_α and CML_2.

Based on our current studies, yearly π_α in the range of tens to possibly one or two hundred basis points are achievable, depending upon the portfolio under management.

New ratios: from Sharpe to STARI and STARR

The *Sharpe Ratio* for a given portfolio p is defined as follows:

$$SR_p = \frac{ER_p - r_f}{\sigma_p} \tag{2}$$

where ER_p is the portfolio expected return, σ_p is the portfolio return standard deviation as a measure of portfolio risk, and r_f is the risk-free rate. While the Sharpe ratio is the single most widely used portfolio performance measure, it has several disadvantages due to its use of the standard deviation as risk measure:

- σ_p is a symmetric measure that does not focus on downside risk.
- σ_p is not a coherent measure of risk (see Artzner *et al*, 1999).
- The classical estimate of σ_p is a highly unstable measure of risk when the portfolio has a heavy-tailed distribution.
- σ_p and has infinite value for non-Gaussian Stable distributions.

Stable Tail Adjusted Return Indicator. As an alternative performance measure that does not suffer these disadvantages, we propose the *Stable Tail Adjusted Return Indicator* (*STARI*) defined as:

$$STARI_p(\varepsilon) = \frac{SER_p - r_f}{SETL_p(\varepsilon)}. \tag{3}$$

Referring to Figure 16, one sees that the overall maximum *STARI* is attained by the ϕ_α optimal portfolio, and is the slope of the ϕ_α capital market line CML_α. On the other hand, the maximum STARI of Markowitz's mean-variance optimal portfolio is equal to the slope of the mean-variance capital market line CML_2. CML_2 is always dominated by CML_α, and CML_2 is equal to CML_α only when the returns distribution is multivariate normal in which case $\alpha_i = 2$ for all risk factors i.

We conclude that the risk adjusted return of the ϕ_α optimal portfolio ω_α is generally superior to the risk adjusted return of Markowitz's mean variance optimal portfolio ω_2. The ϕ_α paradigm results in improved investment performance.

Stable Tail Adjusted Return Ratio. While *STARI* provides a natural measure of return per unit risk, the numerical values obtained are not in a range familiar to users of the Sharpe ratio, even in the case where the returns are multivariate normal. However, it is easy to rescale *STARI* so that when the returns are normally distributed the rescaled STARI is equal to the Sharpe ratio. We use the term *Stable Tail Adjusted Return Ratio* (*STARR*) for this rescaled ratio, and its formula is:

$$STARR_p(\varepsilon) = \frac{SER_p - r_f}{SETL_p(\varepsilon)/NETL^{0,1}(\varepsilon)} \quad (4)$$

where $NETL^{0,1}(\varepsilon)$ is the ETL for a *standard normal distribution* at tail probability ε.

It is easy to check that $STARR_p(\varepsilon)$ coincides with the Sharpe ratio SR_p when the portfolio has a normal distribution. First one easily verifies that:

$$NETL^{0,1}(\varepsilon) = \frac{1}{\varepsilon\sqrt{2\pi}} e^{-\frac{(N^{0,1}VaR(\varepsilon))^2}{2}}$$

where $N^{0,1}VaR(\varepsilon) = -\Phi^{-1}(\varepsilon)$ is the VaR of a *standard normal distribution* at tail probability ε and Φ is the standard normal cumulative distribution function, i.e., $N^{0,1}VaR(\varepsilon)$ is the ε-quantile of the *standard normal distribution*. Now suppose that the loss L_p of portfolio p has a normal distribution standard deviation σ_p, recall that the loss has zero expected value, and call the corresponding expected tail loss $NETL_p(\varepsilon)$. Then:

1) $NETL_p(\varepsilon) = \sigma_p \; NETL^{0,1}(\varepsilon)$ (easy to verify)
2) $SETL_p(\varepsilon) = NETL_p(\varepsilon)$
3) $SER_p = ER_p$ (in any event).

Using *NTARR* to denote the resulting *STARR*, we have:

$$NTARR_p = SR_p$$

which is now independent of ε.

The choice of tail probability

We mentioned earlier that when using $SETL_p(\varepsilon)$ rather than $VaR_p(\varepsilon)$, risk managers and portfolio optimizers may wish to use other values of ε than the conventional VaR values of 0.01 or 0.05, for example values such as 0.1, 0.15, 0.2, 0.25 and 0.5 may be of interest. The choice of a particular ε amounts to a choice of particular risk measure in the SETL family of measures, and such a choice is akin to the choice of a utility function. The tail probability parameter ε is at the asset managers disposal to choose according to his/her asset management and risk control objectives.

Note choosing a tail probability ε is not the same as choosing a risk aversion parameter. Maximizing:

$$SER_p - c \cdot SETL_p(\varepsilon)$$

for various choices of risk aversion parameter c for a fixed value of ε merely corresponds to choosing different points along the ϕ_α efficient frontier. On the other hand changing ε results in different shapes and locations of the ϕ_α efficient frontier, and corresponding different excess profits π_α relative to a mean-variance portfolio.

It is intuitively clear that increasing ε will decrease the degree to which a ϕ_α optimal portfolio depends on extreme tail losses. Where $\varepsilon = 0.5$, which may well be of interest to some managers since it uses the average loss below zero of L_p as its penalty function, small to moderate losses are mixed in with extreme losses in determining the optimal portfolio. Our studies to date show that some of the excess profit π_α advantage relative to Markowitz mean-variance optimal portfolios will be given up as ε increases, and that not surprisingly, this effect is most noticeable for portfolios with smaller Stable tail index values (i.e. fatter tails).

In summary: the smaller the tail probability ε, i.e. the more concentrated in the tail that the manager calculates risk, the higher (in general) the expected excess mean return π_α of the ϕ_α optimal portfolio over the mean-variance optimal portfolio.

It will be interesting to see what values of ε will be used by fund managers of various types and styles in the future.

The Cognity implementation of the ϕ_α paradigm

The ϕ_α Paradigm described in this section has been implemented in the ***Cognity***^{TM} Risk Management and Portfolio Optimization product. This product contains separate Market Risk, Credit Risk and Portfolio Optimization modules, with integrated Market and Credit Risk, and implemented in Java based architecture to support Web delivery.

Acknowledgements

The authors gratefully acknowledge the extensive help provided by Borjana Racheva-Jotova, Stoyan Stoyanov and Stephen Elston in the preparation of this paper.

REFERENCES

■ Artzner, P., Delbean, F., Eber, J. M. and Heath, D. (1999) 'Coherent measures of risk', *Mathematical Finance* 9, 203–228.

■ Bradley, B. O. and Taqqu, M. S. (2003) Financial Risk and Heavy Tails. In *Handbook of Heavy Tailed Distributions in Finance*, edited by S. T. Rachev. Elsevier/North-Holland: Amsterdam.

■ Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events in Insurance and Finance*. Springer, Berlin.

■ Embrechts, P., Lindskog, and McNeil, A. (2003) Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance*, Rachev, S. T. (ed.). Handbooks in Finance, Vol. 1, Elsevier Science, Amsterdam, 329, 328, 384.

■ Fama, E. (1963) Mandelbrot and the stable paretian hypothesis. *Journal of Business* 36, 420–429.

■ Mandelbrot, B. B. (1963) The variation in certain speculative prices. *Journal of Business* 36, 394–419.

■ Rachev, S. and Mittnik, S. (2000) *Stable Paretian Models in Finance*. Wiley, Chichester.

■ Rachev, S., Schwartz, E. and Khindanova, I. (2003) Stable modeling of market and credit value at risk. In *Handbook of Heavy Tailed Distributions in Finance*, Rachev, S. T. (ed.). Handbooks in Finance, Vol. 1, Elsevier Science, Amsterdam, 249–328.
■ Rockafellar, R. T. and Uryasev, S. (2000) Optimization of conditional value-at-risk. *Journal of Risk* 3, 21–41.

18
Managing Smile Risk

Patrick S. Hagan,* Deep Kumar,† Andrew S. Lesniewski* and Diana E. Woodward§

Wilmott magazine, September 2002

European options are often priced and hedged using Black's model, or, equivalently, the Black–Scholes model. In Black's model there is a one-to-one relation between the price of a European option and the volatility parameter σ_B. Consequently, option prices are often quoted by stating the *implied volatility* σ_B, the unique value of the volatility which yields the option's dollar price when used in Black's model. In theory, the volatility σ_B in Black's model is a constant. In practice, options with different strikes K require different volatilities σ_B to match their market prices. See Figure 1. Handling these market *skews* and *smiles* correctly is critical to fixed income and foreign exchange desks, since these desks usually have large exposures across a wide range of strikes. Yet the inherent contradiction of using different volatilities for different options makes it difficult to successfully manage these risks using Black's model.

The development of *local volatility models* by Dupire (1994, 1997) and Derman and Kani (1994, 1998) was a major advance in handling smiles and skews. Local volatility models are self-consistent, arbitrage-free, and can be calibrated to precisely match observed market smiles and skews. Currently these models are the most popular way of managing smile and skew risk. However, as we shall discover in the next section, the *dynamic* behavior of smiles and skews predicted by local vol models is exactly *opposite* the behavior observed in the marketplace: when the price of the underlying asset *decreases*, local vol models predict that the smile shifts to *higher* prices; when the price *increases*, these models predict that the smile shifts to *lower* prices. In reality, asset prices and market smiles move in the *same* direction. This contradiction between the model and the marketplace tends to de-stabilize the delta and vega hedges derived from local volatility models, and often these hedges perform worse than the naive Black–Scholes' hedges.

Contact addresses: *Quantitative Research & Development, Bloomberg LP, 499 Park Avenue, New York, NY 10022, USA.
E-mail: phagan1@bloomberg.net
*Ellington Management Group, 53 Forest Av., Old Greenwich, CT 06870, USA.
†AVM Ltd, 250 Australian Av., Suite 600, West Palm Beach, FL 33401, USA.
§Societe Generale, 1221 Avenue of the Americas, New York, NY 10020, USA.

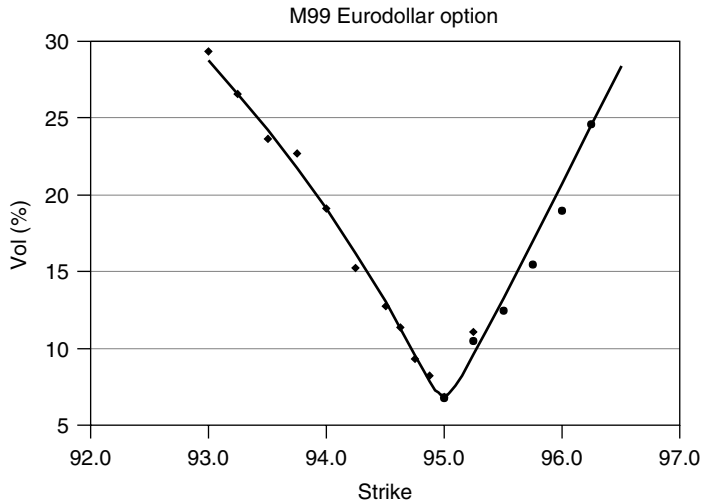


Figure 1: Implied volatility for the June 99 Eurodollar options. Shown are close-of-day values along with the volatilities predicted by the SABR model. Data taken from Bloomberg information services, 23 March 1999

To resolve this problem, we derive the *SABR* model, a stochastic volatility model in which the asset price and volatility are correlated. Singular perturbation techniques are used to obtain the prices of European options under the SABR model, and from these prices we obtain a closed-form algebraic formula for the implied volatility as a function of today's forward price f and the strike K. This closed-form formula for the implied volatility allows the market price and the market risks, including *vanna* and *volga* risks, to be obtained immediately from Black's formula. It also provides good, and sometimes spectacular, fits to the implied volatility curves observed in the marketplace (see Figure 1). More importantly, the formula shows that the SABR model captures the correct dynamics of the smile, and thus yields stable hedges.

Reprise

Consider a European call option on an asset $\mathcal{A}$ with *exercise date* t_{ex}, *settlement date* t_{set}, and *strike* K. If the holder exercises the option on t_{ex}, then on the settlement date t_{set} he receives the underlying asset $\mathcal{A}$ and pays the strike K. To derive the value of the option, define $\hat{F}(t)$ to be the forward price of the asset for a forward contract that matures on the settlement date t_{set}, and define $f = \hat{F}(0)$ to be today's forward price. Also let $D(t)$ be the *discount factor* for date t; that is, let $D(t)$ be the value today of \$1 to be delivered on date t. Martingale pricing theory (Harrison and Kreps, 1979; Harrison and Pliska, 1981; Karatzas *et al*, 1991; Steele, 2001) asserts that under the "usual conditions," there is a measure, known as the forward measure, under which the value of a European option can be written as the expected value of the payoff. The value of a call options is:

$$V_{call} = D(t_{set})E\left\{[\hat{F}(t_{ex}) - K]^{+}|\Im_0\right\} \tag{1a}$$

and the value of the corresponding European put is:

$$\begin{aligned} V_{put} &= D(t_{set})E\{[K - \hat{F}(t_{ex})]^+ | \mathfrak{F}_0\} \\ &\equiv V_{call} + D(t_{set})[K - f] \end{aligned} \tag{1b}$$

Here the expectation E is over the forward measure, and "$|\mathfrak{F}_0$" can be interpreted as "given all information available at $t = 0$". Martingale pricing theory also shows that the forward price $\hat{F}(t)$ is a Martingale under this measure, so the Martingale representation theorem shows that $\hat{F}(t)$ obeys:

$$d\hat{F} = C(t, *)\, dW, \qquad \hat{F}(0) = f \tag{1c}$$

for some coefficient $C(t, *)$, where dW is Brownian motion in this measure. The coefficient $C(t, *)$ may be deterministic or random, and may depend on any information that can be resolved by time t. This is as far as the fundamental theory of arbitrage free pricing goes. In particular, one cannot determine the coefficient $C(t, *)$ on purely theoretical grounds. Instead one must postulate a mathematical *model* for $C(t, *)$.

European swaptions fit within an identical framework. Consider a European swaption with exercise date t_{ex} and fixed rate (strike) R_{fix}. Let $\hat{R}_s(t)$ be the swaption's forward swap rate as seen at date t, and let $R_0 = \hat{R}_s(0)$ be the forward swap rate as seen today. Jamshidean (1997) shows that one can choose a measure in which the value of a payer swaption is:

$$V_{pay} = L_0 E\left\{[\hat{R}_s(t_{ex}) - R_{fix}]^+ | \mathcal{F}_0\right\} \tag{2a}$$

and the value of a receiver swaption is:

$$\begin{aligned} V_{rec} &= L_0 E\{[R_{fix} - \hat{R}_s(t_{ex})]^+ | \mathcal{F}_0\} \\ &\equiv V_{pay} + L_0[R_{fix} - R_0] \end{aligned} \tag{2b}$$

Here the *level* L_0 is today's value of the annuity, which is a known quantity, and E is the expectation over the *level measure* of Jamshidean (1997). In this article it is also shown that the forward swap rate, $\hat{R}_s(t)$, is a Martingale in this measure, so once again we have:

$$d\hat{R}_s = C(t, *)\, dW, \qquad \hat{R}_s(0) = R_0 \tag{2c}$$

where dW is Brownian motion. As before, the coefficient $C(t, *)$ may be deterministic or random, and cannot be determined from fundamental theory. Apart from notation, this is identical to the framework provided by equations (1a–1c) for European calls and puts. Caplets and floorlets can also be included in this picture, since they are just one period payer and receiver swaptions. For the remainder of the paper, we adopt the notation of (1a–1c) for general European options.

Black's model and implied volatilities

To go any further requires postulating a model for the coefficient $C(t, *)$. Black (1976) postulated that the coefficient $C(t, *)$ is $\sigma_B \hat{F}(t)$, where the *volatilty* σ_B is a constant. The forward

price $\hat{F}(t)$ is then geometric Brownian motion:

$$d\hat{F} = \sigma_B \hat{F}(t)\, dW, \qquad \hat{F}(0) = f \tag{3}$$

Evaluating the expected values in (2.1a, 2.1b) under this model then yields Black's formula:

$$V_{call} = D(t_{set})\{f\mathcal{N}(d_1) - K\mathcal{N}(d_2)\} \tag{4a}$$

$$V_{put} = V_{call} + D(t_{set})[K - f] \tag{4b}$$

where

$$d_{1,2} = \frac{\log f/K \pm \frac{1}{2}\sigma_B^2 t_{ex}}{\sigma_B \sqrt{t_{ex}}} \tag{4c}$$

for the price of European calls and puts, as is well known (Black, 1976; Hull, 1997; Wilmott, 2000).

All parameters in Black's formula are easily observed, except for the volatility σ_B. An option's *implied volatility* is the value of σ_B that needs to be used in Black's formula so that this formula matches the market price of the option. Since the call (and put) prices in (4a–4c) are increasing functions of σ_B, the volatility σ_B implied by the market price of an option is unique. Indeed, in many markets it is standard practice to quote prices in terms of the implied volatility σ_B; the option's dollar price is then recovered by substituting the agreed upon σ_B into Black's formula.

The derivation of Black's formula presumes that the volatility σ_B is a constant for each underlying asset $\mathcal{A}$. However, the implied volatility needed to match market prices nearly always varies with both the strike K and the time-to-exercise t_{ex} (see Figure 2). Changing the volatility σ_B means that a *different* model is being used for the underlying asset for each K and t_{ex}. This causes several problems managing large books of options.

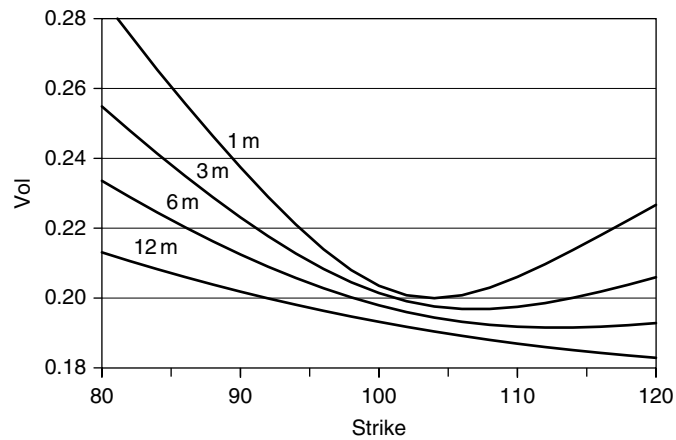


Figure 2: Implied volatility $\sigma_B(K)$ as a function of the strike K for 1 month, 3 month, 6 month, and 12 month European options on an asset with forward price 100

The first problem is pricing exotics. Suppose one needs to price a call option with strike K_1 which has, say, a down-and-out knock-out at $K_2 < K_1$. Should we use the implied volatility at the call's strike K_1, the implied volatility at the barrier K_2, or some combination of the two to price this option? Clearly, this option cannot be priced without a single, self-consistent, model that works for all strikes without "adjustments".

The second problem is hedging. Since *different* models are being used for *different* strikes, it is not clear that the delta and vega risks calculated at one strike are consistent with the same risks calculated at other strikes. For example, suppose that our 1 month option book is long high strike options with a total Δ risk of $+\$1\ MM$, and is long low strike options with a Δ of $-\$1\ MM$. Is our option book really Δ-neutral, or do we have residual delta risk that needs to be hedged? Since different models are used at each strike, it is not clear that the risks offset each other. Consolidating vega risk raises similar concerns. Should we assume parallel or proportional shifts in volatility to calculate the total vega risk of our book? More explicitly, suppose that σ_B is 20% at $K = 100$ and 24% at $K = 90$, as shown for the 1 m options in Figure 2. Should we calculate vega by bumping σ_B by, say, 0.2% for both options? Or by bumping σ_B by 0.2% for the first option and by 0.24% for the second option? These questions are critical to effective book management, since this requires consolidating the delta and vega risks of *all* options on a given asset before hedging, so that only the net exposure of the book is hedged. Clearly one cannot answer these questions without a model that works for all strikes K.

The third problem concerns evolution of the implied volatility curve $\sigma_B(K)$. Since the implied volatility σ_B depends on the strike K, it is likely to also depend on the current value f of the forward price: $\sigma_B = \sigma_B(f, K)$. In this case there would be systematic changes in σ_B as the forward price f of the underlying changes. See Figure 2. Some of the vega risks of Black's model would actually be due to changes in the price of the underlying asset, and should be hedged more properly (and cheaply) as delta risks.

Local volatility models

An apparent solution to these problems is provided by the local volatility model of Dupire (1994), which is also attributed to Derman and Kani (1994, 1998). In an insightful work, Dupire essentially argued that Black was too bold in setting the coefficient $C(t, *)$ to $\sigma_B \hat{F}$. Instead one should only assume that C is Markovian: $C = C(t, \hat{F})$. Re-writing $C(t, \hat{F})$ as $\sigma_{loc}(t, \hat{F})\hat{F}$ then yields the "local volatility model," where the forward price of the asset is:

$$d\hat{F} = \sigma_{loc}(t, \hat{F})\hat{F}\, dW, \qquad \hat{F}(0) = f \tag{5a}$$

in the forward measure. Dupire argued that instead of theorizing about the unknown local volatility function $\sigma_{loc}(t, \hat{F})$, one should obtain $\sigma_{loc}(t, \hat{F})$ directly from the marketplace by "calibrating" the local volatility model to market prices of liquid European options.

In calibration, one starts with a given local volatility function $\sigma_{loc}(t, \hat{F})$, and evaluates:

$$V_{call} = D(t_{set})E\left\{[\hat{F}(t_{ex}) - K]^+ | \hat{F}(0) = f\right\} \tag{5b}$$

$$\equiv V_{put} + D(t_{set})(f - K) \tag{5c}$$

to obtain the theoretical prices of the options; one then varies the local volatility function $\sigma_{loc}(t, \hat{F})$ until these theoretical prices match the actual market prices of the option for each

strike K and exercise date t_{ex}. In practice liquid markets usually exist only for options with specific exercise dates $t_{ex}^1, t_{ex}^2, t_{ex}^3, \ldots$; for example, for 1 m, 2 m, 3 m, 6 m, and 12 m from today. Commonly the local vols $\sigma_{loc}(t, \hat{F})$ are taken to be piecewise constant in time:

$$\begin{aligned}
&\sigma_{loc}(t, \hat{F}) = \sigma_{loc}^{(1)}(\hat{F}) \qquad \text{for } t < t_{ex}^1 \\
&\sigma_{loc}(t, \hat{F}) = \sigma_{loc}^{(j)}(\hat{F}) \qquad \text{for } t_{ex}^{j-1} < t < t_{ex}^j \qquad j = 2, 3, \ldots J \\
&\sigma_{loc}(t, \hat{F}) = \sigma_{loc}^{(J)}(\hat{F}) \qquad \text{for } t > t_{ex}^J
\end{aligned} \tag{6}$$

One first calibrates $\sigma_{loc}^{(1)}(\hat{F})$ to reproduce the option prices at t_{ex}^1 for all strikes K, then calibrates $\sigma_{loc}^{(2)}(\hat{F})$ to reproduce the option prices at t_{ex}^2, for all K, and so forth. This calibration process can be greatly simplified by using the results in Hagan and Woodward (1999) and Hagan *et al* (in preparation). There we solve to obtain the prices of European options under the local volatility model (5a–5c), and from these prices we obtain explicit algebraic formulas for the implied volatility of the local vol models.

Once $\sigma_{loc}(t, \hat{F})$ has been obtained by calibration, the local volatility model is a single, self-consistent model which correctly reproduces the market prices of calls (and puts) for all strikes K and exercise dates t_{ex} without "adjustment". Prices of exotic options can now be calculated from this model without ambiguity. This model yields consistent delta and vega risks for all options, so these risks can be consolidated across strikes. Finally, perturbing f and re-calculating the option prices enables one to determine how the implied volatilities change with changes in the underlying asset price. Thus, the local volatility model thus provides *a* method of pricing and hedging options in the presence of market smiles and skews. It is perhaps the most popular method of managing exotic equity and foreign exchange options. Unfortunately, the local volatility model predicts the *wrong dynamics* of the implied volatility curve, which leads to inaccurate and often unstable hedges.

To illustrate the problem, consider the special case in which the local vol is a function of $\hat{F}$ only:

$$d\hat{F} = \sigma_{loc}(\hat{F})\hat{F}\, dW, \qquad \hat{F}(0) = f \tag{7}$$

In Hagan and Woodward (1999) and Hagan *et al* (in preparation). singular perturbation methods were used to analyze this model. There it was found that European call and put prices are given by Black's formula (4a–4c) with the implied volatility:

$$\sigma_B(K, f) = \sigma_{loc}\left(\frac{1}{2}[f + K]\right)\left\{1 + \frac{1}{24}\frac{\sigma''_{loc}\left(\frac{1}{2}[f + K]\right)}{\sigma_{loc}\left(\frac{1}{2}[f + K]\right)}(f - K)^2 + \cdots\right. \tag{8}$$

On the right-hand side, the first term dominates the solution and the second term provides a much smaller correction The omitted terms are very small, usually less than 1% of the first term.

The behavior of local volatility models can be largely understood by examining the first term in (8). The implied volatility depends on both the strike K and the current forward price f. So suppose that today the forward price is f_0 and the implied volatility curve seen in the

marketplace is $\sigma_B^0(K)$. Calibrating the model to the market clearly requires choosing the local volatility to be

$$\sigma_{loc}(\hat{F}) = \sigma_B^0(2\hat{F} - f_0)\{1 + \cdots\} \tag{9}$$

Now that the model is calibrated, let us examine its predictions. Suppose that the forward value changes from f_0 to some new value f. From (8), (9) we see that the model predicts that the new implied volatility curve is:

$$\sigma_B(K, f) = \sigma_B^0(K + f - f_0)\{1 + \cdots\} \tag{10}$$

for an option with strike K, given that the current value of the forward price is f. In particular, if the forward price f_0 *increases* to f, the implied volatility curve moves to the *left*; if f_0 *decreases* to f, the implied volatility curve moves to the *right. Local volatility models predict that the market smile/skew moves in the* opposite *direction as the price of the underlying asset*. This is opposite to typical market behavior, in which smiles and skews move in the *same* direction as the underlying.

To demonstrate the problem concretely, suppose that today's implied volatility is a perfect smile:

$$\sigma_B^0(K) = \alpha + \beta[K - f_0]^2 \tag{11a}$$

around today's forward price f_0. Then equation (8) implies that the local volatility is:

$$\sigma_{loc}(\hat{F}) = \alpha + 3\beta(\hat{F} - f_0)^2 + \cdots \tag{11b}$$

As the forward price f evolves away from f_0 due to normal market fluctuations, equation (8) predicts that the implied volatility is:

$$\sigma_B(K, f) = \alpha + \beta[K - (\tfrac{3}{2}f_0 - \tfrac{1}{2}f)]^2 + \tfrac{3}{4}\beta(f - f_0)^2 + \cdots \tag{11c}$$

The implied volatility curve not only moves in the opposite direction as the underlying, but the curve also shifts upward regardless of whether f increases or decreases. Exact results are illustrated in Figures 3–5. There we assumed that the local volatility $\sigma_{loc}(\hat{F})$ was given by (11b), and used finite difference methods to obtain essentially exact values for the option prices, and thus implied volatilities.

Hedges calculated from the local volatility model are wrong. To see this, let $BS(f, K, \sigma_B, t_{ex})$ be Black's formula (4a–4c) for, say, a call option. Under the local volatility model, the value of a call option is given by Black's formula:

$$V_{call} = BS(f, K, \sigma_B(K, f), t_{ex}) \tag{12a}$$

with the volatility $\sigma_B(K, f)$ given by (8). Differentiating with respect to f yields the Δ risk:

$$\Delta \equiv \frac{\partial V_{call}}{\partial f} = \frac{\partial BS}{\partial f} + \frac{\partial BS}{\partial \sigma_B}\frac{\partial \sigma_B(K, f)}{\partial f} \tag{12b}$$

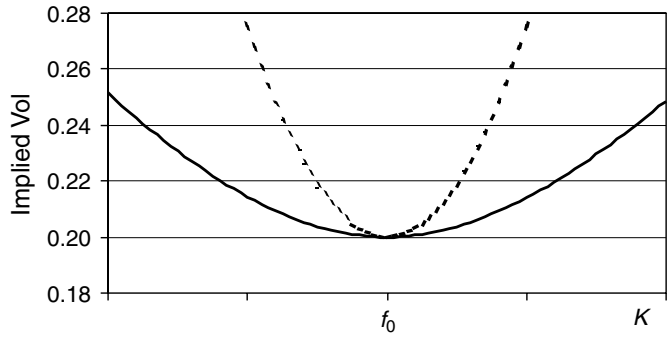


Figure 3: Exact implied volatility $\sigma_B(K, f_0)$ (solid line) obtained from the local volatility $\sigma_{loc}(\hat{F})$ (dashed line)

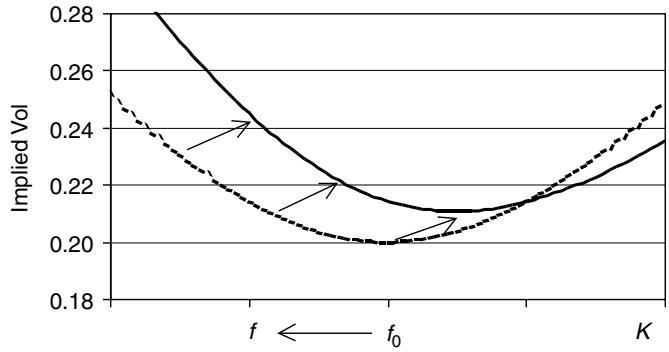


Figure 4: Implied volatility $\sigma_B(K, f)$ if the forward price *decreases* from f_0 to f (solid line)

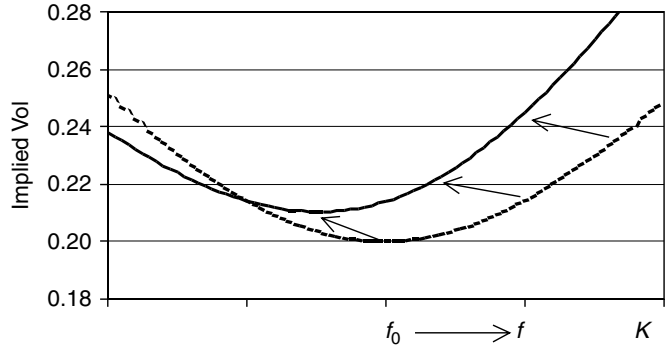


Figure 5: Implied volatility $\sigma_B(K, f)$ if the forward prices *increases* from f_0 to f (solid line)

predicted by the local volatility model. The first term is clearly the Δ risk one would calculate from Black's model using the implied volatility from the market. The second term is the local volatility model's correction to the Δ risk, which consists of the Black vega risk multiplied by the *predicted* change in σ_B due to changes in the underlying forward price f. In real markets the implied volatility moves in the *opposite* direction as the direction predicted by the model. Therefore, the correction term needed for real markets should have the *opposite sign* as the correction predicted by the local volatility model. *The original Black model yields more accurate hedges than the local volatility model, even though the local vol model is self-consistent across strikes and Black's model is inconsistent*.

Local volatility models are also peculiar theoretically. Using any function for the local volatility $\sigma_{loc}(t, \hat{F})$ *except* for a power law:

$$C(t, *) = \alpha(t)\hat{F}^{\beta} \tag{13}$$

$$\sigma_{loc}(t, \hat{F}) = \alpha(t)\hat{F}^{\beta}/\hat{F} = \alpha(t)/\hat{F}^{1-\beta} \tag{14}$$

introduces an intrinsic "length scale" for the forward price $\hat{F}$ into the model. That is, the model becomes inhomogeneous in the forward price $\hat{F}$. Although intrinsic length scales are theoretically possible, it is difficult to understand the financial origin and meaning of these scales (Wan, 1991), and one naturally wonders whether such scales should be introduced into a model without specific theoretical justification.

The SABR model

The failure of the local volatility model means that we cannot use a Markovian model based on a single Brownian motion to manage our smile risk. Instead of making the model non-Markovian, or basing it on non-Brownian motion, we choose to develop a two factor model. To select the second factor, we note that most markets experience both relatively quiescent and relatively chaotic periods. This suggests that volatility is not constant, but is itself a random function of time. Respecting the preceding discussion, we choose the unknown coefficient $C(t, *)$ to be $\hat{\alpha}\hat{F}^{\beta}$, where the "volatility" $\hat{\alpha}$ is itself a stochastic process. Choosing the simplest reasonable process for $\hat{\alpha}$ now yields the "stochastic-$\alpha\beta\rho$ model," which has become known as the *SABR model*. In this model, the forward price and volatility are

$$d\hat{F} = \hat{\alpha}\hat{F}^{\beta}\, dW_1, \qquad \hat{F}(0) = f \tag{15a}$$

$$d\hat{\alpha} = \nu\hat{\alpha}\, dW_2, \qquad \hat{\alpha}(0) = \alpha \tag{15b}$$

under the forward measure, where the two processes are correlated by:

$$dW_1\, dW_2 = \rho\, dt \tag{15c}$$

Many other stochastic volatility models have been proposed (e.g., Hull and White, 1987; Heston, 1993; Lewis, 2000; Fouque *et al*, 2000). However, the SABR model has the virtue of being the simplest stochastic volatility model which is homogenous in $\hat{F}$ and $\hat{\alpha}$. We shall find that the SABR model can be used to accurately fit the implied volatility curves observed in the marketplace for any single exercise date t_{ex}. More importantly, it predicts the correct dynamics of the implied volatility curves. This makes the SABR model an effective means to manage the

smile risk in markets where each asset only has a single exercise date; these markets include the swaption and caplet/floorlet markets.

As written, the SABR model may or may not fit the observed *volatility surface* of an asset which has European options at several different exercise dates; such markets include foreign exchange options and most equity options. Fitting volatility surfaces requires the *dynamic SABR model* which is discussed in the Appendix.

It has been claimed by many authors that stochastic volatility models are models of incomplete markets, because the stochastic volatility risk cannot be hedged. This is not true. It is true that the risk to changes in $\hat{\alpha}$ (the vega risk) cannot be hedged by buying or selling the underlying asset. However, vega risk can be hedged by buying or selling options on the asset in exactly the same way that Δ-hedging is used to neutralize the risks to changes in the price $\hat{F}$. In practice, vega risks are hedged by buying and selling options as a matter of routine, so whether the market would be complete if these risks were not hedged is a moot question.

The SABR model (15a–15c) is analyzed in Appendix B. There singular perturbation techniques are used to obtain the prices of European options. From these prices, the options' implied volatility $\sigma_B(K, f)$ is then obtained. The upshot of this analysis is that under the SABR model, the price of European options is given by Black's formula:

$$V_{call} = D(t_{set})\{f\mathcal{N}(d_1) - K\mathcal{N}(d_2)\} \tag{16a}$$

$$V_{put} = V_{call} + D(t_{set})[K - f] \tag{16b}$$

with:

$$d_{1,2} = \frac{\log f/K \pm \frac{1}{2}\sigma_B^2 t_{ex}}{\sigma_B\sqrt{t_{ex}}} \tag{16c}$$

where the implied volatility $\sigma_B(K, f)$ is given by:

$$\begin{aligned}\sigma_B(K, f) &= \frac{\alpha}{(fK)^{(1-\beta)/2}\left\{1 + \frac{(1-\beta)^2}{24}\log^2 f/K + \frac{(1-\beta)^4}{1920}\log^4 f/K + \cdots\right\}} \cdot \left(\frac{z}{x(z)}\right) \\ &\quad \cdot \left\{1 + \left[\frac{(1-\beta)^2}{24}\frac{\alpha^2}{(fK)^{1-\beta}} + \frac{1}{4}\frac{\rho\beta\nu\alpha}{(fK)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2\right]t_{ex} + \cdots\right.\end{aligned} \tag{17a}$$

Here

$$z = \frac{\nu}{\alpha}(fK)^{(1-\beta)/2}\log f/K \tag{17b}$$

and $x(z)$ is defined by:

$$x(z) = \log\left\{\frac{\sqrt{1 - 2\rho z + z^2} + z - \rho}{1 - \rho}\right\} \tag{17c}$$

For the special case of at-the-money options, options struck at $K = f$, this formula reduces to:

$$\sigma_{ATM} = \sigma_B(f, f) = \frac{\alpha}{f^{(1-\beta)}} \left\{ 1 + \left[\frac{(1-\beta)^2}{24} \frac{\alpha^2}{f^{2-2\beta}} + \frac{1}{4} \frac{\rho\beta\alpha\nu}{f^{(1-\beta)}} + \frac{2-3\rho^2}{24} \nu^2 \right] t_{ex} + \cdots \right. \tag{18}$$

These formulas are the main result of this paper. Although it appears formidable, the formula is explicit and only involves elementary trigonometric functions. Implementing the SABR model for vanilla options is very easy, since once this formula is programmed, we just need to send the options to a Black pricer. In the next section we examine the qualitative behavior of this formula, and how it can be used to manage smile risk.

The complexity of the formula is needed for accurate pricing. Omitting the last line of (17a), for example, can result in a relative error that exceeds 3 per cent in extreme cases. Although this error term seems small, it is large enough to be required for accurate pricing. The omitted terms "$+\cdots$" are *much, much* smaller. Indeed, even though we have derived more accurate expressions by continuing the perturbation expansion to higher order, (17a–17c) is the formula we use to value and hedge our vanilla swaptions, caps, and floors. We have not implemented the higher order results, believing that the increased precision of the higher order results is superfluous.

There are two special cases of note: $\beta = 1$, representing a stochastic log normal model), and $\beta = 0$, representing a stochastic normal model. The implied volatility for these special cases is obtained in the last section of Appendix B.

Managing smile risk

The complexity of the above formula for $\sigma_B(K, f)$ obscures the qualitative behavior of the SABR model. To make the model's phenomenology and dynamics more transparent, note that formula (17a–17c) can be approximated as:

$$\sigma_B(K, f) = \frac{\alpha}{f^{1-\beta}} \left\{ 1 - \frac{1}{2}(1-\beta-\rho\lambda) \log K/f + \frac{1}{12}[(1-\beta)^2 + (2-3\rho^2)\lambda^2] \log^2 K/f + \cdots \right. \tag{19a}$$

provided that the strike K is not too far from the current forward f. Here the ratio:

$$\lambda = \frac{\nu}{\alpha} f^{1-\beta} \tag{19b}$$

measures the strength ν of the volatility of volatility (the "volvol") compared to the local volatility $\alpha/f^{1-\beta}$ at the current forward. Although equations (19a–19b) should not be used to price real deals, they are accurate enough to depict the qualitative behavior of the SABR model faithfully.

As f varies during normal trading, the curve that the ATM volatility $\sigma_B(f, f)$ traces is known as the *backbone*, while the *smile* and *skew* refer to the implied volatility $\sigma_B(K, f)$ as

a function of strike K for a fixed f. That is, the market smile/skew gives a snapshot of the market prices for different strikes K at a given instance, when the forward f has a specific price. Figures 6 and 7 show the dynamics of the smile/skew predicted by the SABR model.

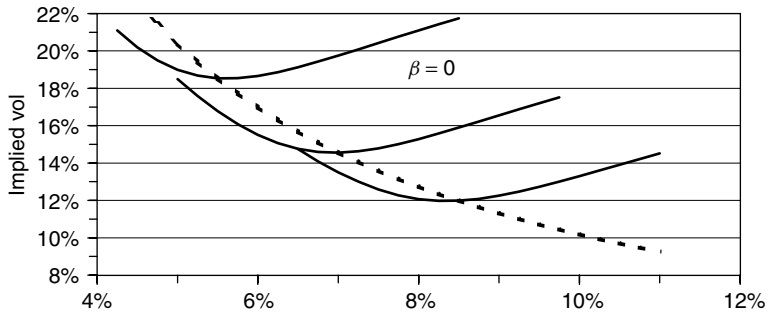


Figure 6: Backbone and smiles for $\beta = 0$. As the forward f varies, the implied volatility $\sigma_B(f, f)$ of ATM options traverses the backbone (dashed curve). Shown are the smiles $\sigma_B(K, f)$ for three different values of the forward. Volatility data from 1 into 1 swaption on 4/28/00, courtesy of Cantor-Fitzgerald

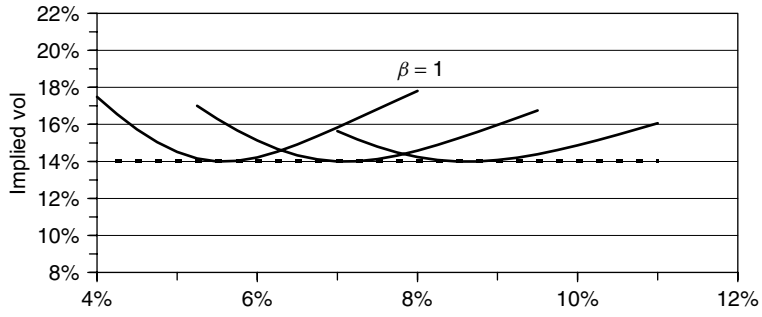


Figure 7: Backbone and smiles as above, but for $\beta = 1$

Let us now consider the implied volatility $\sigma_B(K, f)$ in detail. The first factor $\alpha/f^{1-\beta}$ in (19a) is the implied volatility for at-the-money (ATM) options, options whose strike K equals the current forward f. So the backbone traversed by ATM options is essentially $\sigma_B(f, f) = \alpha/f^{1-\beta}$ for the SABR model. The backbone is almost entirely determined by the exponent β, with the exponent $\beta = 0$ (a stochastic Gaussian model) giving a steeply downward sloping backbone, and the exponent $\beta = 1$, giving a nearly flat backbone.

The second term $-\frac{1}{2}(1 - \beta - \rho\lambda) \log K/f$ represents the skew, the slope of the implied volatility with respect to the strike K. The $-\frac{1}{2}(1 - \beta) \log K/f$ part is the *beta skew*, which is downward sloping since $0 \leq \beta \leq 1$. It arises because the "local volatility" $\hat{\alpha}\hat{F}^{\beta}/\hat{F}^{1} = \hat{\alpha}/\hat{F}^{1-\beta}$ is a decreasing function of the forward price. The second part $\frac{1}{2}\rho\lambda \log K/f$ is the *vanna* skew, the skew caused by the correlation between the volatility and the asset price. Typically

the volatility and asset price are negatively correlated, so on average, the volatility α would decrease (increase) when the forward f increases (decreases). It thus seems unsurprising that a negative correlation ρ causes a downward sloping vanna skew.

It is interesting to compare the skew to the slope of the backbone. As f changes to f' the ATM vol changes to:

$$\sigma_B(f', f') = \frac{\alpha}{f^{1-\beta}} \left\{ 1 - (1-\beta)\frac{f'-f}{f} + \cdots \right\} \tag{20a}$$

Near $K = f$, the β component of skew expands as

$$\sigma_B(K, f) = \frac{\alpha}{f^{1-\beta}} \left\{ 1 - \frac{1}{2}(1-\beta)\frac{K-f}{f} + \cdots \right\} \tag{20b}$$

so the slope of the backbone $\sigma_B(f, f)$ is twice as steep as the slope of the smile $\sigma_B(K, f)$ due to the β-component of the skew.

The last term in (19a) also contains two parts. The first part $\frac{1}{12}(1-\beta)^2 \log^2 K/f$ appears to be a smile (quadratic) term, but it is dominated by the downward sloping beta skew, and, at reasonable strikes K, it just modifies this skew somewhat. The second part $\frac{1}{12}(2-3\rho^2)\lambda^2 \log^2 K/f$ is the smile induced by the *volga* (vol-gamma) effect. Physically this smile arises because of "adverse selection": unusually large movements of the forward $\hat{F}$ happen more often when the volatility α increases, and less often when α decreases, so strikes K far from the money represent, on average, high volatility environments.

Fitting market data

The exponent β and correlation ρ affect the volatility smile in similar ways. They both cause a downward sloping skew in $\sigma_B(K, f)$ as the strike K varies. From a single market snapshot of $\sigma_B(K, f)$ as a function of K at a given f, it is difficult to distinguish between the two parameters. This is demonstrated by Figure 8. There we fit the SABR parameters α, ρ, ν with $\beta = 0$ and then re-fit the parameters α, ρ, ν with $\beta = 1$. Note that there is no substantial difference in the quality of the fits, despite the presence of market noise. This matches our general experience: market smiles can be fit equally well with any specific value of β. In particular, β cannot be determined by fitting a market smile since this would clearly amount to "fitting the noise."

Figure 8 also exhibits a common data quality issue. Options with strikes K away from the current forward f trade less frequently than at-the-money and near-the-money options. Consequently, as K moves away from f, the volatility quotes become more suspect because they are more likely to be out-of-date and not represent bona fide offers to buy or sell options.

Suppose for the moment that the exponent β is known or has been selected. Taking a snapshot of the market yields the implied volatility $\sigma_B(K, f)$ as a function of the strike K at the current forward price f. With β given, fitting the SABR model is a straightforward procedure. The three parameters α, ρ, and ν have different effects on the curve: the parameter α mainly controls the overall height of the curve, changing the correlation ρ controls the curve's skew, and changing the vol of vol ν controls how much smile the curve exhibits. Because of the widely separated roles these parameters play, the fitted parameter values tend to be very stable, even in the presence of large amounts of market noise.

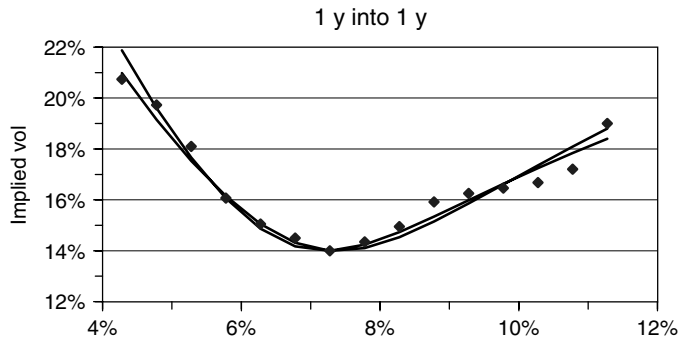


Figure 8: Implied volatilities as a function of strike. Shown are the curves obtained by fitting the SABR model with exponent $\beta = 0$ and with $\beta = 1$ to the 1y into 1y swaption vol observed on 4/28/00. As usual, both fits are equally good. Data courtesy of Cantor-Fitzgerald

The exponent β can be determined from historical observations of the "backbone" or selected from "aesthetic considerations." Equation (18) shows that the implied volatility of ATM options is:

$$\log \sigma_B(f, f) = \log \alpha - (1-\beta)\log f + \log\left\{1 + \left[\frac{(1-\beta)^2}{24}\frac{\alpha^2}{f^{2-2\beta}} + \frac{1}{4}\frac{\rho\beta\alpha\nu}{f^{(1-\beta)}} + \frac{2-3\rho^2}{24}\nu^2\right]t_{ex} + \cdots\right\} \tag{21}$$

The exponent β can be extracted from a log plot of historical observations of f, σ_{ATM} pairs. Since both f and α are stochastic variables, this fitting procedure can be quite noisy, and as the $[\cdots]t_{ex}$ term is typically less than 1 or 2 per cent, it is usually ignored in fitting β.

Selecting β from "aesthetic" or other *a priori* considerations usually results in $\beta = 1$ (stochastic lognormal), $\beta = 0$ (stochastic normal), or $\beta = \frac{1}{2}$ (stochastic CIR) models. Proponents of $\beta = 1$ cite log normal models as being "more natural." or believe that the horizontal backbone best represents their market. These proponents often include desks trading foreign exchange options. Proponents of $\beta = 0$ usually believe that a normal model, with its symmetric break-even points, is a more effective tool for managing risks, and would claim that $\beta = 0$ is essential for trading markets like Yen interest rates, where the forwards f can be negative or near zero. Proponents of $\beta = \frac{1}{2}$ are usually US interest rate desks that have developed trust in CIR models.

It is usually more convenient to use the at-the-money volatility σ_{ATM}, β, ρ, and ν as the SABR parameters instead of the original parameters α, β, ρ, ν. The parameter α is then found whenever needed by inverting (18) on the fly; this inversion is numerically easy since the $[\cdots]t_{ex}$ term is small. With this parameterization, fitting the SABR model requires fitting ρ and ν to the implied volatility curve, with σ_{ATM} and β given. In many markets, the ATM volatilities need to be updated frequently, say once or twice a day, while the smiles and skews need to be updated infrequently, say once or twice a month. With the new parameterization, σ_{ATM} can be updated as often as needed, with ρ, ν (and β) updated only as needed.

Let us apply SABR to options on US dollar interest rates. There are three key groups of European options on US rates: Eurodollar future options, caps/floors, and European swaptions. Eurodollar future options are exchange-traded options on the 3 month Libor rate; like interest rate futures, EDF options are quoted on $100(1 - r_{Libor})$. Figure 1 fits the SABR model (with $\beta = 1$) to the implied volatility for the June 99 contracts, and Figures 9–12 fit the model (also with $\beta = 1$) to the implied volatility for the September 99, December 99, March 00 and June 00 contracts. All prices were obtained from Bloomberg Information Services on 23 March 1999. Two points are shown for the same strike where there are quotes for both puts and calls. Note that market liquidity dries up for the later contracts, and for strikes that are too far from the money. Consequently, more market noise is seen for these options.

Caps and floors are sums of caplets and floorlets; each caplet and floorlet is a European option on the 3 month Libor rate. We do not consider the cap/floor market here because the broker-quoted cap prices must be "stripped" to obtain the caplet volatilities before SABR can be applied.

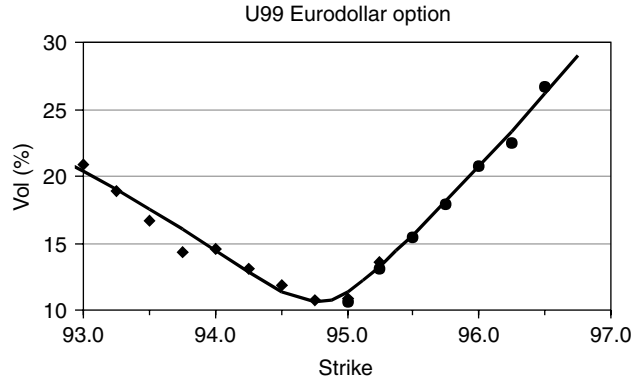


Figure 9: Volatility of the Sep 99 EDF options

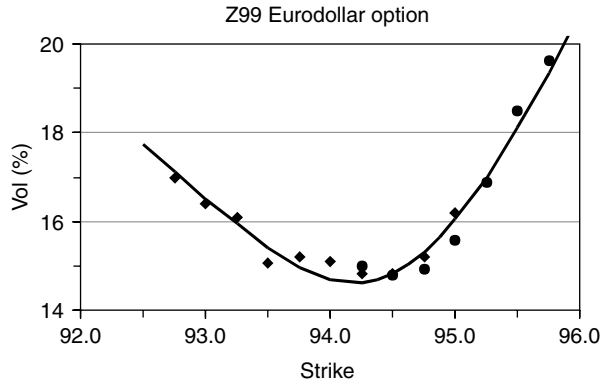


Figure 10: Volatility of the Dec 99 EDF options

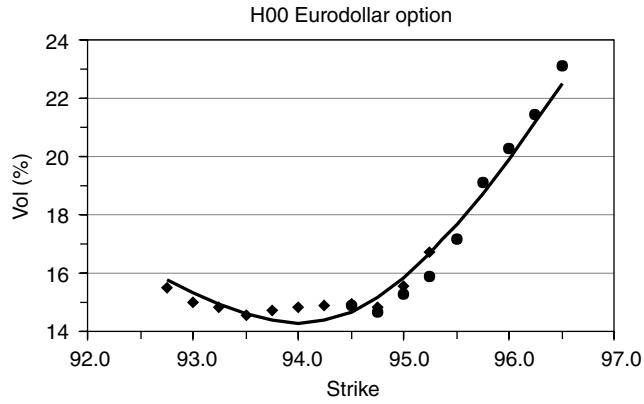


Figure 11: Volatility of the Mar 00 EDF options

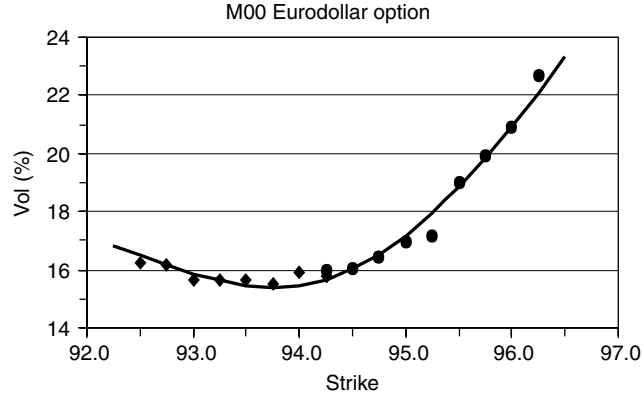


Figure 12: Volatility of the Jun 00 EDF options

A *m year into n year* swaption is a European option with m years to the exercise date (the maturity); if it is exercised, then one receives an n year swap (the tenor, or underlying) on the 3 month Libor rate. For almost all maturities and tenors, the US swaption market is liquid for at-the-money swaptions, but is ill-liquid for swaptions struck away from the money. Hence, market data is somewhat suspect for swaptions that are not struck near the money. Figures 13–16 fits the SABR model (with $\beta = 1$) to the prices of m into 5Y swaptions observed on 28 April 2000. Data supplied courtesy of Cantor-Fitzgerald.

We observe that the smile and skew depend heavily on the time-to-exercise for Eurodollar future options and swaptions. The smile is pronounced for short-dated options and flattens for longer dated options; the skew is overwhelmed by the smile for short-dated options, but is important for long-dated options. This picture is confirmed Tables 1 and 2. These tables show the values of the vol of vol ν and correlation ρ obtained by fitting the smile and skew of each "m into n" swaption, again using the data from 28 April 2000. Note that the vol of vol ν is very high for short dated options, and decreases as the time-to-exercise increases, while the correlations starts near zero and becomes substantially negative. Also note that there is little

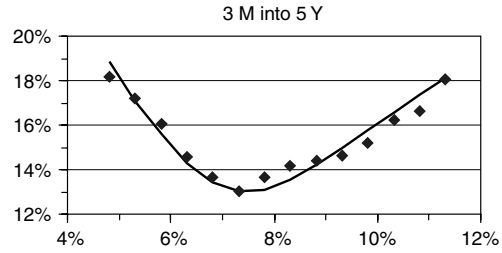


Figure 13: Volatilities of 3 month into 5 year swaption

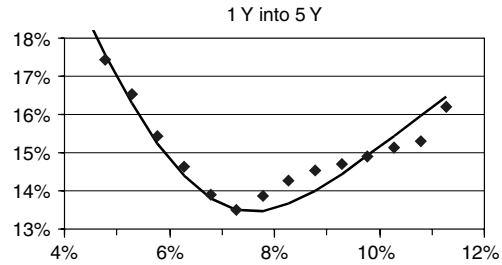


Figure 14: Volatilities of 1 year into 5 year swaptions

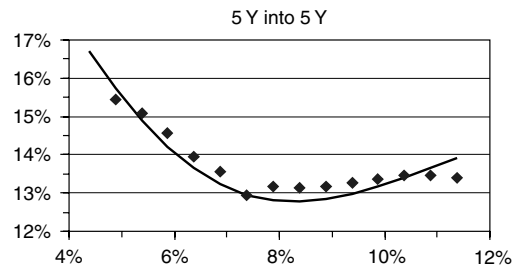


Figure 15: Volatilities of 5 year into 5 year swaptions

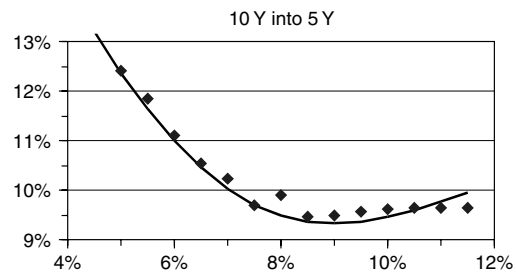


Figure 16: Volatilities of 10 year into 5 year options

TABLE 1: VOLATILITY OF VOLATILITY ν FOR EUROPEAN SWAPTIONS. ROWS ARE TIME–TO–EXERCISE; COLUMNS ARE TENOR OF THE UNDERLYING SWAP

	1Y	2Y	3Y	4Y	5Y	7Y	10Y
1M	76.2%	75.4%	74.6%	74.1%	75.2%	73.7%	74.1%
3M	65.1%	62.0%	60.7%	60.1%	62.9%	59.7%	59.5%
6M	57.1%	52.6%	51.4%	50.8%	49.4%	50.4%	50.0%
1Y	59.8%	49.3%	47.1%	46.7%	46.0%	45.6%	44.7%
3Y	42.1%	39.1%	38.4%	38.4%	36.9%	38.0%	37.6%
5Y	33.4%	33.2%	33.1%	32.6%	31.3%	32.3%	32.2%
7Y	30.2%	29.2%	29.0%	28.2%	26.2%	27.2%	27.0%
10Y	26.7%	26.3%	26.0%	25.6%	24.8%	24.7%	24.5%

TABLE 2: MATRIX OF CORRELATIONS ρ BETWEEN THE UNDERLYING AND THE VOLATILITY FOR EUROPEAN SWAPTIONS

	1Y	2Y	3Y	4Y	5Y	7Y	10Y
1M	4.2%	−0.2%	−0.7%	−1.0%	−2.5%	−1.8%	−2.3%
3M	2.5%	−4.9%	−5.9%	−6.5%	−6.9%	−7.6%	−8.5%
6M	5.0%	−3.6%	−4.9%	−5.6%	−7.1%	−7.0%	−8.0%
1Y	−4.4%	−8.1%	−8.8%	−9.3%	−9.8%	−10.2%	−10.9%
3Y	−7.3%	−14.3%	−17.1%	−17.1%	−16.6%	−17.9%	−18.9%
5Y	−11.1%	−17.3%	−18.5%	−18.8%	−19.0%	−20.0%	−21.6%
7Y	−13.7%	−22.0%	−23.6%	−24.0%	−25.0%	−26.1%	−28.7%
10Y	−14.8%	−25.5%	−27.7%	−29.2%	−31.7%	−32.3%	−33.7%

dependence of the market skew/smile on the length of the underlying swap; both ν and ρ are fairly constant across each row. This matches our general experience: in most markets there is a strong smile for short-dated options which relaxes as the time-to-expiry increases; consequently the volatility of volatility ν is large for short dated options and smaller for long-dated options, regardless of the particular underlying. Our experience with correlations is less clear: in some markets a nearly flat skew for short maturity options develops into a strongly downward sloping skew for longer maturities. In other markets there is a strong downward skew for all option maturities, and in still other markets the skew is close to zero for all maturities.

Managing smile risk

After choosing β and fitting ρ, ν, and either α or σ_{ATM}, the SABR model

$$d\hat{F} = \hat{\alpha}\hat{F}^{\beta}\,dW_1, \qquad \hat{F}(0) = f \tag{22a}$$

$$d\hat{\alpha} = \nu\hat{\alpha}\,dW_2, \qquad \hat{\alpha}(0) = \alpha \tag{22b}$$

with

$$dW_1\,dW_2 = \rho\,dt \tag{22c}$$

fits the smiles and skews observed in the market quite well, especially considering the quality of price quotes away from the money. Let us take for granted that it fits well enough. Then we have a single, self-consistent model that fits the option prices for all strikes K without "adjustment", so we can use this model to price exotic options without ambiguity. The SABR model also predicts that whenever the forward price f changes, the implied volatility curve shifts in the *same* direction and by the *same* amount as the price f. This predicted dynamics of the smile matches market experience. If $\beta < 1$, the "backbone" is downward sloping, so the shift in the implied volatility curve is not purely horizontal. Instead, this curve shifts up and down as the at-the-money point traverses the backbone. Our experience suggests that the parameters ρ and ν are very stable (β is assumed to be a given constant), and need to be re-fit only every few weeks. This stability may be because the SABR model reproduces the usual dynamics of smiles and skews. In contrast, the at-the-money volatility σ_{ATM} or, equivalently, α may need to be updated every few hours in fast-paced markets.

Since the SABR model is a single self-consistent model for all strikes K, the risks calculated at one strike are consistent with the risks calculated at other strikes. Therefore the risks of all the options on the same asset can be added together, and only the residual risk needs to be hedged.

Let us set aside the Δ risk for the moment, and calculate the other risks. Let $BS(f, K, \sigma_B, t_{ex})$ be Black's formula (4a–4c) for, say, a call option. According to the SABR model, the value of a call is:

$$V_{call} = BS(f, K, \sigma_B(K, f), t_{ex}) \tag{23}$$

where the volatility $\sigma_B(K, f) \equiv \sigma_B(K, f; \alpha, \beta, \rho, \nu)$ is given by equations (17a–17c). Differentiating (in practice risks are calculated by finite differences: valuing the option at α, re-valuing the option after bumping the forward to $\alpha + \delta$, and then subtracting to determine the risk. This saves differentiating complex formulas such as (17a–17c)). with respect to α yields the vega risk, the risk to overall changes in volatility:

$$\frac{\partial V_{call}}{\partial \alpha} = \frac{\partial BS}{\partial \sigma_B} \cdot \frac{\partial \sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial \alpha}. \tag{24}$$

This risk is the change in value when α changes by a unit amount. It is traditional to scale vega so that it represents the change in value when the ATM volatility changes by a unit amount.

Since $\delta\sigma_{ATM} = (\partial\sigma_{ATM}/\partial\alpha)\delta\alpha$, the vega risk is:

$$\text{vega} \equiv \frac{\partial V_{call}}{\partial\alpha} = \frac{\partial BS}{\partial\sigma_B} \cdot \frac{\dfrac{\partial\sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial\alpha}}{\dfrac{\partial\sigma_{ATM}(f; \alpha, \beta, \rho, \nu)}{\partial\alpha}} \tag{25a}$$

where $\sigma_{ATM}(f) = \sigma_B(f, f)$ is given by (18). Note that to leading order, $\partial\sigma_B/\partial\alpha \approx \sigma_B/\alpha$ and $\partial\sigma_{ATM}/\partial\alpha \approx \sigma_{ATM}/\alpha$, so the vega risk is roughly given by:

$$\text{vega} \approx \frac{\partial BS}{\partial\sigma_B} \cdot \frac{\sigma_B(K, f)}{\sigma_{ATM}(f)} = \frac{\partial BS}{\partial\sigma_B} \cdot \frac{\sigma_B(K, f)}{\sigma_B(f, f)}. \tag{25b}$$

Qualitatively, then, vega risks at different strikes are calculated by bumping the implied volatility at each strike K by an amount that is proportional to the implied volatility $\sigma_B(K, f)$ at that strike. That is, in using equation (25a), we are essentially using proportional, and not parallel, shifts of the volatility curve to calculate the total vega risk of a book of options.

Since ρ and ν are determined by fitting the implied volatility curve observed in the marketplace, the SABR model has risks to ρ and ν changing. Borrowing terminology from foreign exchange desks, vanna is the risk to ρ changing and volga (vol gamma) is the risk to ν changing:

$$\text{vanna} = \frac{\partial V_{call}}{\partial\rho} = \frac{\partial BS}{\partial\sigma_B} \cdot \frac{\partial\sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial\rho}, \tag{26a}$$

$$\text{volga} = \frac{\partial V_{call}}{\partial\nu} = \frac{\partial BS}{\partial\sigma_B} \cdot \frac{\partial\sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial\nu}. \tag{26b}$$

Vanna basically expresses the risk to the skew increasing, and volga expresses the risk to the smile becoming more pronounced. These risks are easily calculated by using finite differences on the formula for σ_B in equations (17a–17c). If desired, these risks can be hedged by buying or selling away-from-the-money options.

The delta risk expressed by the SABR model depends on whether one uses the parameterization α, β, ρ, ν or σ_{ATM}, β, ρ, ν. Suppose first we use the parameterization α, β, ρ, ν, so that $\sigma_B(K, f) \equiv \sigma_B(K, f; \alpha, \beta, \rho, \nu)$. Differentiating with respect to f yields the Δ risk:

$$\Delta \equiv \frac{\partial V_{call}}{\partial f} = \frac{\partial BS}{\partial f} + \frac{\partial BS}{\partial\sigma_B} \frac{\partial\sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial f}. \tag{27}$$

The first term is the ordinary Δ risk one would calculate from Black's model. The second term is the SABR model's correction to the Δ risk. It consists of the Black vega times the *predicted* change in the implied volatility σ_B caused by the change in the forward f. As discussed above, the predicted change consists of a sideways movement of the volatility curve in the same direction (and by the same amount) as the change in the forward price f. In addition, if $\beta < 1$ the volatility curve rises and falls as the at-the-money point traverses up and down the backbone. There may also be minor changes to the shape of the skew/smile due to changes in f.

Now suppose we use the parameterization σ_{AMT}, β, ρ, ν. Then α is a function of σ_{ATM} and f defined implicitly by (18). Differentiating (23) now yields the Δ risk:

$$\Delta \equiv \frac{\partial BS}{\partial f} + \frac{\partial BS}{\partial \sigma_B}\left\{\frac{\partial \sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial f} + \frac{\partial \sigma_B(K, f; \alpha, \beta, \rho, \nu)}{\partial \alpha}\frac{\partial \alpha(\sigma_{ATM}, f)}{\partial f}\right\}. \tag{28}$$

The delta risk is now the risk to changes in f with σ_{ATM} held fixed. The last term is just the change in α needed to keep σ_{ATM} constant while f changes. Clearly this last term must just cancel out the vertical component of the backbone, leaving only the sideways movement of the implied volatility curve. Note that this term is zero for $\beta = 1$.

Theoretically one should use the Δ from equation (27) to risk manage option books. In many markets, however, it may take several days for volatilities σ_B to change following significant changes in the forward price f. In these markets, using Δ from (28) is a *much* more effective hedge. For suppose one used Δ from equation (27). Then, when the volatility σ_{ATM} did not immediately change following a change in f, one would be forced to re-mark α to compensate, and this re-marking would change the Δ hedges. As σ_{ATM} equilibrated over the next few days, one would mark α back to its original value, which would change the Δ hedges back to their original value. This "hedging chatter" caused by market delays can prove to be costly.

Appendix A. Analysis of the SABR model

Here we use singular perturbation techniques to price European options under the SABR model. Our analysis is based on a small volatility expansion, where we take both the volatility $\hat{\alpha}$ and the "volvol" ν to be small. To carry out this analysis in a systematic fashion, we re-write $\hat{\alpha} \longrightarrow \varepsilon\hat{\alpha}$, and $\nu \longrightarrow \varepsilon\nu$, and analyze:

$$d\hat{F} = \varepsilon\hat{\alpha}C(\hat{F})\,dW_1, \tag{A.1a}$$

$$d\hat{\alpha} = \varepsilon\nu\hat{\alpha}\,dW_2, \tag{A.1b}$$

with:

$$dW_1\,dW_2 = \rho\,dt, \tag{A.1c}$$

in the limit $\varepsilon \ll 1$. This is the *distinguished limit* (Cole, 1968; Kevorkian and Cole, 1985) in the language of singular perturbation theory. After obtaining the results we replace $\varepsilon\hat{\alpha} \longrightarrow \hat{\alpha}$, and $\varepsilon\nu \longrightarrow \nu$ to get the answer in terms of the original variables. We first analyze the model with a general $C(\hat{F})$, and then specialize the results to the power law $\hat{F}^{\beta}$. This is notationally simpler than working with the power law throughout, and the more general result may prove valuable in some future application.

We first use the forward Kolmogorov equation to simplify the option pricing problem. Suppose the economy is in state $\hat{F}(t) = f, \hat{\alpha}(t) = \alpha$ at date t. Define the probability density $p(t, f, \alpha; T, F, A)$ by:

$$p(t, f, \alpha; T, F, A)\,dF\,dA = \text{prob}\left\{F < \hat{F}(T) < F + dF.A < \hat{\alpha}(T) < A + dA|\hat{F}(t) = f,\ \hat{\alpha}(t) = \alpha\right\}. \tag{A.2}$$

As a function of the *forward variables* T, F, A, the density p satisfies the forward Kolmogorov equation (the Fŏkker–Planck equation):

$$\begin{aligned} pT &= \tfrac{1}{2}\varepsilon^2 A^2 [C^2(F)p]_{FF} \\ &\quad + \varepsilon^2 \rho\nu [A^2 C(F)p]_{FA} + \tfrac{1}{2}\varepsilon^2 \nu^2 [A^2 p]_{AA} \qquad \text{for } T > t \end{aligned} \tag{A.3a}$$

with:

$$p = \delta(F - f)\delta(A - \alpha) \quad \text{at } T = t, \tag{A.3b}$$

as is well-known (Karatzas and Shreve, 1988; Okdendal, 1998; Musiela and Rutkowski, 1998). Here, and throughout, we use subscripts to denote partial derivatives.

Let $V(t, f, \alpha)$ be the value of a European call option at date t, when the economy is in state $\hat{F}(t) = f, \hat{\alpha}(t) = \alpha$. Let t_{ex} be the option's exercise date, and let K be its strike. Omitting the discount factor $D(t_{set})$, which factors out exactly, the value of the option is:

$$\begin{aligned} V(t, f, \alpha) &= E\left\{[\hat{F}(t_{ex}) - K]^+ | \hat{F}(t) = f, \hat{\alpha}(t) = \alpha\right\} \\ &= \int_{-\infty}^{\infty}\int_{K}^{\infty} (F - K)\, p(t, f, \alpha; t_{ex}, F, A)\, dF\, dA. \end{aligned} \tag{A.4}$$

See (1a). Since:

$$p(t, f, \alpha; t_{ex}, F, A) = \delta(F - f)\delta(A - \alpha) + \int_t^{t_{ex}} p_T(t, f, \alpha; T, F, A)\, dT \tag{A.5}$$

we can re-write $V(t, f, \alpha)$ as:

$$V(t, f, \alpha) = [f - K]^+ + \int_t^{t_{ex}}\int_K^{\infty}\int_{-\infty}^{\infty} (F - K)\, p_T(t, f, \alpha; T, F, A)\, dA\, dF\, dT \tag{A.6}$$

We substitute (A.3a) for p_T into (A.6). Integrating the A derivatives $\varepsilon^2\rho\nu[A^2C(F)p]_{FA}$ and $\frac{1}{2}\varepsilon^2\nu^2[A^2p]_{AA}$ over all A yields zero. Therefore our option price reduces to:

$$V(t, f, \alpha) = [f - K]^+ + \frac{1}{2}\varepsilon^2 \int_t^{t_{ex}}\int_{-\infty}^{\infty}\int_K^{\infty} A^2 (F - K)[C^2(F)p]_{FF}\, dF\, dA\, dT \tag{A.7}$$

where we have switched the order of integration. Integrating by parts twice with respect to F now yields:

$$V(t, f, \alpha) = [f - K]^+ + \frac{1}{2}\varepsilon^2 C^2(K) \int_t^{t_{ex}}\int_{-\infty}^{\infty} A^2 p(t, f, \alpha; T, K, A)\, dA\, dT \tag{A.8}$$

The problem can be simplified further by defining:

$$P(t, f, \alpha; T, K) = \int_{-\infty}^{\infty} A^2 p(t, f, \alpha; T, K, A)\, dA \tag{A.9}$$

Then P satisfies the backward's Kolmogorov equation (Karatzas and Shrene, 1988; Okdendal, 1998; Musiela and Rutkowski; 1998)

$$P_t + \tfrac{1}{2}\varepsilon^2\alpha^2 C^2(f)P_{ff} + \varepsilon^2\rho\nu\alpha^2 C(f)P_{f\alpha} + \tfrac{1}{2}\varepsilon^2\nu^2\alpha^2 P_{\alpha\alpha} = 0 \qquad \text{for } t < T \tag{A.10a}$$

$$P = \alpha^2\delta(f - K), \qquad \text{for } t = T \tag{A.10b}$$

Since t does not appear explicitly in this equation, P depends only on the combination $T - t$, and not on t and T separately. So define:

$$\tau = T - t, \qquad \tau_{ex} = t_{ex} - t \tag{A.11}$$

Then our pricing formula becomes:

$$V(t, f, \alpha) = [f - K]^+ + \frac{1}{2}\varepsilon^2 C^2(K)\int_0^{\tau_{ex}} P(\tau, f, \alpha; K)\, d\tau \tag{A.12}$$

where $P(\tau, f, \alpha; K)$ is the solution of the problem:

$$P_\tau = \tfrac{1}{2}\varepsilon^2\alpha^2 C^2(f)P_{ff} + \varepsilon^2\rho\nu\alpha^2 C(f)P_{f\alpha} + \tfrac{1}{2}\varepsilon^2\nu^2\alpha^2 P_{\alpha\alpha}, \quad \text{for } \tau > 0 \tag{A.13a}$$

$$P = \alpha^2\delta(f - K), \qquad \text{for } \tau = 0 \tag{A.13b}$$

In this Appendix we solve (A.13a), (A.13b) to obtain $P(\tau, f, \alpha; K)$, and then substitute this solution into (A.12) to obtain the option value $V(t, f, \alpha)$. This yields the option price under the SABR model, but the resulting formulas are awkward and not very useful. To cast the results in a more usable form, we re-compute the option price under the normal model:

$$d\hat{F} = \sigma_N\, dW, \tag{A.14a}$$

and then equate the two prices to determine which normal volatility σ_N needs to be used to reproduce the option's price under the SABR model. That is, we find the "implied normal volatility" of the option under the SABR model. By doing a second comparison between option prices under the log normal model:

$$d\hat{F} = \sigma_B\hat{F}\, dW \tag{A.14b}$$

and the normal model, we then convert the implied normal volatility to the usual implied log-normal (Black–Scholes) volatility. That is, we quote the option price predicted by the SABR model in terms of the option's implied volatility.

Singular perturbation expansion

Using a straightforward perturbation expansion would yield a Gaussian density to leading order:

$$P = \frac{\alpha}{\sqrt{2\pi\varepsilon^2 C^2(K)\tau}} e^{-\frac{(f-K)^2}{2\varepsilon^2\alpha^2 C^2(K)\tau}}\{1 + \cdots\}. \tag{A.15a}$$

Since the "$+\cdots$" involves powers of $(f - K)/\varepsilon\alpha C(K)$, this expansion would become inaccurate as soon as $(f - K)C'(K)/C(K)$ becomes a significant fraction of 1; i.e., as soon as $C(f)$

and $C(K)$ are significantly different. Stated differently, small changes in the exponent cause much greater changes in the probability density. A better approach is to re-cast the series as:

$$P = \frac{\alpha}{\sqrt{2\pi\varepsilon^2 C^2(K)\tau}} \frac{(f-K)^2}{e^{2\varepsilon^2\alpha^2 C^2(K)\tau}}\{1+\cdots\} \tag{A.15b}$$

and expand the exponent, since one expects that only small changes to the exponent will be needed to effect the much larger changes in the density. This expansion also describes the basic physics better – P is essentially a Gaussian probability density which tails off faster or slower depending on whether the "diffusion coefficient" $C(f)$ decreases or increases.

We can refine this approach by noting that the exponent is the integral:

$$\frac{(f-K)^2}{2\varepsilon^2\alpha^2 C^2(K)\tau}\{1+\cdots\} = \frac{1}{2\tau}\left(\frac{1}{\varepsilon\alpha}\int_K^f \frac{df'}{C(f')}\right)^2\{1+\cdots\} \tag{A.16}$$

Suppose we define the new variable:

$$z = \frac{1}{\varepsilon\alpha}\int_K^f \frac{df'}{C(f')} \tag{A.17}$$

so that the solution P is essentially $e^{-z^2/2}$. To leading order, the density is Gaussian in the variable z, which is determined by how "easy" or "hard" it is to diffuse from K to f, which closely matches the underlying physics. The fact that the Gaussian changes by orders of magnitude as z^2 increases should be largely irrelevant to the quality of the expansion. This approach is directly related to the geometric optics technique that is so successful in wave propagation and quantum electronics (Kevorkian and Cole, 1985; Whitham, 1974). To be more specific, we shall use the near identity transform method to carry out the geometric optics expansion. This method, pioneered in Neu (1978), transforms the problem order-by-order into a simple canonical problem, which can then be solved trivially. Here we obtain the solution only through $O(\varepsilon^2)$, truncating all higher order terms.

Let us change variables from f to:

$$z = \frac{1}{\varepsilon\alpha}\int_K^f \frac{df'}{C(f')} \tag{A.18a}$$

and to avoid confusion, we define:

$$B(\varepsilon\alpha z) = C(f) \tag{A.18b}$$

Then:

$$\frac{\partial}{\partial f} \longrightarrow \frac{1}{\varepsilon\alpha C(f)}\frac{\partial}{\partial z} = \frac{1}{\varepsilon\alpha B(\varepsilon\alpha z)}\frac{\partial}{\partial z} \qquad \frac{\partial}{\partial\alpha} \longrightarrow \frac{\partial}{\partial\alpha} - \frac{z}{\alpha}\frac{\partial}{\partial z} \tag{A.19a}$$

and:

$$\frac{\partial^2}{\partial f^2} \longrightarrow \frac{1}{\varepsilon^2\alpha^2 B^2(\varepsilon\alpha z)}\left\{\frac{\partial^2}{\partial z^2} - \varepsilon\alpha\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}\frac{\partial}{\partial z}\right\} \tag{A.19b}$$

$$\frac{\partial^2}{\partial f\,\partial\alpha} \longrightarrow \frac{1}{\varepsilon\alpha B(\varepsilon\alpha z)}\left\{\frac{\partial^2}{\partial z\partial\alpha} - \frac{z}{\alpha}\frac{\partial^2}{\partial z^2} - \frac{1}{\alpha}\frac{\partial}{\partial z}\right\} \tag{A.19c}$$

$$\frac{\partial^2}{\partial\alpha^2} \longrightarrow \frac{\partial^2}{\partial\alpha^2} - \frac{2z}{\alpha}\frac{\partial^2}{\partial z\partial\alpha} + \frac{z^2}{\alpha^2}\frac{\partial^2}{\partial z^2} + \frac{2z}{\alpha^2}\frac{\partial}{\partial z} \tag{A.19d}$$

Also:

$$\delta(f - K) = \delta(\varepsilon\alpha z C(K)) = \frac{1}{\varepsilon\alpha C(K)}\delta(z) \tag{A.19e}$$

Therefore, (A.12) through (A.13b) become:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon^2 C^2(K) \int_0^{\tau_{ex}} P, (\tau, z, \alpha)\, d\tau \tag{A.20}$$

where $P(\tau, z, \alpha)$ is the solution of:

$$P_\tau = \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2) P_{zz} - \frac{1}{2}\varepsilon\alpha\frac{B'}{B} P_z + (\varepsilon\rho\nu - \varepsilon^2\nu^2 z)(\alpha P_{ex} - P_z)$$
$$+ \frac{1}{2}\varepsilon^2\nu^2\alpha^2 P_{\alpha a} \qquad \text{for } \tau > 0 \tag{A.21a}$$

$$P = \frac{\alpha}{\varepsilon C(K)}\delta(z) \qquad \text{at } \tau = 0 \tag{A.21b}$$

Accordingly, let us define $\hat{P}(\tau, z, \alpha)$ by:

$$\hat{P} = \frac{\varepsilon}{\alpha} C(K) P \tag{A.22}$$

In terms of $\hat{P}$, we obtain:

$$\mathrm{V}(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha C(K) \int_0^{\tau_{ex}} \hat{P}(\tau, z, \alpha)\, d\tau \tag{A.23}$$

where $\hat{P}(\tau, z, \alpha)$ is the solution of:

$$\hat{P}_\tau = \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2) \hat{P}_{zz} - \frac{1}{2}\varepsilon a\frac{B'}{B} \hat{P}_z + (\varepsilon\rho\nu - \varepsilon^2\nu^2 z)\alpha \hat{P}_{z\alpha}$$
$$+ \frac{1}{2}\varepsilon^2\nu^2(\alpha^2 \hat{P}_{\alpha\alpha} + 2\alpha \hat{P}_\alpha) \qquad \text{for } \tau > 0 \tag{A.24a}$$

$$\hat{P} = \delta(z) \qquad \text{at } \tau = 0. \tag{A.24b}$$

To leading order $\hat{P}$ is the solution of the standard diffusion problem $\hat{P}_\tau = \frac{1}{2}\hat{P}_{zz}$ with $\hat{P} = \delta(z)$ at $\tau = 0$. So it is a Gaussian to leading order. The next stage is to transform the problem to the standard diffusion problem through $O(\varepsilon)$, and then through $O(\varepsilon^2), \ldots$. This is the near identify transform method which has proved so powerful in near-Hamiltonian systems (Neu, 1978).

Note that the variable α does not enter the problem for $\hat{P}$ until $O(\varepsilon)$, so:

$$\hat{P}(\tau, z, \alpha) = \hat{P}_0(\tau, z) + \hat{P}_1(\tau, z, \alpha) + \cdots \tag{A.25}$$

Consequently, the derivatives $\hat{P}_{z\alpha}$, $\hat{P}_{\alpha\alpha}$, and $\hat{P}_{\alpha}$ are all $O(\varepsilon)$. Recall that we are only solving for $\hat{P}$ through $O(\varepsilon^2)$. So, through this order, we can re-write our problem as:

$$\hat{P}_\tau = \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2)\hat{P}_{zz} - \frac{1}{2}\varepsilon a\frac{B'}{B}\hat{P}_z + \varepsilon\rho\nu\alpha\hat{P}_{z\alpha} \qquad \text{for } \tau > 0 \tag{A.26a}$$

$$\hat{P}'' ='' \delta(z) \qquad \text{at } \tau = 0 \tag{A.26b}$$

Let us now eliminate the $\frac{1}{2}\varepsilon a(B'/B)\hat{P}_z$ term. Define $H(\tau, z, \alpha)$ by:

$$\hat{P} = \sqrt{C(f)/C(K)}H \equiv \sqrt{B(\varepsilon\alpha z)/B(0)}H. \tag{A.27}$$

Then:

$$\hat{P}_z = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_z + \frac{1}{2}\varepsilon\alpha\frac{B'}{B}H\right\}; \tag{A.28a}$$

$$\hat{P}_{zz} = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_{zz} + \varepsilon\alpha\frac{B'}{B}H_z + \varepsilon^2\alpha^2\left[\frac{B''}{2B} - \frac{B'^2}{4B^2}\right]H\right\} \tag{A.28b}$$

$$\hat{P}_{z\alpha} = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_{z\alpha} + \frac{1}{2}\varepsilon z\frac{B'}{B}H_z + \frac{1}{2}\varepsilon\alpha\frac{B'}{B}H_\alpha + \frac{1}{2}\varepsilon\frac{B'}{B}H + O(\varepsilon^2)\right\} \tag{A.28c}$$

The option price now becomes:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}\int_0^{\tau_{ex}} H(\tau, z, \alpha)\, d\tau \tag{A.29}$$

where:

$$H_\tau = \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2)H_{zz} - \frac{1}{2}\varepsilon^2\rho\nu\alpha\frac{B'}{B}(zH_z - H)$$
$$+ \varepsilon^2\alpha^2\left(\frac{1}{4}\frac{B''}{B} - \frac{3}{8}\frac{B'^2}{B^2}\right)H + \varepsilon\rho\nu\alpha\left(H_{z\alpha} + \frac{1}{2}\varepsilon\alpha\frac{B'}{B}H_\alpha\right) \qquad \text{for } \tau > 0 \tag{A.30a}$$

$$H = \delta(z) \qquad \text{at } \tau = 0 \tag{A.30b}$$

Equations (A.30a), (A.30b) are independent of α to leading order, and at $O(\varepsilon)$ they depend on α only through the last term $\varepsilon\rho\nu\alpha(H_{z\alpha} + \frac{1}{2}\varepsilon\alpha\frac{B'}{B}; H_\alpha)$. As above, since (A.30a) is independent of α to leading order, we can conclude that the α derivatives H_α and $H_{z\alpha}$ are no larger than $O(\varepsilon)$, and so the last term is actually no larger than $O(\varepsilon^2)$. Therefore, H is independent of α until $O(\varepsilon^2)$ and the α derivatives are actually no larger than $O(\varepsilon^2)$. Thus, the last term is actually

only $O(\varepsilon^3)$, and can be neglected since we are only working through $O(\varepsilon^2)$. So:

$$\begin{aligned} H_\tau &= \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2)H_{zz} - \frac{1}{2}\varepsilon^2\rho\nu\alpha\frac{B'}{B}(zH_z - H) \\ &\quad + \varepsilon^2\alpha^2\left(\frac{1}{4}\frac{B''}{B} - \frac{3}{8}\frac{B'^2}{B^2}\right)H \qquad \text{for } \tau > 0 \end{aligned} \tag{A.31a}$$

$$H = \delta(z) \qquad \text{at } \tau = 0 \tag{A.31b}$$

There are no longer any α derivatives, so we can now treat α as a parameter instead of as an independent variable. That is, we have succeeded in effectively reducing the problem to one dimension.

Let us now remove the H_z term through $O(\varepsilon^2)$. To leading order, $B'(\varepsilon\alpha z)/B(\varepsilon\alpha z)$ and $B''(\varepsilon\alpha z)/B(\varepsilon\alpha z)$ are constant. We can replace these ratios by:

$$b_1 = B'(\varepsilon\alpha z_0)/B(\varepsilon\alpha z_0), \qquad b_2 = B''(\varepsilon\alpha z_0)/B(\varepsilon\alpha z_0) \tag{A.32}$$

committing only an $O(\varepsilon)$ error, where the constant z_0 will be chosen later. We now define $\hat{H}$ by:

$$H = e^{\varepsilon^2\rho\nu\alpha b_1 z^2/4}\hat{H} \tag{A.33}$$

Then our option price becomes:

$$V(t, f, a) = [f - K]^+ + \tfrac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}e^{\varepsilon^2\rho\nu\alpha b_1 z^2/4}\int_0^{\tau_{ex}}\hat{H}(\tau, z)\,d\tau \tag{A.34}$$

where $\hat{H}$ is the solution of:

$$\begin{aligned} \hat{H}_\tau &= \frac{1}{2}(1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2)\hat{H}_{zz} + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)\hat{H} \\ &\quad + \frac{3}{4}\varepsilon^2\rho\nu\alpha b_1\hat{H} \qquad \text{for } \tau > 0 \end{aligned} \tag{A.35a}$$

$$\hat{H} = \delta(z) \qquad \text{at } \tau = 0 \tag{A.35b}$$

We've almost beaten the equation into shape. We now define:

$$\begin{aligned} x &= \frac{1}{\varepsilon\nu}\int_0^{\varepsilon\nu z}\frac{d\zeta}{\sqrt{1 - 2\rho\zeta + \zeta^2}} \\ &= \frac{1}{\varepsilon\nu}\log\left(\frac{\sqrt{1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2} - \rho + \varepsilon\nu z}{1 - \rho}\right), \end{aligned} \tag{A.36a}$$

which can be written implicitly as:

$$\varepsilon\nu z = \sinh\varepsilon\nu x - \rho(\cosh\varepsilon\nu x - 1). \tag{A.36b}$$

In terms of x, our problem is:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}e^{\varepsilon^2\rho\nu\alpha b_1 z^2/4}\int_0^{\tau_{ex}} \hat{H}(\tau, x)\, d\tau \tag{A.37}$$

with:

$$\begin{aligned}\hat{H}_\tau = {} & \frac{1}{2}\hat{H}_{xx} - \frac{1}{2}\varepsilon\nu I'(\varepsilon\nu z)\hat{H}_x + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)\hat{H} \\ & + \frac{3}{4}\varepsilon^2\rho\nu\alpha b_1\hat{H} \qquad \text{for } \tau > 0\end{aligned} \tag{A.38a}$$

$$\hat{H} = \delta(x) \qquad \text{at } \tau = 0 \tag{A.38b}$$

Here:

$$I(\zeta) = \sqrt{1 - 2\rho\zeta + \zeta^2} \tag{A.39}$$

The final step is to define Q by:

$$\hat{H} = I^{1/2}(\varepsilon\nu z(x))Q = (1 - 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2)^{1/4}Q \tag{A.40}$$

Then:

$$\hat{H}_x = I^{1/2}(\varepsilon\nu z)\left[Q_x + \frac{1}{2}\varepsilon\nu I'(\varepsilon\nu z)Q\right] \tag{A.41a}$$

$$\hat{H}_{xx} = I^{1/2}(\varepsilon\nu z)\left[Q_{xx} + \varepsilon\nu I' Q_x + \varepsilon^2\nu^2\left(\frac{1}{2}I''I + \frac{1}{4}I'I'\right)Q\right] \tag{A.41b}$$

and so:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}I^{1/2}(\varepsilon\nu z)e^{\frac{1}{4}\varepsilon^2\rho\nu\alpha b_1 z^2}\int_0^{\tau_{ex}} Q, (\tau, x)\, d\tau \tag{A.42}$$

where Q is the solution of:

$$Q_\tau = \frac{1}{2}Q_{xx} + \varepsilon^2\nu^2\left(\frac{1}{4}I''I - \frac{1}{8}I'I'\right)Q + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)Q + \frac{3}{4}\varepsilon^2\rho\nu\alpha b_1 Q \tag{A.43a}$$

for $\tau > 0$, with:

$$Q = \delta(x) \qquad \text{at } \tau = 0 \tag{A.43b}$$

As above, we can replace $I(\varepsilon\nu z)$, $I'(\varepsilon\nu z)$, $I''(\varepsilon\nu z)$ by the constants $I(\varepsilon\nu z_0)$, $I'(\varepsilon\nu z_0)$, $I''(\varepsilon\nu z_0)$, and commit only $O(\varepsilon)$ errors. Define the constant κ by:

$$\begin{aligned}\kappa = {} & \nu^2\left(\frac{1}{4}I''(\varepsilon\nu z_0)I(\varepsilon\nu z_0) - \frac{1}{8}[I'(\varepsilon\nu z_0)]^2\right) \\ & + \alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right) + \frac{3}{4}\rho\nu\alpha b_1\end{aligned} \tag{A.44}$$

where z_0 will be chosen later. Then through $O(\varepsilon^2)$, we can simplify our equation to:

$$Q_\tau = \frac{1}{2}Q_{xx} + \varepsilon^2 \kappa Q \qquad \text{for } \tau > 0 \tag{A.45a}$$

$$Q = \delta(x) \qquad \text{at } \tau = 0 \tag{A.45b}$$

The solution of (A.45a, A.45b) is clearly:

$$Q = \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau} e^{\varepsilon^2\kappa\tau} = \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau} \frac{1}{\left(1 - \frac{2}{3}\kappa\varepsilon^2\tau + \cdots\right)^{3/2}} \tag{A.46}$$

through $O(\varepsilon^2)$.

This solution yields the option price:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}\, I^{1/2}(\varepsilon\nu z) e^{\frac{1}{4}\varepsilon^2\rho\nu\alpha b_1 z^2} \cdot \int_0^{\tau_{ex}} \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau} e^{\varepsilon^2\kappa\tau}\, d\tau \tag{A.47}$$

Observe that this can be written as:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\frac{f - K}{x} \int_0^{\tau_{ex}} \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau} e^{\varepsilon^2\theta} e^{\varepsilon^2\kappa\tau}\, d\tau \tag{A.48a}$$

where:

$$\varepsilon^2\theta = \log\left(\frac{\varepsilon\alpha z}{f - K}\sqrt{B(0)B(\varepsilon\alpha z)}\right) + \log\left(\frac{x I^{1/2}(\varepsilon\nu z)}{z}\right) + \frac{1}{4}\varepsilon^2\rho\nu\alpha b_1 z^2 \tag{A.48b}$$

Moreover, quite amazingly:

$$\mathrm{e}^{\varepsilon^2\kappa\tau} = \frac{1}{\left(1 - \frac{2}{3}\kappa\varepsilon^2\tau\right)^{3/2}} = \frac{1}{\left(1 - 2\varepsilon^2\tau\frac{\theta}{x^2}\right)^{3/2}} + O(\varepsilon^4) \tag{A.48c}$$

through $O(\varepsilon^2)$. This can be shown by expanding $\varepsilon^2\theta$ through $O(\varepsilon^2)$, and noting that $\varepsilon^2\theta/x^2$ matches $\kappa/3$. Therefore our option price is:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\frac{f - K}{x} \int_0^{\tau_{ex}} \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau} e^{\varepsilon^2\theta} \frac{d\tau}{\left(1 - \frac{2\tau}{x^2}\varepsilon^2\theta\right)^{3/2}} \tag{A.49}$$

and changing integration variables to:

$$q = \frac{x^2}{2\tau} \tag{A.50}$$

reduces this to:

$$V(t,\, f,\, a) = [f - K]^+ + \frac{|f - K|}{4\sqrt{\pi}} \int_{\frac{x^2}{2\tau_{ex}}}^{\infty} \frac{e^{-q+\varepsilon^2\theta}}{(q - \varepsilon^2\theta)^{3/2}}\, dq \tag{A.51}$$

That is, the value of a European call option is given by:

$$V(t,\, f,\, a) = [f - K]^+ + \frac{|f - K|}{4\sqrt{\pi}} \int_{\frac{x^2}{2\tau_{ex}} - \varepsilon^2\theta}^{\infty} \frac{e-q}{q^{3/2}}\, dq \tag{A.52a}$$

with:

$$\varepsilon^2\theta = \log\left(\frac{\varepsilon\alpha z}{f - K}\sqrt{B(0)B(\varepsilon\alpha z)}\right) + \log\left(\frac{xI^{1/2}(\varepsilon\nu z)}{z}\right) + \frac{1}{4}\varepsilon^2\rho\nu\alpha b_1 z^2 \tag{A.52b}$$

through $O(\varepsilon^2)$.

Equivalent normal volatility

Equations (A.52a) and (A.52b) are a formula for the dollar price of the call option under the SABR model. The utility and beauty of this formula is not overwhelmingly apparent. To obtain a useful formula, we convert this dollar price into the equivalent implied volatilities. We first obtain the implied *normal* volatility, and then the standard *log normal* (Black) volatility.

Suppose we repeated the above analysis for the ordinary *normal model*:

$$d\hat{F} = \sigma_N\, dW, \qquad \hat{F}(0) = f \tag{A.53a}$$

where the normal volatility σ_N is constant, not stochastic. (This model is also called the *absolute* or *Gaussian* model). We would find that the option value for the normal model is exactly:

$$V(t,\, f) = [f - K]^+ + \frac{|f - K|}{4\sqrt{\pi}} \int_{\frac{(f-K)^2}{2\sigma_N^2\tau_{ex}}}^{\infty} \frac{e^{-q}}{q^{3/2}}\, dq \tag{A.53b}$$

This can be seen by setting $C(f)$ to 1, setting $\varepsilon\alpha$ to σ_N and setting ν to 0 in (A.52a, A.52b). Working out this integral then yields the exact European option price:

$$V(t,\, f) = (f - K)\mathcal{N}\left(\frac{f - K}{\sigma_N\sqrt{\tau_{ex}}}\right) + \sigma_N\sqrt{\tau_{ex}}\,\mathcal{G}\left(\frac{f - K}{\sigma_N\sqrt{\tau_{ex}}}\right) \tag{A.54a}$$

for the normal model, where $\mathcal{N}$ is the normal distribution and $\mathcal{G}$ is the Gaussian density:

$$\mathcal{G}(q) = \frac{1}{\sqrt{2\pi}} e^{-q^2/2} \tag{A.54b}$$

From (A.53b) it is clear that the option price under the normal model matches the option price under the SABR model (A.52a, A.52b) if and only if we choose the normal volatility σ_N to be

$$\frac{1}{\sigma_N^2} = \frac{x^2}{(f - K)^2}\left\{1 - 2\varepsilon^2\frac{\theta}{x^2}\tau_{ex}\right\} \tag{A.55}$$

Taking the square root now shows the option's *implied normal* (absolute) *volatility* is given by:

$$\sigma_N = \frac{f-K}{x}\left\{1+\varepsilon^2\frac{\theta}{x^2}\tau_{ex}+\cdots\right\} \tag{A.56}$$

through $O(\varepsilon^2)$.

Before continuing to the implied *log normal* volatility, let us seek the simplest possible way to re-write this answer which is correct through $O(\varepsilon^2)$. Since $x = z[1 + O(\varepsilon)]$, we can re-write the answer as:

$$\sigma_N = \left(\frac{f-K}{z}\right)\left(\frac{z}{x(z)}\right)\{1+\varepsilon^2\ (\phi_1+\phi_2+\phi_3)\ \tau_{ex}+\cdots\} \tag{A.57a}$$

where

$$\frac{f-K}{z} = \frac{\varepsilon\alpha(f-K)}{\int_K^f \frac{df'}{C(f')}} = \left(\frac{1}{f-K}\int_K^f \frac{df'}{\varepsilon\alpha C(f')}\right)^{-1}$$

This factor represents the average difficulty in diffusing from today's forward f to the strike K, and would be present even if the volatility were not stochastic.

The next factor is:

$$\frac{z}{x(z)} = \frac{\zeta}{\log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right)} \tag{A.57b}$$

where

$$\zeta = \varepsilon\nu z = \frac{\nu}{\alpha}\int_K^f \frac{df'}{C(f')} = \frac{\nu}{\alpha}\frac{f-K}{C(f_{av})}\{1+O(\varepsilon^2)\} \tag{A.57c}$$

Here $f_{av} = \sqrt{fK}$ is the geometric average of f and K. (The arithmetic average could have been used equally well at this order of accuracy). This factor represents the main effect of the stochastic volatility.

The coefficients ϕ_1, ϕ_2, and ϕ_3 provide relatively minor corrections. Through $O(\varepsilon^2)$ these corrections are:

$$\begin{aligned}\varepsilon^2\phi_1 &= \frac{1}{z^2}\log\left(\frac{\varepsilon\alpha z}{f-K}\sqrt{C(f)\ C(K)}\right)\\ &= \frac{2\gamma_2-\gamma_1^2}{24}\varepsilon^2\alpha^2C^2\ (f_{av})+\cdots\end{aligned} \tag{A.57d}$$

$$\varepsilon^2\phi_2 = \frac{1}{z^2}\log\left(\frac{x}{z}[1-2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2]^{1/4}\right) = \frac{2-3\rho^2}{24}\varepsilon^2\nu^2+\cdots \tag{A.57e}$$

$$\varepsilon^2\phi_3 = \frac{1}{4}\varepsilon^2\rho\alpha\nu\frac{B'(\varepsilon\nu z_0)}{B(\varepsilon\nu z_0)} = \frac{1}{4}\varepsilon^2\rho\nu\alpha\gamma_1C(f_{av})+\cdots \tag{A.57f}$$

where

$$\gamma_1 = \frac{C'(f_{av})}{C(f_{av})}, \qquad \gamma_2 = \frac{C''(f_{av})}{C(f_{av})} \tag{A.57g}$$

Let us briefly summarize before continuing. Under the *normal model*, the value of a European call option with strike K and exercise date τ_{ex} is given by (A.54a), (A.54b). For the SABR model:

$$d\hat{F} = \varepsilon\hat{\alpha}C(\hat{F})\,dW_1 \qquad \hat{F}(0) = f \tag{A.58a}$$

$$d\hat{\alpha} = \varepsilon\nu\hat{\alpha}\,dW_2 \qquad \hat{\alpha}(0) = \alpha \tag{A.58b}$$

$$dW_1\,dW_2 = \rho\,dt \tag{A.58c}$$

the value of the call option is given by the same formula, at least through $O(\varepsilon^2)$, provided we use the implied normal volatility:

$$\sigma_N(K) = \frac{\varepsilon\alpha(f-K)}{\displaystyle\int_K^f \frac{df'}{C(f')}} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{2\gamma_2 - \gamma_1^2}{24}\alpha^2 C^2(f_{av}) + \frac{1}{4}\rho\nu\alpha\gamma_1 C(f_{av}) + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.59a}$$

Here:

$$f_{av} = \sqrt{fK}, \qquad \gamma_1 = \frac{C'(f_{av})}{C(f_{av})}, \qquad \gamma_2 = \frac{C''(f_{av})}{C(f_{av})}, \tag{A.59b}$$

$$\zeta = \frac{\nu}{\alpha}\frac{f-K}{C(f_{av})}, \qquad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right). \tag{A.59c}$$

The first two factors provide the dominant behavior, with the remaining factor $1 + [\cdots]\varepsilon^2\tau_{ex}$ usually providing corrections of around 1% or so.

One can repeat the analysis for a European put option, or simply use call/put parity. This shows that the value of the put option under the SABR model is:

$$V_{put}(f,\alpha,K) = (K-f)\mathcal{N}\left(\frac{K-f}{\sigma_N\sqrt{\tau_{ex}}}\right) + \sigma_N\sqrt{\tau_{ex}}\,\mathcal{G}\left(\frac{K-f}{\sigma_N\sqrt{\tau_{ex}}}\right) \tag{A.60}$$

where the implied normal volatility σ_N is given by the same formulas (A.59a–A.59c) as the call.

We can revert to the original units by replacing $\varepsilon\alpha \longrightarrow \alpha, \varepsilon\nu \longrightarrow \nu$ everywhere in the above formulas; this is equivalent to setting ε to 1 everywhere.

Equivalent Black volatility

With the exception of JPY traders, most traders prefer to quote prices in terms of Black (log normal) volatilities, rather than normal volatilities. To derive the implied Black volatility, consider Black's model:

$$d\hat{F} = \varepsilon\sigma_B\hat{F}\,dW, \qquad \hat{F}(0) = f, \tag{A.61}$$

where we have written the volatility as $\varepsilon\sigma_B$ to stay consistent with the preceding analysis. For Black's model, the value of a European call with strike K and exercise date τ_{ex} is:

$$V_{call} = f\mathcal{N}(d_1) - K\mathcal{N}(d_2) \tag{A.62a}$$

$$V_{put} = V_{call} + D(t_{set})[K - f] \tag{A.62b}$$

with:

$$d_{1,2} = \frac{\log f/K \pm \frac{1}{2}\varepsilon^2\sigma_B^2\tau_{ex}}{\varepsilon\sigma_B\sqrt{\tau_{ex}}} \tag{A.62c}$$

where we are omitting the overall factor $D(t_{set})$ as before.

We can obtain the implied normal volatility for Black's model by repeating the preceding analysis for the SABR model with $C(f) = f$ and $\nu = 0$. Setting $C(f) = f$ and $\nu = 0$ in (A.59a–A.59c) shows that the normal volatility is:

$$\sigma_N(K) = \frac{\varepsilon\sigma_B(f - K)}{\log f/K}\left\{1 - \frac{1}{24}\varepsilon^2\sigma_B^2\tau_{ex} + \cdots\right\} \tag{A.63}$$

through $O(\varepsilon^2)$. Indeed, in Hagan *et al.* (in preparation) it is shown that the implied normal volatility for Black's model is:

$$\sigma_N(K) = \varepsilon\sigma_B\sqrt{fK}\frac{1 + \frac{1}{24}\log^2 f/K + \frac{1}{1920}\log^4 f/K + \cdots}{1 + \frac{1}{24}\left(1 - \frac{1}{120}\log^2 f/K\right)\varepsilon^2\sigma_B^2\tau_{ex} + \frac{1}{5760}\varepsilon^4\sigma_B^4\tau_{ex}^2 + \cdots} \tag{A.64}$$

through $O(\varepsilon^4)$. We can find the implied Black vol for the SABR model by setting σ_N obtained from Black's model in equation (A.63) equal to σ_N obtained from the SABR model in (A.59a–B.59c). Through $O(\varepsilon^2)$ this yields:

$$\begin{aligned}\sigma_B(K) = {} & \frac{\alpha\log f/K}{\int_K^f \frac{df'}{C(f')}} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right)\left\{1 + \left[\frac{2\gamma_2 - \gamma_1^2 + 1/f_{av}^2}{24}\alpha^2C^2(f_{av}) + \frac{1}{4}\rho\nu\alpha\gamma_1C(f_{av})\right.\right. \\ & \left.\left. + \frac{2 - 3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\}\end{aligned} \tag{A.65}$$

This is the main result of this article. As before, the implied log normal volatility for puts is the same as for calls, and this formula can be re-cast in terms of the original variables by simply setting ε to 1.

Stochastic β model

As originally stated, the SABR model consists of the special case $C(f) = f^\beta$:

$$d\hat{F} = \varepsilon\hat{\alpha}\hat{F}^\beta\, dW_1 \qquad \hat{F}(0) = f \tag{A.66a}$$

$$d\hat{\alpha} = \varepsilon\nu\hat{\alpha}\, dW_2 \qquad \hat{\alpha}(0) = \alpha \tag{A.66b}$$

$$dW_1\, dW_2 = \rho\, dt \tag{A.66c}$$

Making this substitution in (A.58a–A.58b) shows that the implied *normal* volatility for this model is:

$$\sigma_N(K) = \frac{\varepsilon\alpha(1-\beta)(f-K)}{f^{1-\beta}-K^{1-\beta}} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{-\beta(2-\beta)\alpha^2}{24 f_{av}^{2-2\beta}} + \frac{\rho\alpha\nu\beta}{4 f_{av}^{1-\beta}} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.67a}$$

through $O(\varepsilon^2)$, where $f_{av} = \sqrt{fK}$ as before and:

$$\zeta = \frac{\nu}{\alpha}\frac{f-K}{f_{av}^{\beta}}, \qquad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right). \tag{A.67b}$$

We can simplify this formula by expanding:

$$f - K = \sqrt{fK}\log f/K\left\{1 + \frac{1}{24}\log^2 f/K + \frac{1}{1920}\log^4 f/K + \cdots\right. \tag{A.68a}$$

$$f^{1-\beta} - K^{1-\beta} = (1-\beta)(fK^{(1-\beta)/2})\log f/K \cdot \left\{1 + \frac{(1-\beta)^2}{24}\log^2 f/K + \frac{(1-\beta)^4}{1920}\log^4 f/K + \cdots\right. \tag{A.68b}$$

and neglecting terms higher than fourth order. This expansion reduces the implied normal volatility to:

$$\sigma_N(K) = \varepsilon\alpha(fK)^{\beta/2}\frac{1 + \frac{1}{24}\log^2 f/K + \frac{1}{1920}\log^4 f/K + \cdots}{1 + \frac{(1-\beta)^2}{24}\log^2 f/K + \frac{(1-\beta)^4}{1920}\log^4 f/K + \cdots} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{-\beta(2-\beta)\alpha^2}{24(fK)^{1-\beta}} + \frac{\rho\alpha\nu\beta}{4(fK)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.69a}$$

where:

$$\zeta = \frac{\nu}{\alpha}(fK)^{(1-\beta)/2}\log f/K, \qquad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right) \tag{A.69b}$$

This is the formula we use in pricing European calls and puts.

To obtain the implied Black volatility, we equate the implied normal volatility $\sigma_N(K)$ for the SABR model obtained in (A.69a–A.69b) to the implied normal volatility for Black's model

obtained in (A.63). This shows that the implied Black volatility for the SABR model is:

$$\sigma_B(K) = \frac{\varepsilon\alpha}{(fK)^{(1-\beta)/2}} \frac{1}{1 + \frac{(1-\beta)^2}{24}\log^2 f/K + \frac{(1-\beta)^4}{1920}\log^4 f/K + \cdots} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{(1-\beta)^2\alpha^2}{24(fK)^{1-\beta}} + \frac{\rho\alpha\nu\beta}{4(fK)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.69c}$$

through $O(\varepsilon^2)$, where ζ and $\hat{x}(\zeta)$ are given by (A.69b) as before. Apart from setting ε to 1 to recover the original units, this is the formula quoted in the Reprise section, and fitted to the market in the section on Managing smile risk.

Special cases

Two special cases are worthy of special treatment: the stochastic normal model ($\beta = 0$) and the stochastic log normal model ($\beta = 1$). Both these models are simple enough that the expansion can be continued through $O(\varepsilon^4)$. For the stochastic *normal* model ($\beta = 0$) the implied volatilities of European calls and puts are:

$$\sigma_N(K) = \varepsilon\alpha\left(\frac{\zeta}{\hat{x}(\zeta)}\right)\left\{1 + \frac{2-3\rho^2}{24}\varepsilon^2\nu^2\tau_{ex} + \cdots\right\} \tag{A.70a}$$

$$\sigma_B(K) = \varepsilon\alpha\frac{\log f/K}{f-K}\cdot\left(\frac{\zeta}{\hat{x}(\zeta)}\right)\cdot\left\{1 + \left[\frac{\alpha^2}{24fK} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.70b}$$

through $O(\varepsilon^4)$, where:

$$\zeta = \frac{\nu}{\alpha}\sqrt{fK}\log f/K, \qquad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right) \tag{A.70c}$$

For the stochastic log normal model ($\beta = 1$) the implied volatilities are:

$$\sigma_N(K) = \varepsilon\alpha\frac{f-K}{\log f/K}\cdot\left(\frac{\zeta}{\hat{x}(\zeta)}\right)\cdot\left\{1 + \left[-\frac{1}{24}\alpha^2 + \frac{1}{4}\rho\alpha\nu + \frac{1}{24}(2-3\rho^2)\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.71a}$$

$$\sigma_B(K) = \varepsilon\alpha\cdot\left(\frac{\zeta}{\hat{x}(\zeta)}\right)\cdot\left\{1 + \left[\frac{1}{4}\rho\alpha\nu + \frac{1}{24}(2-3\rho^2)\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\} \tag{A.71b}$$

through $O(\varepsilon^4)$, where:

$$\zeta = \frac{\nu}{\alpha}\log f/K, \qquad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right) \tag{A.71c}$$

Appendix B. Analysis of the dynamic SABR model

We use *effective medium theory* (Clouet, 1998) to extend the preceding analysis to the dynamic SABR model. As before, we take the volatility $\gamma(t)\hat{\alpha}$ and "volvol" $\nu(t)$ to be small, writing $\gamma(t) \longrightarrow \varepsilon\gamma(t)$, and $\nu(t) \longrightarrow \varepsilon\nu(t)$, and analyze:

$$d\hat{F} = \varepsilon\gamma(t)\hat{\alpha}C(\hat{F})\,dW_1 \tag{B.1a}$$

$$d\hat{\alpha} = \varepsilon\nu(t)\hat{\alpha}\,dW_2 \tag{B.1b}$$

with:

$$dW_1\,dW_2 = \rho(t)\,dt \tag{B.1c}$$

in the limit $\varepsilon \ll 1$. We obtain the prices of European options, and from these prices we obtain the implied volatility of these options. After obtaining the results, we replace $\varepsilon\gamma(t) \longrightarrow \gamma(t)$ and $\varepsilon\nu(t) \longrightarrow \nu(t)$ to get the answer in terms of the original variables.

Suppose the economy is in state $\hat{F}(t) = f, \hat{\alpha}(t) = \alpha$ at date t. Let $V(t, f, \alpha)$ be the value of, say, a European call option with strike K and exercise date t_{ex}. As before, define the transition density $p(t, f, \alpha; T, F, A)$ by:

$$p(t, f, \alpha; T, F, A)\,dF\,dA \equiv \text{prob}\Big\{F < \hat{F}(T) < F + dF,\ A < \hat{\alpha}(T) < A + dA | \hat{F}(t) = f, \hat{\alpha}(t) = \alpha\Big\} \tag{B.2a}$$

and define:

$$P(t, f, \alpha; T, K) = \int_{-\infty}^{\infty} A^2 p(t, f, \alpha; T, K, A)\,dA \tag{B.2b}$$

Repeating the analysis in Appendix B through equations (A.10a), (A.10b) now shows that the option price is given by:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon^2 C^2(K) \int_t^{t_{ex}} \gamma^2(T) P(t, f, \alpha; T, K)\,dT \tag{B.3}$$

where $P(t, f, \alpha; T, K)$ is the solution of the backwards problem:

$$P_t + \tfrac{1}{2}\varepsilon^2 \left\{\gamma^2\alpha^2 C^2(f) P_{ff} + 2\rho\gamma\nu\alpha^2 C(f) P_{f\alpha} + \nu^2\alpha^2 P_{\alpha\alpha}\right\} = 0 \qquad \text{for } t < T \tag{B.4a}$$

$$P = \alpha^2\delta(f - K), \text{ for } t = T \tag{B.4b}$$

We eliminate $\gamma(t)$ by defining the new time variable:

$$s = \int_0^t \gamma^2(t')\,dt', \quad s' = \int_0^T \gamma^2(t')\,dt', \quad s_{ex} = \int_0^{t_{ex}} \gamma^2(t')\,dt' \tag{B.5}$$

Then the option price becomes:

$$V(t,\, f, a) = [f - K]^{+} + \frac{1}{2}\varepsilon^2 C^2(K) \int_s^{s_{ex}} P(s,\, f, \alpha; s',\, K)\, ds' \tag{B.6}$$

where $P(s,\, f, \alpha; s',\, K)$ solves the backward problem:

$$P_s + \tfrac{1}{2}\varepsilon^2 \left\{\alpha^2 C^2(f) P_{ff} + 2\eta(s)\alpha^2 C(f) P_{f\alpha} + \upsilon^2(s)\alpha^2 P_{\alpha\alpha}\right\} = 0 \qquad \text{for } s < s' \tag{B.7a}$$

$$P = \alpha^2 \delta(f - K) \text{ for } s = s' \tag{B.7b}$$

Here:

$$\eta(s) = \rho(t)\nu(t)/\gamma(t), \qquad \upsilon(s) = \nu(t)/\gamma(t) \tag{B.8}$$

We solve this problem by using an *effective media* strategy (Clouet, 1998). In this strategy our objective is to determine which constant values $\overline{\eta}$ and $\overline{\upsilon}$ yield the same option price as the time dependent coefficients $\eta(s)$ and $\upsilon(s)$. If we could find these constant values, this would reduce the problem to the non-dynamic SABR model solved in Appendix A.

We carry out this strategy by applying the same series of *time-independent* transformations that was used to solve the non-dynamic SABR model in Appendix A, defining the transformations in terms of the (as yet unknown) constants $\overline{\eta}$ and $\overline{\upsilon}$. The resulting problem is relatively complex, more complex than the canonical problem obtained in Appendix A. We use a regular perturbation expansion to solve this problem, and once we have solved this problem, we choose $\overline{\eta}$ and $\overline{\upsilon}$ so that all terms arising from the time dependence of $\eta(t)$ and $\upsilon(t)$ cancel out. As we shall see, this simultaneously determines the "effective" parameters and allows us to use the analysis in Appendix A to obtain the implied volatility of the option.

Transformation

As in Appendix A, we change independent variables to:

$$z = \frac{1}{\varepsilon\alpha} \int_K^f \frac{df'}{C(f')} \tag{B.9a}$$

and define:

$$B(\varepsilon\alpha z) = C(f) \tag{B.9b}$$

We then change dependent variables from P to $\hat{P}$, and then to H:

$$\hat{P} = \frac{\varepsilon}{\alpha} C(K) P \tag{B.9c}$$

$$H = \sqrt{C(K)/C(f)}\,\hat{P} \equiv \sqrt{B(0)/B(\varepsilon\alpha z)}\,\hat{P} \tag{B.9d}$$

Following the reasoning in Appendix A, we obtain:

$$V(t,\, f, a) = [f - K]^{+} + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)} \int_s^{s_{ex}} H(s, z, \alpha; s')\, ds', \tag{B.10}$$

where $H(s, z, \alpha; s')$ is the solution of:

$$H_s + \frac{1}{2}(1 - 2\varepsilon\eta z + \varepsilon^2\upsilon^2 z^2)H_{zz} - \frac{1}{2}\varepsilon^2\eta\alpha\frac{B'}{B}(zH_z - H) + \varepsilon^2\alpha^2\left(\frac{1}{4}\frac{B''}{B} - \frac{3}{8}\frac{B'^2}{B^2}\right)H = 0 \tag{B.11a}$$

for $s < s'$, and:

$$H = \delta(z) \quad \text{at } s = s' \tag{B.11b}$$

through $O(\varepsilon^2)$. See (A.29), (A.31a), and (A.31b). There are no α derivatives in equations (B.11a), (B.11b), so we can treat α as a parameter instead of a variable. Through $O(\varepsilon^2)$ we can also treat B'/B and B''/B as constants:

$$b_1 \equiv \frac{B'(\varepsilon\alpha z_0)}{B(\varepsilon\alpha z_0)}, \qquad b_2 \equiv \frac{B''(\varepsilon\alpha z_0)}{B(\varepsilon\alpha z_0)} \tag{B.12}$$

where z_0 will be chosen later. Thus we must solve:

$$H_s + \frac{1}{2}(1 - 2\varepsilon\eta z + \varepsilon^2\upsilon^2 z^2)H_{zz} - \frac{1}{2}\varepsilon^2\eta\alpha b_1(zH_z - H) + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)H = 0 \quad \text{for } s < s' \tag{B.13a}$$

$$H = \delta(z) \quad \text{at } s = s' \tag{B.13b}$$

At this point we would like to use a time-independent transformation to remove the zH_z term from equation (B.13a). It is not possible to cancel this term exactly, since the coefficient $\eta(s)$ is time dependent. Instead we use the transformation:

$$H = e^{\frac{1}{4}\varepsilon^2\alpha b_1\delta z^2}\hat{H} \tag{B.14}$$

where the constant δ will be chosen later. This transformation yields:

$$\hat{H}_s + \frac{1}{2}(1 - 2\varepsilon\eta z + \varepsilon^2\upsilon^2 z^2)\hat{H}_{zz} - \frac{1}{2}\varepsilon^2\alpha b_1(\eta - \delta)z\hat{H}_z + \frac{1}{4}\varepsilon^2\alpha b_1(2\eta + \delta)\hat{H} + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)\hat{H} = 0 \quad \text{for } s < s'$$

$$\hat{H} = \delta(z) \quad \text{at } s = \mathrm{s}' \tag{B.15a}$$

through $O(\varepsilon^2)$. Later the constant δ will be selected so that the change in the option price caused by the term $\frac{1}{2}\varepsilon^2\alpha b_1\eta z\hat{H}_z$ is exactly offset by the change in price due to $\frac{1}{2}\varepsilon^2\alpha b_1\delta z\hat{H}_z$ term. In this way to the transformation cancels out the zH_z term "on average".

In a similar vein we define:

$$I(\varepsilon\overline{\upsilon}z) = \sqrt{1 - 2\varepsilon\overline{\eta}z + \varepsilon^2\overline{\upsilon}^2 z^2} \tag{B.16a}$$

and:

$$x = \frac{1}{\varepsilon\overline{\upsilon}}\int_0^{\varepsilon\overline{\upsilon}z}\frac{d\zeta}{I(\zeta)} = \frac{1}{\varepsilon\overline{\upsilon}}\log\left(\frac{\sqrt{1 - 2\varepsilon\overline{\eta}z + \varepsilon^2\overline{\upsilon}^2 z^2} - \overline{\eta}/\overline{\upsilon} + \varepsilon\overline{\upsilon}z}{1 - \overline{\eta}/\overline{\upsilon}}\right) \tag{B.16b}$$

where the constants $\overline{\eta}$ and $\overline{\upsilon}$ will be chosen later. This yields:

$$\hat{H}_s + \frac{1}{2}\left(\frac{1 - 2\varepsilon\eta z + \varepsilon^2\upsilon^2 z^2}{1 - 2\varepsilon\overline{\eta}z + \varepsilon^2\overline{\upsilon}^2 z^2}\right)\left(\hat{H}_{xx} - \varepsilon\overline{\upsilon}I'(\varepsilon\overline{\upsilon}z)\hat{H}_x\right) - \frac{1}{2}\varepsilon^2\alpha b_1(\eta - \delta)x\hat{H}_x$$
$$+ \frac{1}{4}\varepsilon^2\alpha b_1(2\eta + \delta)\hat{H} + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)\hat{H} = 0 \quad \text{for } s < s' \tag{B.17a}$$

$$\hat{H} = \delta(x) \quad \text{at } s = s' \tag{B.17b}$$

through $O(\varepsilon^2)$. Here we used $z = x + \cdots$ and $z\hat{H}_z = x\hat{H}_x + \cdots$ to leading order to simplify the results. Finally, we define:

$$\hat{H} = I^{1/2}(\varepsilon\overline{\upsilon}z)Q. \tag{B.18}$$

Then the price of our call option is:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}I^{1/2}$$
$$\cdot (\varepsilon\overline{\upsilon z})e^{\frac{1}{4}\varepsilon^2\alpha b_1\delta z^2}\int_s^{s_{ex}} Q(s, x, s')\,ds' \tag{B.19}$$

where $Q(s, x; s')$ is the solution of:

$$Q + \frac{1}{2}\left(\frac{1 - 2\varepsilon\eta z + \varepsilon^2\upsilon^2 z^2}{1 - 2\varepsilon\overline{\eta}z + \varepsilon^2\overline{\upsilon}^2 z^2}Q_{xx} - \frac{1}{2}\varepsilon^2\alpha b_1(\eta - \delta)xQ_x + \frac{1}{4}\varepsilon^2\alpha b_1(2\eta + \delta)Q\right.$$
$$\left.+\varepsilon^2\upsilon^{-2}\right)\left(\frac{1}{4}I''I - \frac{1}{8}I'I'\right)Q + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)Q = 0 \quad \text{for } s < s' \tag{B.20a}$$

$$Q = \delta(x) \quad \text{at } s = s' \tag{B.20b}$$

Using:

$$z = x - \tfrac{1}{2}\varepsilon\overline{\eta}x^2 + \cdots \tag{B.21}$$

we can simplify this to:

$$Q_s + \frac{1}{2}Q_{xx} = \varepsilon(\eta - \overline{\eta})xQ_{xx} - \frac{1}{2}\varepsilon^2[\upsilon^2 - \overline{\upsilon}^2 - 3\overline{\eta}(\eta - \overline{\eta})]x^2Q_{xx}$$
$$+ \frac{1}{2}\varepsilon^2\alpha b_1(\eta - \delta)(xQ_x - Q)$$

$$-\frac{3}{4}\varepsilon^2\alpha b_1\delta Q-\varepsilon^2\overline{\upsilon}^2\left(\frac{1}{4}I''I-\frac{1}{8}I'I'\right)Q$$

$$-\varepsilon^2\alpha^2\left(\frac{1}{4}b_2-\frac{3}{8}b_1^2\right)Q \quad \text{for } s<s' \tag{B.22a}$$

$$Q=\delta(x) \quad \text{at } s=s' \tag{B.22b}$$

through $O(\varepsilon^2)$. Note that I, I', and I'' can be replaced by the constants $I(\varepsilon\overline{\upsilon}z_0)$, $I'(\varepsilon\overline{\upsilon}z_0)$, and $I''(\varepsilon\overline{\upsilon}z_0)$ through $O(\varepsilon^2)$.

Perturbation expansion

Suppose we were to expand $Q(s,x;s')$ as a power series in ε:

$$Q(s,x;s')=Q^{(0)}(s,x;s')+\varepsilon Q^{(1)}(s,x;s')+\varepsilon^2Q^{(2)}(s,x;s')+\cdots \tag{B.23}$$

Substituting this expansion into (B.22a), (B.22b) yields the following hierarchy of equations. To leading order we have:

$$Q_s^{(0)}+\frac{1}{2}Q_{xx}^{(0)}=0 \quad \text{for } s<s' \tag{B.24a}$$

$$Q^{(0)}=\delta(x) \quad \text{at } s=s' \tag{B.24a}$$

At $O(\varepsilon)$ we have:

$$Q_s^{(1)}+\tfrac{1}{2}Q_{xx}^{(1)}=(\eta-\overline{\eta})xQ_{xx}^{(0)} \quad \text{for } s<s' \tag{B.25a}$$

$$Q^{(1)}=0 \quad \text{at } s=s' \tag{B.25b}$$

At $O(\varepsilon^2)$ we can break the solution into:

$$Q^{(2)}=Q^{(2s)}+Q^{(2d)}+Q^{(2b)} \tag{B.26}$$

where:

$$Q_s^{(2s)}+\frac{1}{2}Q_{xx}^{(2s)}=-\frac{3}{4}ab_1\delta Q^{(0)}-\overline{\upsilon}^2\left(\frac{1}{4}I''I-\frac{1}{8}I'I'\right)Q^{(0)}$$

$$-\alpha^2\left(\frac{1}{4}b_2-\frac{3}{8}b_1^2\right)Q^{(0)} \text{ for } s<s' \tag{B.27a}$$

$$Q^{(2s)}=0 \quad \text{at } s=s' \tag{B.27b}$$

where:

$$Q_s^{(2a)}+\tfrac{1}{2}Q_{xx}^{(2a)}=\tfrac{1}{2}\alpha b_1(\eta-\delta)\left(xQ_x^{(0)}\right)-Q^{(0)} \quad \text{for } s<s' \tag{B.28a}$$

$$Q^{(2a)}=0 \text{ at } s=s' \tag{B.28b}$$

and where:

$$Q_s^{(2b)} + \tfrac{1}{2} Q_{xx}^{(2b)} = (\eta - \overline{\eta}) x Q_{xx}^{(1)} - \tfrac{1}{2}[\upsilon^2 - \overline{\upsilon}^2 - 3\overline{\eta}(\eta - \overline{\eta})] x^2 Q_{xx}^{(0)} \quad \text{for } s < s' \tag{B.29a}$$

$$Q^{(2b)} = 0 \quad \text{at } s = s' \tag{B.29b}$$

Once we have solved these equations, then the option price is then given by:

$$V(t, f, a) = [f - K]^+ + \tfrac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}\, I^{1/2}(\varepsilon z) e^{\frac{1}{4}\varepsilon^2 a b_1 \delta s^2} J \tag{B.30a}$$

where:

$$J = \int_s^{s_{ex}} Q^{(0)}(s, x; s')\, ds' + \varepsilon \int_s^{s_{ex}} Q^{(1)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2s)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2a)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2b)}(s, x; s')\, ds' + \cdots \tag{B.30b}$$

The terms $Q^{(1)}$, $Q^{(2a)}$, and $Q^{(2b)}$ arise from the time-dependence of the coefficients $\eta(s)$ and $\upsilon(s)$. Indeed, if $\eta(s)$ and $\upsilon(s)$ were constant in time, we would have $Q^{(1)} \equiv Q^{(2a)} \equiv Q^{(2b)} \equiv 0$, and the solution would be just $Q^{(s)} \equiv Q^{(0)} + \varepsilon^2 Q^{(2s)}$. Therefore, we will first solve for $Q^{(1)}$, $Q^{(2a)}$, and $Q^{(2b)}$, and then *try* to choose the constants δ, $\overline{\eta}$, and $\overline{\upsilon}$ so that the last three integrals are zero for all x. In this case, the option price would be given by:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)B(\varepsilon\alpha z)}\, I^{1/2}(\varepsilon z) \cdot e^{\frac{1}{4}\varepsilon^2 a b_1 \delta z^2} \int_s^{s_{ex}} Q^{(s)}(s, x; s')\, ds' \tag{B.31a}$$

and, through $O(\varepsilon^2)$, $Q^{(s)}$ would be the solution of the static problem:

$$Q_s^{(s)} + \frac{1}{2} Q_{xx}^{(s)} = -\frac{3}{4}\varepsilon^2 a b_1 \delta Q^{(s)} - \varepsilon^2 \overline{\upsilon}^2 \left(\frac{1}{4} I'' I - \frac{1}{8} I' I'\right) Q^{(s)} - \varepsilon^2 \alpha^2 \left(\frac{1}{4} b_2 - \frac{3}{8} b_1^2\right) Q^{(s)} \quad \text{for } s < s' \tag{B.31b}$$

$$Q^{(s)} = \delta(x) \quad \text{at } s = s' \tag{B.31c}$$

This is exactly the time-independent problem solved in Appendix A. See equations (A.42), (A.43a), and (A.43b). So *if* we can carry out this strategy, we can obtain option prices for the dynamic SABR model by reducing them to the previously-obtained prices for the static model.

Leading order analysis The solution of (B.24a), (B.24b) is Gaussian:

$$Q^{(0)} = G(x/\sqrt{\Delta}) \tag{B.32a}$$

where:

$$G(x/\sqrt{\Delta}) = \frac{1}{\sqrt{2\pi\,\Delta}} e^{-x^2/2\Delta}, \quad \Delta = s' - s \tag{B.32b}$$

For future reference, note that:

$$G_x = -\frac{x}{\Delta} G; \quad G_{xx} = \frac{x^2 - \Delta}{\Delta^2} G; \quad G_{xxx} = -\frac{x^3 - 3\Delta x}{\Delta^3} G \tag{B.33a}$$

$$G_{xxxx} = \frac{x^4 - 6\Delta x^2 + 3\Delta^2}{\Delta^4} G; \quad G_{xxxxx} = -\frac{x^5 - 10\Delta x^3 + 15\Delta^2 x}{\Delta^5} G \tag{B.33b}$$

$$G_{xxxxxx} = \frac{x^6 - 15\Delta x^4 + 45\Delta^2 x^2 - 15\Delta^3}{\Delta^6} G \tag{B.33c}$$

Order ε Substituting $Q^{(0)}$ into the equation for $Q^{(1)}$ and using (B.33a) yields:

$$\begin{aligned} Q_s^{(1)} + \frac{1}{2} Q_{xx}^{(1)} &= (\eta - \overline{\eta}) \frac{x^3 - \Delta x}{\Delta^2} G \\ &= -(s' - s)(\eta - \overline{\eta})\; G_{xxx} - 2(\eta - \overline{\eta}) G_x \quad \text{for } s < s' \end{aligned} \tag{B.34}$$

with the "initial" condition $Q^{(1)} = 0$ at $s = s'$. The solution is:

$$Q^{(1)} = A(s, s') G_{xxx} + 2A_{s'}(s, s') G_x \tag{B.35a}$$

$$= \frac{\partial}{\partial s'} \left\{ 2A(s, s') G_x (x/\sqrt{s' - s} \right\} \tag{B.35b}$$

where:

$$A(s, s') = \int_s^{s'} (s' - \tilde{s})[\eta(\tilde{s}) - \overline{\eta}]\, d\tilde{s}; \qquad A_{s'}(s, s') = \int_s^{s'} [\eta(\tilde{s}) - \overline{\eta}]\, d\tilde{s} \tag{B.35c}$$

This term contributes:

$$\int_s^{s_{ex}} Q^{(1)}(s, x; s')\, ds' = 2A(s, s_{ex}) G_x (x/\sqrt{s_{ex} - s}) \tag{B.36}$$

to the option price. See equations (B.30a), (B.30b). To eliminate this contribution, we chose $\overline{\eta}$ so that $A(s, s_{ex}) = 0$:

$$\overline{\eta} = \frac{\displaystyle\int_s^{s_{ex}} (s_{ex} - \tilde{s}) \eta(\tilde{s})\, d\tilde{s}}{\frac{1}{2}(s_{ex} - s)^2} \tag{B.37}$$

The $\varepsilon^2 Q^{(2a)}$ *term* From equation (B.28a) we obtain:

$$\begin{aligned} Q_s^{(2a)} + \frac{1}{2} Q_{xx}^{(2a)} &= -\frac{1}{2} \alpha b_1 (\eta - \delta) \frac{x^2 + \Delta}{\Delta} G \\ &= -\frac{1}{2} \alpha b_1 (\eta - \delta) \Delta G_{xx} - \alpha b_1 (\eta - \delta) G \end{aligned} \tag{B.38}$$

for $s < s'$, with $Q^{(2a)} = 0$ at $s = s'$. Solving then yields:

$$Q^{(2a)} = \frac{\partial}{\partial s'} \left\{ \alpha b_1 \int_s^{s'} (s' - \tilde{s})[\eta(\tilde{s}) - \delta]\, d\tilde{s} G(x/\sqrt{s' - s}) \right\} \tag{B.39}$$

This term makes a contribution of:

$$\int_s^{s_{ex}} Q^{(2a)}(s, x; s')\, ds' = \alpha b_1 \left(\int_s^{s_{ex}} (s_{ex} - \tilde{s})[\eta(\tilde{s}) - \delta]\, d\tilde{s} \right) G(x/\sqrt{s_{ex} - s}) \tag{B.40}$$

to the option price, so we choose:

$$\delta = \overline{\eta} = \frac{\displaystyle\int_s^{s_{ex}} (s_{ex} - \tilde{s})[\eta(\tilde{s}) - \delta]\, d\tilde{s}}{\frac{1}{2}(s_{ex} - \tilde{s})^2} \tag{B.41}$$

to eliminate this contribution.

The $\varepsilon^2 Q^{(2b)}$ term Substituting $Q^{(1)}$ and $Q^{(0)}$ into equation (B.29a), we obtain:

$$Q_s^{(2b)} + \tfrac{1}{2} Q_{xx}^{(2b)} = (\eta - \overline{\eta}) A x G_{xxxxx} + 2(\eta - \overline{\eta}) A_{s'} x G_{xxx} - \tfrac{1}{2} \kappa x^2 G_{xx} \tag{B.42a}$$

for $s < s'$, where:

$$\kappa = \upsilon^2(s) - \overline{\upsilon}^2 - 3\overline{\eta}[\eta(s) - \overline{\eta}] \tag{B.42b}$$

This can be re-written as:

$$\begin{aligned} Q_s^{(2b)} + \tfrac{1}{2} Q_{xx}^{(2b)} = {} & -(\eta - \overline{\eta}) A[\Delta G_{xxxxxx} + 5G_{xxxx}] - 2(\eta - \overline{\eta}) A_{s'}[\Delta G_{xxxx} \\ & + 3G_{xx}] - \tfrac{1}{2}\kappa[\Delta^2 G_{xxxx} + 5\Delta G_{xx} + 2G] \end{aligned} \tag{B.43}$$

Solving this with the initial condition $Q^{(2b)} = 0$ at $s = s'$ yields:

$$\begin{aligned} Q^{(2b)} = {} & \frac{1}{2} A^2(s, s') G_{xxxxxx} + 2A(s, s') A_{s'}(s, s') G_{xxxx} \\ & + 3 \int_s^{s'} [\eta(\tilde{s}) - \overline{\eta}] A(\tilde{s}, s') d\tilde{s} G_{xxxx} + 3 A_{s'}^2(s, s') G_{xx} \\ & + \frac{1}{2} \int_s^{s'} [s' - \tilde{s}]^2 \kappa(\tilde{s}) d\tilde{s} G_{xxxx} + \frac{5}{2} \int_s^{s'} [s' - \tilde{s}] \kappa(\tilde{s}) d\tilde{s} G_{xx} + \int_s^{s'} \kappa(\tilde{s}) d\tilde{s} G \end{aligned} \tag{B.44}$$

This can be written as:

$$\begin{aligned} Q^{(2b)} = \frac{\partial}{\partial s'} \Bigg\{ & 4A^2(s, s') G_{ss} - 12 \int_s^{s'} [\eta(\tilde{s}) - \overline{\eta}] A(\tilde{s}, s')\, d\tilde{s} G_s \\ & - 2 \int_s^{s'} (s' - \tilde{s})^2 \kappa(\tilde{s})\, d\tilde{s} G_s + \int_s^{s'} (s' - \tilde{s}) \kappa(\tilde{s})\, d\tilde{s} G \Bigg\} \end{aligned} \tag{B.45}$$

Recall that $\overline{\eta}$ was chosen above so that $A(s, s_{ex}) = 0$. Therefore the contribution of $Q^{(2b)}$ to the option price is:

$$\begin{aligned}\int_s^{s_{ex}} Q^{(2b)}(s, x; s')\, ds' = -\Bigg(&12 \int_s^{s_{ex}} [\eta(\tilde{s}) - \overline{\eta}] A(\tilde{s}, s_{ex})\, d\tilde{s} \\ &+2 \int_s^{s_{ex}} (s_{ex} - \tilde{s})^2 \kappa(\tilde{s})\, d\tilde{s}\Bigg) G_s(x/\sqrt{s_{ex} - s}) \\ &+\left(\int_s^{s_{ex}} (s_{ex} - \tilde{s}) \kappa(\tilde{s})\, d\tilde{s}\right) G(x/\sqrt{s_{ex} - s})\end{aligned} \tag{B.46}$$

where $\kappa = \upsilon^2(s) - \overline{\upsilon}^2 - 3\overline{\eta}[\eta(s) - \overline{\eta}]$.

We can choose the remaining "effective media" parameter $\overline{\upsilon}$ to set either the coefficient of $G_s(x/\sqrt{s_{ex} - s})$ or the coefficient of $G(x/\sqrt{s_{ex} - s})$ to zero, but cannot set both to zero to completely eliminate the contribution of the term $Q^{(2b)}$. We choose $\overline{\upsilon}$ to set the coefficient of $G_s(x/\sqrt{s_{ex} - s})$ to zero, for reasons that will become apparent in a moment:

$$\begin{aligned}\overline{\upsilon}^2 = \frac{1}{\frac{1}{3}(s_{ex} - \tilde{s})^3} \Bigg\{&\int_s^{s_{ex}} (s_{ex} - \tilde{s})^2 \upsilon^2(\tilde{s})\, d\tilde{s} - 3\overline{\eta} \int_s^{s_{ex}} (s_{ex} - \tilde{s})^2 [\eta(\tilde{s}) - \overline{\eta}]\, d\tilde{s} \\ &- 6 \int_s^{s_{ex}} \int_s^{s_1} s_2 [\eta(s_1) - \overline{\eta}][\eta(s_2) - \overline{\eta}]\, ds_2\, ds_1\end{aligned} \tag{B.47}$$

Then the remaining contribution to the option price is:

$$\begin{aligned}\int_s^{s_{ex}} Q^{(2b)}(s, x; s')\, ds' &= \frac{1}{2}\overline{\kappa}(s_{ex} - s)^2\ G(x/\sqrt{s_{ex} - s}) \\ &= \frac{1}{2}\overline{\kappa}(s_{ex} - s)^2\ Q^{(0)}(s, x; s_{ex})\end{aligned} \tag{B.48a}$$

where:

$$\overline{\kappa} = \frac{1}{\frac{1}{2}(s_{ex} - s)^2} \int_s^{s_{ex}} (s_{ex} - \tilde{s})[\upsilon^2(\tilde{s}) - \overline{\upsilon}^2]\, d\tilde{s} \tag{B.48b}$$

Here we have used $\int_s^{s_{ex}} (s_{ex} - \tilde{s})(\eta(\tilde{s}) - \overline{\eta}) d\tilde{s} = 0$ to simplify (B.48b).

Equivalent volatilities

We can now determine the implied volatility for the dynamic model by mapping the problem back to the static model of Appendix A. Recall from (B.30a), (B.30b) that the value of the option is:

$$V(t, f, a) = [f - K]^+ + \tfrac{1}{2}\varepsilon\alpha\sqrt{B(0)\ B(\varepsilon\alpha z)}\, I^{1/2}(\varepsilon z)\ e^{\frac{1}{4}\varepsilon^2 a b_1 \overline{\eta} z^2} J \tag{B.49a}$$

where:

$$J = \int_s^{s_{ex}} Q^{(0)}(s, x; s')\, ds' + \varepsilon \int_s^{s_{ex}} Q^{(1)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2s)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2a)}(s, x; s')\, ds' + \varepsilon^2 \int_s^{s_{ex}} Q^{(2b)}(s, x; s')\, ds' + \cdots \tag{B.49b}$$

and where we have used $\delta = \overline{\eta}$. We chose the "effective parameters" $\overline{\eta}$ and $\overline{\upsilon}$ so that the integrals of $Q^{(1)}$, $Q^{(2a)}$ contribute nothing to J. The integral of $Q^{(2b)}$ then contributed $\frac{1}{2}\varepsilon^2\overline{\kappa}(s_{ex} - s)^2 Q^{(0)}(s, x; s_{ex})$. The option price is:

$$\begin{aligned} J &= \int_s^{s_{ex}} \{Q^{(0)}(s, x; s') + \varepsilon^2 Q^{(2s)}(s, x; s')\}\, ds' \\ &\quad + \frac{1}{2}\varepsilon^2\overline{\kappa}(s_{ex} - s)^2\ Q^{(0)}(s, x; s_{ex}) + \cdots \\ &= \int_s^{\hat{s}_{ex}} \{Q^{(0)}(s, x; s') + \varepsilon^2 Q^{(2s)}(s, x; s')\}\, ds' + \cdots \end{aligned} \tag{B.50a}$$

through $O(\varepsilon^2)$, where:

$$\hat{s}_{ex} = s_{ex} + \tfrac{1}{2}\varepsilon^2\overline{\kappa}(s_{ex} - s)^2 + \cdots \tag{B.50b}$$

Through $O(\varepsilon^2)$ we can combine $Q^{(s)} = Q^{(0)}(s, x; s') + \varepsilon^2 Q^{(2s)}(s, x; s')$, where $Q^{(s)}$ solves the static problem:

$$Q_s^{(s)} + \frac{1}{2}Q_{xx}^{(s)} = -\frac{3}{4}\varepsilon^2 a b_1 \delta Q^{(s)} - \varepsilon^2\overline{\upsilon}^2\left(\frac{1}{4}I''I - \frac{1}{8}I'I'\right) \cdot Q^{(s)} - \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right)Q^{(s)} \qquad \text{for} \quad s < s' \tag{B51a}$$

$$Q^{(s)} = \delta(s - s') \qquad \text{at} \quad s = s' \tag{B.51b}$$

This problem is homogeneous in the time s, so its solution $Q^{(s)}$ depends only on the time difference $\tau = s' - s$. The option price is therefore:

$$V(t, f, a) = [f - K]^+ + \frac{1}{2}\varepsilon\alpha\sqrt{B(0)\ B(\varepsilon\alpha z)}I^{1/2} (\varepsilon z) e^{\frac{1}{4}\varepsilon^2 a b_1 \overline{\eta} z^2} \int_0^{\hat{s}_{ex-s}} Q^{(s)}(\tau, x)\, d\tau, \tag{B.52}$$

where $Q^s(\tau, x)$ is the solution of:

$$Q^{(s)}_\tau - \frac{1}{2}Q^{(s)}_{xx} = \frac{3}{4}\varepsilon^2 a b_1 \overline{\eta} Q^{(s)} + \varepsilon^2 \overline{\upsilon}^2 \left(\frac{1}{4}I''I - \frac{1}{8}I'I'\right) Q^{(s)} + \varepsilon^2\alpha^2\left(\frac{1}{4}b_2 - \frac{3}{8}b_1^2\right) Q^{(s)} \qquad \text{for} \quad \tau > 0 \tag{B.53a}$$

$$Q^s = \delta(x) \qquad \text{at} \quad \tau = 0 \tag{B.53b}$$

The option price defined by (B.52), (B.53a), and (B.53b) is identical to the static model's option price defined by (A.42), (A.43a), and (A.43b), provided we make the identifications:

$$\nu \to \overline{\upsilon} \qquad \rho \to \overline{\eta}/\overline{\upsilon} \tag{B.54}$$

$$\tau_{ex} \to \hat{s}_{ex} - s = s_{ex} - s + \tfrac{1}{2}\varepsilon^2\overline{\kappa}(s_{ex} - s)^2 \tag{B.55}$$

in Appendix A for the original non-dynamic SABR model, provided we make the identifications:

$$\tau_{ex} = \tau + \varepsilon^2 \int_0^\tau \tilde{\tau}[\upsilon^2(\tilde{\tau}) - \overline{\upsilon}^2]\, d\tilde{\tau} \tag{B.56a}$$

$$\nu \to \overline{\eta}/\overline{\upsilon} \qquad \nu \to \overline{\upsilon} \tag{B.56b}$$

See equations (A.42–A.43b). Following the reasoning in the preceding Appendix now shows that the European call price is given by the formula:

$$V(t, f, K) = (f - K)\mathcal{N}\left(\frac{f-K}{\sigma_N\sqrt{\tau_{ex}}}\right) + \sigma_N\sqrt{\tau_{ex}}\,\mathcal{G}\left(\frac{f-K}{\sigma_N\sqrt{\tau_{ex}}}\right) \tag{B.57}$$

with the implied *normal* volatility:

$$\sigma_N(K) = \frac{\varepsilon\alpha(f-K)}{\int_K^f \frac{df'}{C(f')}} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{2\gamma_2 - \gamma_1^2}{24}\alpha^2 C^2(f_{av}) + \tfrac{1}{4}\overline{\eta}\alpha\gamma_1 C(f_{av}) + \frac{2\overline{\upsilon}^2 - 3\overline{\eta}^2}{24} + \frac{1}{2}\overline{\theta}\right]\varepsilon^2\tau_{ex} + \cdots\right. \tag{B.58a}$$

where:

$$\zeta = \frac{\overline{\upsilon}}{\alpha}\frac{f-K}{C(f_{av})}, \quad \hat{x}(\zeta) = \log\left(\frac{\sqrt{1 - 2\overline{\eta}\zeta/\overline{\upsilon} + \zeta^2} - \overline{\eta}/\overline{\upsilon} + \zeta}{1 - \overline{\eta}/\overline{\upsilon}}\right) \tag{B.58b}$$

$$f_{av} = \sqrt{fK}, \qquad \gamma_1 = \frac{C'(f_{av})}{C(f_{av})} \qquad \gamma_2 = \frac{C''(f_{av})}{C(f_{av})} \tag{B.58c}$$

$$\overline{\theta} = \frac{\int_0^\tau \tilde{\tau}[\upsilon^2(\tilde{\tau}) - \overline{\upsilon}^2]d\tilde{\tau}}{\frac{1}{2}\tau^2} \tag{B.58d}$$

Equivalently, the option prices are given by Black's formula with the effective Black volatility of:

$$\sigma_B(K) = \frac{\alpha \log f/K}{\int_K^f \frac{df'}{C(f')}} \cdot \left(\frac{\zeta}{\hat{x}(\zeta)}\right) \cdot \left\{1 + \left[\frac{2\gamma_2 - \gamma_1^2 + 1/f_{av}^2}{24}\alpha^2 C^2(f_{av}) + \frac{1}{4}\overline{\eta}\alpha\gamma_1 C(f_{av}) + \frac{2\overline{v}^2 - 3\overline{\eta}^2}{24} + \frac{1}{2}\overline{\theta}\right]\varepsilon^2\tau_{ex} + \cdots\right.$$

REFERENCES

- Berner, N. A. BNP Paribas (2000) Private communication.
- Black, F. (1976) The pricing of commodity contracts. *J. Pol. Ec.* 81, 167–179.
- Breeden, D. T. and Litzenberger, R. H. (1994) Prices of state-contingent claims implicit in option prices. *J. Business* 51, 621–651.
- Clouet, J. F. (1998) Diffusion approximation of a transport process in random media. *SIAM J. Appl. Math* 58, 1604–1621.
- Cole, J. D. (1968) *Perturbation Methods in Applied Mathematics*. Ginn-Blaisdell.
- Derman, E. and Kani, I. (1994) Riding on a smile. *Risk*, Feb. 32–39.
- Derman, E. and Kani, I. (1998) Stochastic implied trees: arbitrage pricing with stochastic term and strike structure of volatility. *Int J Theor Appl Finance* 1, 61–110.
- Dupire, B. (1994) Pricing with a smile. *Risk*, Jan. 18–20.
- Dupire, B. (1997) Pricing and hedging with smiles. In *Mathematics of Derivative Securities*, M. A. H. Dempster and S. R. Pliska (eds). Cambridge University Press, Cambridge, pp. 103–111.
- Fouque, J. P., Papanicolaou, G., and Sirclair, K. R. (2000) *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press, Cambridge.
- Hagan, P. S. and Woodward, D. E. (1999) Equivalent black volatilities. *App. Math. Finance* 6, 147–157.
- Hagan, P. S., Lesniewski, A. and Woodward D. (in preparation) Geometrical optics in finance.
- Harrison, J. M. and Kreps, D. (1979) Martingales and arbitrage in multiperiod securities markets. *J. Econ. Theory* 20, 381–408.
- Harrison, J. M. and Pliska, S. (1981) Martingales and stochastic integrals in the theory of continuous trading. *Stoch. Proc. Appl.* 11, 215–260.
- Heston, S. L. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies* 6, 327–343.
- Hull, J. C. (1997) *Options, Futures and Other Derivatives*. Prentice-Hall, Hemel Hempstead.
- Hull, J. and White, A. (1987) The pricing of options on assets with stochastic volatilities. *J. Finance* 42, 281–300.
- Jamshidean, F. (1997) Libor and swap market models and measures. *Fin. Stoch.* 1, 293–330.
- Karatzas, I. and Shreve, S. (1988) *Brownian Motion and Stochastic Calculus*. Springer, Berlin.
- Karatzas, I., Lehoczky, J. P., Shreve, S. E. and Xu, G. L. (1991) Martingale and duality methods for utility maximization in an incomplete market. *SIAM J Control Optim* 29, 702–730.

■ Kevorkian, J. and Cole, J. D. (1985) *Perturbation Methods in Applied Mathematics*. Springer-Verlag, Berlin.
■ Lewis, A. (2000) *Option Valuation Under Stochastic Volatility*. Financial Press.
■ Musiela, M. and Rutkowski, M. (1998) *Martingale Methods in Financial Modelling*. Springer, Berlin.
■ Neu, J. C. (1978) Doctoral Thesis, California Institute of Technology.
■ Okdendal, B. (1998) *Stochastic Differential Equations*. Springer, Berlin.
■ Steele, J. M. (2001) *Stochastic Calculus and Financial Applications*. Springer, Berlin.
■ Wan, F. (1991) *A Beginner's Book of Modeling*. Springer-Verlag, Berlin.
■ Whitham, G. B. (1974) *Linear and Nonlinear Waves*. Wiley, Chichester.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance*. Wiley, Chichester.

19
Adjusters: Turning Good Prices into Great Prices

Patrick S. Hagan

Wilmott magazine, December 2002

I'm sure we've all been there: We need to price and trade an exotic derivative, but because of limitations in our pricing systems, we cannot calibrate on the "natural set" of hedging instruments. Instead we have to calibrate on some other set of vanilla instruments, which provide only a poor representation of the exotic. Consequently, our prices are questionable, and if we are bold enough to trade on these prices, our hedges are unstable, chewing up any profit as bid-ask spread. Here we discuss how to get out of these jams by using "adjusters", a technique for re-expressing the vega risks of an exotic derivative in terms of its "natural hedging instruments". This helps prevent unstable hedges and exotic deal mis-management, and, as a side benefit, leads to significantly better pricing of the exotic.

1 Managing exotics

First let us briefly discuss how we get in these jams. During the normal course of business, the pricing and management of fixed income derivatives depend on two key markets. First is the swap market (*delta market*), which is encapsulated by the yield curve. Swap desks maintain current yield curves by continually stripping and re-stripping a set of liquid swaps, futures, and deposit rates throughout the day. This curve determines all current swap rates, FRA rates, forward swap rates, etc. The yield curve also shows how to hedge all interest rate risks by buying and selling the same swaps, futures, and deposit rates used in the stripping process.

The second market is the vanilla option market (*vega market*) for European swaptions, caps, and floors. Prices of these options are quoted in terms of the volatility σ, which is inserted into Black's 1976 formula to determine the dollar price of the option. European swaptions are defined by three numbers: the exercise date and the *tenor* (length) and *strike* (fixed

Contact address: Quantitative Research & Development, Bloomberg LP, 499 Park Avenue, New York, NY 10022, USA.
E-mail: phagan1@bloomberg.net

rate) of the swap received upon exercise. Keeping track of this market requires maintaining a *volatility cube*, which contains the volatilities σ as a function of the three coordinates. However, the vast majority of swaptions are struck *at-the-money*, at strikes equaling the current swap rate of the underlying forward swap, so desks normally track this market by maintaining a volatility matrix containing the vols of at-the-money swaptions (see Table 1), and a set of auxiliary "smile" matrices showing how much to add/subtract to the volatility for strikes 50 bps, 100 bps, ... above or below the swap rate. Alternatively, some swap desks determine the adjustment by using a smile model, such as the SABR or Heston models. In any case, desks are reasonably confident that they can trade the vanilla instruments at the indicated prices.

TABLE 1: AT-THE-MONEY VOLATILITY MATRIX

European swaptions are defined by the time-to-exercise (row), and length (column) and fixed rate (strike) of the swap received upon exercise. A volatility matrix (as opposed to a volatility cube) contains the volatilities of at-the-money swaptions, swaptions whose fixed rates are equal to the current forward swap rate of the underlying swap. Linear interpolation is used for the volatilities in between grid points. The 3 m column is the caplet column.

σ (in %)	3 m	1 y	2 y	3 y	...	10 y
1 m	5.25	12.25	13.50	14.125	...	14.25
3 m	7.55	13.00	14.125	14.375	...	14.50
6 m	11.44	14.25	14.875	15.00	...	14.75
1 y	16.20	16.75	16.375	16.125	...	15.50
2 y	19.25	17.75	17.125	17.00	...	15.75
⋮	⋮	⋮	⋮	⋮		⋮
10 y	14.00	13.50	13.00	12.50	...	11.00

Now consider the typical management of an *exotic* interest rate derivative, such as a Bermudan swap or a callable range note. During the nightly mark-to-market, the deal will be priced by:

- Selecting an interest rate model, such as Hull–White or Black–Karasinski;
- Selecting a set of vanilla swaptions and/or caplets as the calibration instruments;
- Calibrating the interest rate model so that the model reproduces the market prices of these instruments, either exactly or in a least squares sense;
- Using the calibrated model to find the value of the exotic via finite difference methods, trees, or Monte Carlo.

The exotic's vega risks will then be obtained by:

- Bumping each volatility in the matrix (or cube) one at a time;
- Re-calibrating the model and re-pricing the exotic derivative for each bump; and
- Subtracting to obtain the difference in value for the bumped case versus the base (market) case.

This results in a matrix of vega risks. Each cell represents the deal's dollar gain or loss should the volatility of that particular swaption change. These vega risks are then hedged by buying or selling enough of each underlying swaption so that the total vega risks are zero. Of course the desk first adds up the vega exposure of all deals, and only hedges the net exposure.

Calibration is the only step in this procedure which incorporates information about market volatilities. Under the typical nightly procedure *the exotic derivative will only have vega risks to the set of vanilla swaptions and/or caplets used in calibration*. So regardless of the actual nature of the exotic derivative, the vega hedges will be trying to mimic the exotic derivative as a linear combination of the calibration instruments. If the calibration instruments are "natural hedging instruments" which are "similar" to the exotic, then the hedges probably provide a faithful representation of the exotic. If the calibration instruments are dissimilar to the exotic, having the wrong expiries, tenors, or strikes, then the vega hedges will probably be a poor representation of the exotic. This often causes the hedges to be unstable, which gets expensive as bid-ask spread is continually chewed up in re-hedging the exotic.

For example, consider a cancellable 10 year receiver swap struck at 7.50%, where the first call date is in 3 years (10NC3@7.50). Surely the natural hedging instruments for this Bermudan are the diagonal swaptions: the 3 y into 7 y struck at 7.50%, the 4 y into 6 y struck at 7.50%, . . . , and the 9 y into 1 y struck at 7.50%, since a dynamic combination of these instruments should be capable of accurately replicating the exotic. Indeed, if we do not calibrate on these swaptions, then our calibrated model would not get the correct market prices of these swaptions, and if our prices for the 3 y into 7 y, the 4 y into 6 y (etc.) are incorrect, we don't have a prayer of pricing and hedging the callable swap correctly.

When feasible, best practice is to use autocalibration for managing exotic books. For each exotic derivative on the books, autocalibration first selects the "natural hedging instruments" of the exotic, usually based on some simple scheme of matching the expiries, tenors, and effective strikes of the exotic. It then re-calibrates the model to match these instruments to their market values, and then values the exotic. Autocalibration then picks the next deal out of the book, selects a new set of natural hedging instruments, re-calibrates the model, and re-prices the exotic, and so on.

There are a variety of reasons why autocalibration may not be feasible. If one's interest rate model is too complex, perhaps a several factor affair, one may not have the computational resources to allow frequent calibration. Or if one's calibration software is too "fractious", one may not have the patience to calibrate the model very often. In such cases one would generally calibrate to all swaptions in the vol matrix in a least squares sense, and the calibration would only include at-the-money swaptions. Alternatively, an interest rate model may be more easily calibrated on some instruments than others. For example, a multi-factor BGM model is much easier to calibrate to caplets than to swaptions.

Finally, one's software may not be set up to calibrate on the "natural hedging instruments". A callable range note provides an example. Consider a regular (non-callable) 10 year range note which pays a coupon of, say, $1 each day Libor sets between 2.50% and 6.00%. Apart from minor date differences, the range note is equivalent to being long one digital call at 2.50% and short a digital call at 6.00% for each day in the next 10 years. Since digital calls can be written in terms of ordinary calls, a range note is very, very close to being a vanilla instrument, and can be priced exactly from the swaption volatility matrix (or cube). To price a *callable* range

note, one would like to calibrate on the underlying daily range notes, for if we don't price the underlying range notes correctly, how could we trust our price for the callable range note? Yet many systems are not set up to calibrate on range notes.

2 Risk migration

We now describe a method for moving the vega risk, either all of it, or as much as possible, to the natural hedging instruments. Suppose we have an exotic derivative v which has $h_1, h_2, \ldots, h_m$ as its natural hedging instruments. For example, for the 10NC3 Bermudan struck at 7.50%, the natural hedging instruments are just the 3 y into 7 y swaption struck at 7.50%, the 4 y into 6 y at 7.50%, ..., and the 9 y into 1 y at 7.50%. Suppose that for "operational reasons" one could not calibrate on $h_1, h_2, \ldots, h_m$, but instead were forced to calibrate on the swaptions and/or caplets $S_1, S_2, \ldots, S_n$. Let these instruments have market volatities $\sigma_1, \sigma_2, \ldots, \sigma_n$. Then after calibrating the model, all prices obtained from the model are functions of these volatilities. So let:

$$V^{\text{mod}} = V^{\text{mod}}(\sigma_1, \sigma_2, \ldots, \sigma_n) \tag{1a}$$

be the value of the exotic derivative v obtained from the model. Suppose we use the model to price the natural hedging instruments $h_1, h_2, \ldots, h_m$. Let:

$$H_k^{\text{mod}}(\sigma_1, \sigma_2, \ldots, \sigma_n) \quad k = 1, 2, \ldots, m \tag{1b}$$

be the value of these instruments according to the calibrated model. Finally, let:

$$H_k^{mar} \quad k = 1, 2, \ldots, m \tag{1c}$$

be the market price of the natural hedging instruments.

Let us create an imaginary portfolio consisting of the exotic derivative and its natural hedging instruments:

$$\pi = v - \sum_{k=1}^{m} b_k h_k, \tag{2}$$

where the amounts b_k of the hedging instruments will be selected shortly. Using the calibrated model to price this portfolio yields:

$$\Pi = V^{\text{mod}}(\sigma_1, \sigma_2, \ldots, \sigma_n) - \sum_{k=1}^{m} b_k H_k^{\text{mod}}(\sigma_1, \sigma_2, \ldots, \sigma_n), \tag{3a}$$

According to the calibrated model, this portfolio has the vega risks:

$$\frac{\partial \Pi}{\partial \sigma_j} = \frac{\partial V^{\text{mod}}}{\partial \sigma_j} - \sum_{k=1}^{m} b_k \frac{\partial H_k^{\text{mod}}}{\partial \sigma_j} \tag{3b}$$

to the calibration instruments.

Suppose we have chosen the portfolio weights b_k. (In the next section we show how to choose the amounts b_k so as to eliminate the vega risks, either completely or as completely as

possible). We add and subtract this portfolio of natural hedging instruments to write the exotic derivative v as:

$$v = \left\{ v - \sum_{k=1}^{m} b_k h_k \right\} + \left\{ \sum_{k=1}^{m} b_k h_k \right\}. \tag{4}$$

We now use the calibrated model to value the instruments in the first set of braces, and use the market prices to evaluate the instruments in the second set of braces. This yields the adjusted price:

$$V^{adj} = \left\{ V^{\text{mod}} - \sum_{k=1}^{m} b_k H_k^{\text{mod}} \right\} + \sum_{k=1}^{m} b_k H_k^{mar} \tag{5a}$$

$$= V^{\text{mod}} + \sum_{k=1}^{m} b_k (H_k^{mar} - H_k^{\text{mod}}) \tag{5b}$$

This procedure is generally known as "applying an adjuster". In equation 5a, the terms in braces are evaluated using the calibrated model, so they only have vega risk to the volatilities of the calibration instruments $\sigma_1, \sigma_2, \ldots, \sigma_n$. With the weights b_k chosen to eliminate these risks as nearly as possible, the adjusted price V^{adj} has little or no vega risk to the calibration instruments. Instead, the vega risks of the adjusted price come from the last term:

$$\sum_{k=1}^{m} b_k H_k^{mar} \tag{6a}$$

which only contains the market prices of the natural hedging instruments. So, as claimed, the adjuster has moved the vega risks from the calibration instruments to the natural hedging instruments. In fact, to hedge these risks one must take the opposite position:

$$-\sum_{k=1}^{m} b_k h_k \tag{6b}$$

in the natural hedging instruments of the exotic. For the 10NC3 Bermudan struck at 7.50%, for example, the resulting hedge is a combination of the 3 y into 7 y, the 4 y into 6 y, ..., and the 9 y into 1 y swaptions, all struck at 7.50%, regardless of which set of instruments were used to originally calibrate the model.

Equation 5b gives a different view. It shows the adjusted price as being the model price corrected for the difference between the market price and the model price of the natural hedging instruments.

3 Choosing the portfolio weights

We wish to choose the amounts b_k to minimize the model's vega risks in 3b. This is an exercise in linear algebra. Define the matrix M and vector $\mathbf{U}$ by:

$$M_{jk} = \frac{\partial H_k^{\text{mod}}}{\partial \sigma_j}, \quad U_j = \frac{\partial V}{\partial \sigma_j}, \tag{7a}$$

and let $\mathbf{b}$ be the vector of positions $(b_1, b_2, \ldots, b_m)^T$ so that the vega risks to the calibration instruments are:

$$\mathbf{U} - M\mathbf{b} \tag{7b}$$

There are three cases to consider. First suppose that there are fewer hedging instruments than model calibration instruments. One cannot expect to eliminate n risks with $m < n$ hedging instruments, so one cannot eliminate all the vega risks in 3b in this case. Instead one can minimize the sum of squares of the vega risks:

$$\min(\mathbf{U} - M\mathbf{b})^T(\mathbf{U} - M\mathbf{b}) \tag{8a}$$

Solving this problem yields:

$$\mathbf{b} = (M^T M)^{-1} M^T \quad (\text{if } m < n) \tag{8b}$$

The matrix $(M^T M)^{-1} M^T$ is known as the pseudo-inverse of M. Of course one can use some criterion other than least squares, such as choosing the portfolio $\mathbf{b}$ to eliminate the least liquid calibration instruments first.

If there are exactly as many hedging instruments as calibration instruments, then we can expect to completely eliminate the risk by choosing:

$$\mathbf{b} = M^{-1}\mathbf{U} \quad (\text{if } m = n) \tag{9}$$

Finally, if there are more hedging instruments than model calibration instruments, then we can select the smallest hedge which completely eliminates the vega risks to the calibration instruments:

$$\min \mathbf{b}^T\mathbf{b} \quad \text{subject to } M\mathbf{b} = \mathbf{U}. \tag{10a}$$

This yields:

$$\mathbf{b} = M^T(MM^T)^{-1}\mathbf{U} \quad (\text{if } m > n), \tag{10b}$$

where the matrix $M^T(MM^T)^{-1}$ is also known as the pseudo-inverse of M. As before, one may use a criterion other than least squares for choosing $\mathbf{b}$.

4 Examples

Consider once more the cancellable 10 year receiver swap struck at 7.50%, where the first call date is in 3 years. This derivative is normally booked as a straight 10 year swap, with a Bermudan option to enter into the opposite swap. Here we just price the Bermudan option, the option to enter a payer swaption at 7.50% on any coupon date starting on the third anniversary of the deal. For the purposes of this example, we assume a flat 5% yield curve, and use the Hull–White model with the USD volatility matrix from March 1999.

Clearly the natural hedging instruments are the 3 y into 7 y swaption struck at 7.50%, the 4 y into 6 y swaption at 7.50%, ..., and the 9 y into 1 y swaption at 7.50%. Suppose we calibrate the Hull-White model to these "natural hedging instruments" and then use the calibrated model to price the Bermudan. This leads to a price of:

$$V = 200.18 \text{ bps} \tag{11}$$

This represents the best price available within the one factor, Hull–White framework.

Suppose we calibrate to the same "diagonal" swaptions as before, but instead of calibrating to swaptions struck at 7.50%, we calibrate to swaptions struck at-the-money, at 5.00%. This yields a much lower price:

$$V^{\text{mod}} = 163.31 \text{ bps} \tag{12a}$$

If we add in the adjustor, we obtain the price:

$$V^{\text{mod}} + \sum_{k=1}^{m} b_k (H_k^{mar} - H_k^{\text{mod}}) = 163.31 \text{ bps} + 39.18 \text{ bps} = 202.49 \text{ bps} \tag{12b}$$

a great improvement.

Alternatively, suppose we calibrate the Hull–White model to the caplets starting at 3 years, at 3.25 years, at 3.5 years, ..., and at 9.75 years, with all caplets struck at 7.50%. Now we have the correct strike, but the wrong tenors. The calibrated model yields the price:

$$V^{\text{mod}} = 196.82 \text{ bps} \tag{13a}$$

If we add in the adjustor, we obtain a price of:

$$V^{\text{mod}} + \sum_{k=1}^{m} b_k (H_k^{mar} - H_k^{\text{mod}}) = 196.82 \text{ bps} + 3.12 \text{ bps} = 199.94 \text{ bps} \tag{13b}$$

again a distinct improvement.

5 Nothing is free

At first glance, it appears that using an adjuster greatly increases the computational load. After all, to determine the adjustment requires computing the exotic derivative's vega risk $\partial V^{\text{mod}}/\partial \sigma_j$ to all calibration instruments. These risks are usually found via finite differences, so evaluating these risks would seem to require model calibrations in $n + 1$ separate scenarios (base case, and each σ_j bumped separately). However, these vega risks are needed for hedging purposes, and are nearly always computed as part of the nightly batch, even if one is *not* applying an adjustor. So computing the vega matrix is usually free. The computational load does increase modestly, because for each natural hedging instrument, one has to calculate the model price H_k^{mod} and its vega derivatives $\partial H_k^{\text{mod}}/\partial \sigma_j$. This requires calculating the model price of m vanilla instruments $n + 1$ times. This is the same load as calculating the *calibration error* in each of the $n + 1$ scenarios, clearly much much faster than actually *calibrating* the model in each of the $n + 1$ scenarios.

20
Convexity Conundrums: Pricing CMS Swaps, Caps, and Floors

Patrick S. Hagan

Wilmott magazine, March 2003

I'm sure we've all been there: We're in hot competition with another bank over a deal. As the deal evolves, our trading team starts getting pushed around the market, and it dawns on us that the other bank's pricing is better than ours, at least for this class of deals. Here we focus on a single class of deals, the constant maturity swaps, caps, and floors. We develop a framework that leads to the standard methodology for pricing these deals, and then use this framework to systematically improve the pricing.

Let us start by agreeing on basic notation. In our notation, today is always $t = 0$. We use:

$$Z(t;T) = \text{value at date } t \text{ of a zero coupon bond with maturity } T \tag{1a}$$

$$D(T) \equiv Z(0,T) = \text{today's discount factor for maturity } T \tag{1b}$$

We distinguish between zero coupon bonds and discount factors to remind ourselves that discount factors are not random, we can always obtain the current discount factors $D(T)$ by stripping the yield curve, while zero coupon bonds $Z(t,T)$ remain random until the present catches up to date t. We also use:

$$\text{cvg}(t_{st}, t_{end}, \text{dcb}) \tag{2}$$

Contact address: Quantitative Research & Development, Bloomberg LP, 499 Park Avenue, New York, NY 10022, USA.
E-mail: phagan1@bloomberg.net

to denote the *coverage* (also called the *year fraction* or *day count fraction*) of the period t_{st} to t_{end}, where dcb is the day count basis (Act360, 30360, ...) specified by the contract. So if interest accrues at rate R, then $\text{cvg}(t_{st}, t_{end}, \text{dcb})R$ is the interest accruing in the interval t_{st} to t_{end}.

Deal definition

Consider a CMS swap leg paying, say, the N year swap rate plus a margin m. Let $t_0, t_1, \ldots, t_m$ be the dates of the CMS leg specified in the contract. (These dates are usually quarterly.) For each period j, the CMS leg pays:

$$\delta_j(R_j + m) \quad \text{paid at } t_j \quad \text{for } j = 1, 2, \ldots, m \tag{3a}$$

where R_j is the N year swap rate and:

$$\delta_j = \text{cvg}(t_{j-1}, t_j, \text{dcb}_{pay}) \tag{3b}$$

is the coverage of interval j. If the CMS leg is *set-in-advance* (this is standard), then R_j is the rate for a standard swap that begins at t_{j-1} and ends N years later (Figure 1). This swap rate is fixed on the date τ_j that is *spot lag* business days before the interval begins at t_{j-1}, and pertains throughout the interval, with the accrued interest $\delta_j(R_j + m)$ being paid on the interval's end date, t_j. Although set-in-advance is the market standard, it is not uncommon for contracts to specify CMS legs *set-in-arrears*. Then R_j is the N year swap rate for the swap that begins on the *end date* t_j of the interval, not the start date, and the fixing date τ_j for R_j is *spot lag* business days before the interval *ends* at t_j. As before, δ_j is the coverage for the j^{th} interval using the day count basis dcb_{pay} specified in the contract. Standard practice is to use the 30360 basis for USD CMS legs.

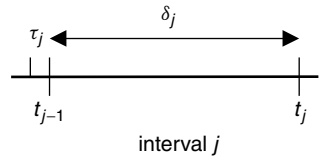


Figure 1: j^{th} interval of a "set-in-advance" CMS leg

CMS caps and floors are constructed in an almost identical fashion. For CMS caps and floors on the N year swap rate, the payments are:

$$\delta_j[R_j - K]^+ \quad \text{paid at } t_j \quad \text{for } j = 1, 2, \ldots, m, \quad \text{(cap)} \tag{4a}$$

$$\delta_j[K - R_j]^+ \quad \text{paid at } t_j \quad \text{for } j = 1, 2, \ldots, m, \quad \text{(floor)} \tag{4b}$$

where the N year swap rate is set-in-advance or set-in-arrears, as specified in the contract.

Reference swap

The value of the CMS swap, cap, or floor is just the sum of the values of each payment. Any margin payments m can also be valued easily. So all we need do is value a single payment of the three types:

$$R_s \quad \text{paid at } t_p \tag{5a}$$

$$[R_s - K]^+ \quad \text{paid at } t_p \tag{5b}$$

$$[K - R_s]^+ \quad \text{paid at } t_p \tag{5c}$$

Here the reference rate R_s is the par rate for a standard swap that starts at date s_0, and ends N years later at s_n. To express this rate mathematically, let $s_1, s_2, \ldots, s_n$ be the swap's (fixed leg) pay dates. Then a swap with rate R_{fix} has the fixed leg payments:

$$\alpha_j R_{fix} \quad \text{paid at } s_j \quad \text{for } j = 1, 2, \ldots, n \tag{6a}$$

where:

$$\alpha_j = \text{cvg}(t_{j-1}, t_j, \text{dcb}_{sw}) \tag{6b}$$

is the coverage (fraction of a year) for each period j, and dcb_{sw} is the standard swap basis. In return for making these payments, the payer receives the floating leg payments. Neglecting any basis spread, the floating leg is worth 1 paid at the start date s_0, minus 1 paid at the end date s_n. At any date t, then, the value of the swap to the payer is:

$$V_{sw}(t) = Z(t; s_0) - Z(t; s_n) - R_{fix} \sum_{j=1}^{n} \alpha_j Z(t; s_j) \tag{7}$$

The *level* of the swap (also called the *annuity*, *PV01*, *DV01*, or *numerical duration*) is defined as:

$$L(t) = \sum_{j=1}^{n} \alpha_j Z(t; s_j) \tag{8}$$

Crudely speaking, the level $L(t)$ represents the value at time t of receiving \$1 per year (paid annually or semiannually, according to the swap's frequency) for N years. With this definition, the value of the swap is:

$$V_{sw}(t) = [R_s(t) - R_{fix}]L(t) \tag{9a}$$

where:

$$R_s(t) = \frac{Z(t; s_0) - Z(t; s_n)}{L(t)} \tag{9b}$$

Clearly the swap is worth zero when R_{fix} equals $R_s(t)$, so $R_s(t)$ is the par swap rate at date t. In particular, today's level:

$$L_0 = L(0) = \sum_{j=1}^{n} \alpha_j D_j = \sum_{j=1}^{n} \alpha_j D(s_j) \tag{10a}$$

and today's (forward) swap rate:

$$R_s^0 = R_s(0) = \frac{D_0 - D_n}{L_0} \tag{10b}$$

are both determined by today's discount factors.

Valuation

According to the theory of arbitrage free pricing, we can choose any freely tradeable instrument as our *numeraire*. Examining (8) shows that the level $L(t)$ is just the value of a collection of zero coupon bonds, since the coverages α_j are just fixed numbers. These are clearly freely tradeable instruments, so we can choose the level $L(t)$ as our numeraire.[1] The usual theorems then guarantee that there exists a probability measure such that the value $V(t)$ of any freely tradeable deal divided by the numeraire is a Martingale. So:

$$V(t) = L(t)\mathrm{E}\left\{ \frac{V(T)}{L(T)} \middle| \mathcal{F}_t \right\} \quad \text{for any } T > t \tag{11}$$

provided there are no cash flows between t and T.

It is helpful to examine the valuation of a plain vanilla swaption. Consider a standard European option on the reference swap. The exercise date of such an option is the swap's fixing date τ, which is spot-lag business days before the start date s_0. At this exercise date, the payoff is the value of the swap, provided this value is positive, so:

$$V_{opt}(\tau) = [R_s(\tau) - R_{fix}]^+ L(\tau) \tag{12}$$

on date τ. Since the Martingale formula (11) holds for any $T > t$, we can evaluate it at $T = \tau$, obtaining:

$$V_{opt}(t) = L(t)\mathrm{E}\left\{ \frac{V_{opt}(\tau)}{L(\tau)} \middle| \mathcal{F}_t \right\} = L(t)\mathrm{E}\{[R_s(\tau) - R_{fix}]^+ | \mathcal{F}_t\} \tag{13}$$

In particular, today's value of the swaption is:

$$V_{opt}(t) = L_0\mathrm{E}\{[R_s(\tau) - R_{fix}]^+ | \mathcal{F}_0\}. \tag{14a}$$

Moreover, (9b) shows that the par swap rate $R_s(t)$ is the value of a freely tradable instrument (two zero coupon bonds) divided by our numeraire. So the swap rate must also be a Martingale, and:

$$\mathrm{E}\{R_s(\tau) | \mathcal{F}_0\} = R_s(0) \equiv R_s^0 \tag{14b}$$

To complete the pricing, one now has to invoke a mathematical model (Black's model, Heston's model, the SABR model, ...) for how $R_s(\tau)$ is distributed around its mean value R_s^0. In Black's model, for example, the swap rate is distributed according to:

$$R_s(\tau) = R_s^0 e^{\sigma x\sqrt{\tau} - \frac{1}{2}\sigma^2\tau} \tag{15}$$

where x is a normal variable with mean zero and unit variance. One completes the pricing by integrating to calculate the expected value.

CMS caplets

The payoff of a CMS caplet is:

$$[R_s(\tau) - K]^+ \quad \text{paid at } t_p \tag{16}$$

On the swap's fixing date τ, the par swap rate R_s is set and the payoff is known to be $[R_s(\tau) - K]^+ Z(\tau; t_p)$, since the payment is made on t_p. Evaluating (11) at $T = \tau$ yields:

$$V_{cap}^{CMS}(t) = L(t)\mathrm{E}\left\{ \frac{[R_s(\tau) - K]^+ Z(\tau; t_p)}{L(\tau)} \middle| \mathcal{F}_t \right\} \tag{17a}$$

In particular, today's value is:

$$V_{cap}^{CMS}(0) = L_0\mathrm{E}\left\{ \frac{[R_s(\tau) - K]^+ Z(\tau; t_p)}{L(\tau)} \middle| \mathcal{F}_0 \right\} \tag{17b}$$

The ratio $Z(\tau; t_p)/L(\tau)$ is (yet another!) Martingale, so it's average value is today's value:

$$\mathrm{E}\{Z(\tau; t_p)/L(\tau)|\mathcal{F}_0\} = D(t_p)/L_0 \tag{18}$$

By dividing $Z(\tau; t_p)/L(\tau)$ by its mean, we obtain:

$$V_{cap}^{CMS}(0) = D(t_p)\mathrm{E}\left\{ [R_s(\tau) - K]^+ \frac{Z(\tau; t_p)/L(\tau)}{D(t_p)/L_0} \middle| \mathcal{F}_0 \right\} \tag{19}$$

which can be written more evocatively as:

$$\begin{aligned} V_{cap}^{CMS}(0) = {} & D(t_p)\mathrm{E}\{[R_s(\tau) - K]^+|\mathcal{F}_0\} \\ & + D(t_p)\mathrm{E}\left\{ [R_s(\tau) - K]^+ \left(\frac{Z(\tau; t_p)/L(\tau)}{D(t_p)/L_0} - 1 \right) \middle| \mathcal{F}_0 \right\} \end{aligned} \tag{20}$$

The first term is exactly the price of a European swaption with notional $D(t_p)/L_0$, regardless of how the swap rate $R_s(\tau)$ is modeled. The last term is the "convexity correction". Since $R_s(\tau)$ is a Martingale and $[Z(\tau; t_p)/L(\tau)]/[Z(t; t_p)/L(t)] - 1$ is zero on average, this term goes to zero linearly with the variance of the swap rate $R_s(\tau)$, and is much, much smaller than the first term.

There are two steps in evaluating the convexity correction. The first step is to *model* the yield curve movements in a way that allows us to re-write the level $L(\tau)$ and the zero coupon bond $Z(\tau;t_p)$ in terms of the swap rate R_s. (One obvious model is to allow only parallel shifts of the yield curve.) Then we can write:

$$Z(\tau;t_p)/L(\tau) = G(R_s(\tau)) \tag{21a}$$

$$D(t_p)/L_0 = G(R_s^0) \tag{21b}$$

for some function $G(R_s)$. The convexity correction is then just the expected value:

$$cc = D(t_p)\mathrm{E}\left\{[R_s(\tau) - K]^+ \left(\frac{G(R_s(\tau))}{G(R_s^0)} - 1\right)\middle| \mathcal{F}_0\right\} \tag{22}$$

over the swap rate $R_s(\tau)$. The second step is to evaluate this expected value.

In the Appendix we start with the street-standard model for expressing $L(\tau)$ and $Z(\tau;t_p)$ in terms of the swap rate R_s. This model uses bond math to obtain:

$$G(R_s) = \frac{R_s}{(1 + R_s/q)^{\Delta}} \frac{1}{1 - \dfrac{1}{(1 + R_s/q)^n}} \tag{23a}$$

Here q is the number of periods per year (1 if the reference swap is annual, 2 if it is semi-annual, ...), and:

$$\Delta = \frac{t_p - s_0}{s_1 - s_0} \tag{23b}$$

is the fraction of a period between the swap's start date s_0 and the pay date t_p. For deals "set-in-arrears" $\Delta = 0$. For deals "set-in-advance", if the CMS leg dates $t_0, t_1, \ldots$ are quarterly, then t_p is 3 months after the start date s_0, so $\Delta = \frac{1}{2}$ if the swap is semiannual and $\Delta = \frac{1}{4}$ if it is annual.

In the Appendix we also consider increasingly sophisticated models for expressing $L(\tau)$ and $Z(\tau;t_p)$ in terms of the swap rate R_s, and obtain increasingly sophisticated functions $G(R_s)$.

We can carry out the second step by *replicating* the payoff in (22) in terms of payer swaptions. For any smooth function $f(R_s)$ with $f(K) = 0$, we can write:

$$f'(K)[R_s - K]^+ + \int_K^{\infty} [R_s - x]^+ f''(x)\, dx = \begin{cases} f(R_s) & \text{for } R_s > K \\ 0 & \text{for } R_s < K \end{cases} \tag{24}$$

Choosing:

$$f(x) \equiv [x - K]\left(\frac{G(x)}{G(R_s^0)} - 1\right) \tag{25}$$

and substituting this into (22), we find that:

$$cc = D(t_p)\left\{f'(K)E\{[R_s(\tau) - K]^+|\mathcal{F}_0\} + \int_K^{\infty} f''(x)E\{[R_s(\tau) - x]^+|\mathcal{F}_0\}\, dx\right\} \tag{26}$$

Together with the first term, this yields:

$$V_{cap}^{CMS}(0) = \frac{D(t_p)}{L_0}\left\{[1 + f'(K)]C(K) + \int_K^\infty C(x)f''(x)\,dx\right\} \tag{27a}$$

as the value of the CMS caplet, where:

$$C(x) = L_0\mathrm{E}\{[R_s(\tau) - x]^+|\mathcal{F}_0\} \tag{27b}$$

is the value of an ordinary payer swaption with strike x.

This formula replicates the value of the CMS caplet in terms of European swaptions at different strikes x. At this point some pricing systems break the integral up into 10bp or so buckets, and re-write the convexity correction as the sum of European swaptions centered in each bucket. These swaptions are then consolidated with the other European swaptions in the vanilla book, and priced in the vanilla pricing system. This "replication method" is the most accurate method of evaluating CMS legs. It also has the advantage of automatically making the CMS pricing and hedging consistent with the desk's handling of the rest of its vanilla book. In particular, it incorporates the desk's smile/skew corrections into the CMS pricing. However, this method is opaque and compute intensive. After briefly considering CMS floorlets and CMS swaplets, we develop simpler approximate formulas for the convexity correction, as an alternative to the replication method.

CMS floorlets and swaplets

Repeating the above arguments shows that the value of a CMS floorlet is given by:

$$V_{floor}^{CMS}(0) = \frac{D(t_p)}{L_0}\left\{[1 + f'(K)]P(K) - \int_{-\infty}^K P(x)f''(x)\,dx\right\} \tag{28a}$$

where $f(x)$ is the same function as before (see equation 25), and where:

$$P(x) = L_0\mathrm{E}\{[x - R_s(\tau)]^+|\mathcal{F}_0\} \tag{28b}$$

is the value of the ordinary receiver swaption with strike x. Thus, the CMS floorlets can also be priced through replication with vanilla receivers. Similarly, the value of a single CMS swap payment is:

$$V_{swap}^{CMS}(0) = D(t_p)R_s^0 + \frac{D(t_p)}{L_0}\left\{\int_{R_s^0}^\infty C(x)f''_{atm}(x)\,dx + \int_{-\infty}^{R_s^0} P(x)f''_{atm}(x)\,dx\right\} \tag{29a}$$

where:

$$f_{atm}(x) \equiv [x - R_s^0]\left(\frac{G(x)}{G(R_s^0)} - 1\right) \tag{29b}$$

is the same as $f(x)$ with the strike K replaced by the par swap rate R_s^0. Here, the first term in (29a) is the value if the payment were exactly equal to the forward swap rate R_s^0 as seen

today. The other terms represent the convexity correction, written in terms of vanilla payer and receiver swaptions. These too can be evaluated by replication.

It should be noted that CMS caplets and floorlets satisfy call-put parity. Since:

$$[R_s(\tau) - K]^+ - [K - R_s(\tau)]^+ = R_s(\tau) - K \quad \text{paid at } t_p \tag{30}$$

the payoff of a CMS caplet minus a CMS floorlet is equal to the payoff of a CMS swaplet minus K. Therefore, the value of this combination must be equal at all earlier times as well:

$$V_{cap}^{CMS}(t) - V_{floor}^{CMS}(t) = V_{swap}^{CMS}(t) - KZ(t;t_p) \tag{31a}$$

In particular:

$$V_{cap}^{CMS}(0) - V_{floor}^{CMS}(0) = V_{swap}^{CMS}(0) - KD(t_p) \tag{31b}$$

Accordingly, we can price an in-the-money caplet or floorlet as a swaplet plus an out-of-the-money floorlet or caplet.

Analytical formulas

The function $G(x)$ is smooth and slowly varying, regardless of the model used to obtain it. Since the probable swap rates $R_s(\tau)$ are heavily concentrated around R_s^0, it makes sense to expand $G(x)$ as:

$$G(x) \approx G(R_s^0) + G'(R_s^0)(x - R_s^0) + \cdots \tag{32a}$$

For the moment, let us limit the expansion to the linear term. This makes $f(x)$ a quadratic function:

$$f(x) \approx \frac{G'(R_s^0)}{G(R_s^0)}(x - R_s^0)(x - K) \tag{32b}$$

and $f''(x)$ a constant. Substituting this into our formula for a CMS caplet (27a), we obtain:

$$V_{cap}^{CMS}(0) = \frac{D(t_p)}{L_0}C(K) + G'(R_s^0)\left\{(K - R_s^0)C(K) + 2\int_K^\infty C(x)\,dx\right\} \tag{33}$$

where we have used $G(R_s^0) = D(t_p)/L_0$. Now, for any K the value of the payer swaption is:

$$C(K) = L_0 \mathrm{E}\{[R_s(\tau) - K]^+ | \mathcal{F}_0\} \tag{34a}$$

so the integral can be re-written as:

$$\begin{aligned}\int_K^\infty C(x)\,dx &= L_0 \mathrm{E}\left\{\int_K^\infty [R_s(\tau) - x]^+\,dx \,\middle|\, \mathcal{F}_0\right\} \\ &= \tfrac{1}{2} L_0 \mathrm{E}\{([R_s(\tau) - K]^+)^2 | \mathcal{F}_0\}\end{aligned} \tag{34b}$$

Putting this together yields:

$$V_{cap}^{CMS}(0) = \frac{D(t_p)}{L_0}C(K) + G'(R_s^0)L_0\mathrm{E}\{[R_s(\tau) - R_s^0][R_s(\tau) - K]^+|\mathcal{F}_0\} \tag{35a}$$

for the value of a CMS caplet, where the convexity correction is now the expected value of a quadratic "payoff". An identical arguments yields the formula:

$$V_{floor}^{CMS}(0) = \frac{D(t_p)}{L_0}P(K) - G'(R_s^0)L_0\mathrm{E}\{[R_s^0 - R_s(\tau)][K - R_s(\tau)]^+|\mathcal{F}_0\} \tag{35b}$$

for the value of a CMS floorlet. Similarly, the value of a CMS swap payment works out to be:

$$V_{swap}^{CMS}(0) = D(t_p)R_s^0 + G'(R_s^0)L_0\mathrm{E}\{(R_s(\tau) - R_s^0)^2|\mathcal{F}_0\} \tag{35c}$$

To finish the calculation, one needs an explicit model for the swap rate $R_s(\tau)$. The simplest model is Black's model, which assumes that the swap rate $R_s(\tau)$ is log normal with a volatility σ. With this model, one obtains:

$$V_{swap}^{CMS}(0) = D(t_p)R_s^0 + G'(R_s^0)L_0(R_s^0)^2[e^{\sigma^2\tau} - 1] \tag{36a}$$

for the CMS swaplets:

$$\begin{aligned} V_{cap}^{CMS}(0) = \frac{D(t_p)}{L_0}C(K) + G'(R_s^0)L_0[&(R_s^0)^2e^{\sigma^2\tau}\mathcal{N}(d_{3/2}) \\ &- R_s^0(R_s^0 + K)\mathcal{N}(d_{1/2}) + R_s^0K\mathcal{N}(d_{-1/2})] \end{aligned} \tag{36b}$$

for CMS caplets, and:

$$\begin{aligned} V_{floor}^{CMS}(0) = \frac{D(t_p)}{L_0}P(K) - G'(R_s^0)L_0[&(R_s^0)^2e^{\sigma^2\tau}\mathcal{N}(-d_{3/2}) \\ &- R_s^0(R_s^0 + K)\mathcal{N}(-d_{1/2}) + R_s^0K\mathcal{N}(-d_{-1/2})] \end{aligned} \tag{36c}$$

for CMS floorlets. Here:

$$d_\lambda = \frac{\ln R_s^0/K + \lambda\sigma^2\tau}{\sigma\sqrt{\tau}}. \tag{36d}$$

The key concern with Black's model is that it does not address the smiles and/or skews seen in the marketplace. This can be partially mitigated by using the correct volatilities. For CMS swaps, the volatility σ_{ATM} for at-the-money swaptions should be used, since the expected value (35) includes high and low strike swaptions equally. For out-of-the-money caplets and floorlets, the volatility σ_K for strike K should be used, since the swap rates $R_s(\tau)$ near K provide the largest contribution to the expected value. For in-the-money options, the largest contributions come from swap rates $R_s(\tau)$ near the mean value R_s^0. Accordingly, call-put parity should be used to evaluate in-the-money caplets and floorlets as a CMS swap payment plus an out-of-the-money floorlet or caplet.

Conclusions

The standard pricing for CMS legs is given by (36a–36d) with $G(R_s)$ given by (23a). These formulas are adequate for many purposes. When finer pricing is required, one can systematically improve these formulas by using the more sophisticated models for $G(R_s)$ developed in the Appendix, and by adding the quadratic and higher order terms in the expansion (32a). In addition, (35a–35b) show that the convexity corrections are essentially swaptions with "quadratic" payoffs. These payoffs emphasize away-from-the-money rates more than standard swaptions, so the convexity corrections can be quite sensitive to the market's skew and smile. CMS pricing can be improved by replacing Black's model with a model that matches the market smile, such as Heston's model or the SABR model. Alternatively, when the very highest accuracy is needed, replication can be used to obtain near perfect results.

Appendix A. Models of the yield curve

Model 1: Standard model

The standard method for computing convexity corrections uses bond math approximations: payments are discounted at a flat rate, and the coverage (day count fraction) for each period is assumed to be $1/q$, where q is the number of periods per year (1 for annual, 2 for semi-annual, etc). At any date t, the level is approximated as:

$$L(t) = Z(t, s_0) \sum_{j=1}^{n} \alpha_j \frac{Z(t, s_j)}{Z(t, s_0)} \approx Z(t, s_0) \sum_{j=1}^{n} \frac{1/q}{[1 + R_s(t)/q]^j} \tag{A.1}$$

which works out to:

$$L(t) = \frac{Z(t, s_0)}{R_s(t)} \left[1 - \frac{1}{(1 + R_s(t)/q)^n} \right] \tag{A.2a}$$

Here the par swap rate $R_s(t)$ is used as the discount rate, since it represents the average rate over the life of the reference swap. In a similar spirit, the zero coupon bond for the pay date t_p is approximated as:

$$Z(t; t_p) \approx \frac{Z(t, s_0)}{(1 + R_s(t)/q)^{\Delta}} \tag{A.2b}$$

where:

$$\Delta = \frac{t_p - s_0}{s_1 - s_0} \tag{A.2c}$$

is the fraction of a period between the swap's start date s_0 and the pay date t_p. Thus the standard "bond math model" leads to:

$$G(R_s) = \frac{Z(t; t_p)}{L(t)} \approx \frac{R_s}{(1 + R_s/q)^{\Delta}} \frac{1}{1 - \dfrac{1}{(1 + R_s/q)^n}} \tag{A.3}$$

This method (a) approximates the schedule and coverages for the reference swaption; (b) assumes that the initial and final yield curves are flat, at least over the tenor of the reference swaption; and (c) assumes a correlation of 100% between rates of differing maturities.

Model 2: "Exact yield" model

We can account for the reference swaption's schedule and day count exactly by approximating:

$$Z(t;s_j) \approx Z(t;s_0)\prod_{k=1}^{j}\frac{1}{1+\alpha_k R_s(t)} \tag{A.4}$$

where α_k is the coverage of the k^{th} period of the reference swaption. At any date t, the level is then:

$$L(t) = \sum_{j=1}^{n}\alpha_j Z(t;s_j) = Z(t;s_0)\sum_{j=1}^{n}\alpha_j\left(\prod_{k=1}^{j}\frac{1}{1+\alpha_k R_s(t)}\right) \tag{A.5}$$

We can establish the following identity by induction:

$$L(t) = \frac{Z(t;s_0)}{R_s(t)}\left(1-\prod_{k=1}^{n}\frac{1}{[1+\alpha_k R_s(t)]}\right) \tag{A.6}$$

In the same spirit, we can approximate:

$$Z(t;t_p) = Z(t;s_0)\frac{1}{(1+\alpha_1 R_s(t))^{\Delta}} \tag{A.7}$$

where $\Delta = (t_p - s_0)/(s_1 - s_0)$ as before. Then:

$$G(R_s) = \frac{Z(t;t_p)}{L(t)} \approx \frac{R_s}{(1+\alpha_1 R_s)^{\Delta}}\frac{1}{1-\prod_{k=1}^{n}\frac{1}{(1+\alpha_j R_s)}} \tag{A.8}$$

This approximates the yield curve as flat and only allows parallel shifts, but has the schedule right.

Model 3: Parallel shifts

This model takes into account the initial yield curve shape, which can be significant in steep yield curve environments. We still only allow parallel yield curve shifts, so we approximate:

$$\frac{Z(t;s_j)}{Z(t;s_0)} \approx \frac{D(s_j)}{D(s_0)}e^{-(s_j-s_0)x} \quad \text{for } j = 1,2,\ldots,n \tag{A.9}$$

where x is the amount of the parallel shift. The level and swap rate R_s are given by:

$$\frac{L(t)}{Z(t;s_0)} = \sum_{j=1}^{n} \alpha_j \frac{D(s_j)}{D(s_0)} e^{-(s_j - s_0)x} \tag{A.10a}$$

$$R_s(t) = \frac{D(s_0) - D(s_n)e^{-(s_n - s_0)x}}{\sum_{j=1}^{n} \alpha_j D(s_j) e^{-(s_j - s_0)x}} \tag{A.10b}$$

Turning this around:

$$R_s \sum_{j=1}^{n} \alpha_j D(s_j) e^{-(s_j - s_0)x} + D(s_n)e^{-(s_n - s_0)x} = D(s_0) \tag{A.11a}$$

determines the parallel shift x implicitly in terms of the swap rate R_s. With x determined by R_s, the level is given by:

$$\frac{L(R_s)}{Z(t;s_0)} = \frac{D(s_0) - D(s_n)e^{-(s_n - s_0)x}}{D(s_0)R_s} \tag{A.11b}$$

in terms of the swap rate. Thus this model yields:

$$G(R_s) = \frac{Z(t;t_p)}{L(t)} \approx \frac{R_s e^{-(t_p - s_0)x}}{1 - \dfrac{D(s_n)}{D(s_0)} e^{-(s_n - s_0)x}} \tag{A.12a}$$

where x is determined implicitly in terms of R_s by:

$$R_s \sum_{j=1}^{n} \alpha_j D(s_j) e^{-(s_j - s_0)x} + D(s_n)e^{-(s_n - s_0)x} = D(s_0) \tag{A.12b}$$

This model's limitations are that it allows only parallel shifts of the yield curve and it presumes perfect correlation between long- and short-term rates.

Model 4: Non-parallel shifts

We can allow non-parallel shifts by approximating:

$$\frac{Z(t;s_j)}{Z(t;s_0)} \approx \frac{D(s_j)}{D(s_0)} e^{-[h(s_j) - h(s_0)]x} \tag{A.13}$$

where x is the amount of the shift, and $h(s)$ is the effect of the shift on maturity s. As above, the shift x is determined implicitly in terms of the swap rate R_s via:

$$R_s \sum_{j=1}^{n} \alpha_j D(s_j) e^{-[h(s_j) - h(s_0)]x} + D(s_n)e^{-[h(s_n) - h(s_0)]x} = D(s_0) \tag{A.14a}$$

Then:

$$\frac{L(R_s)}{Z(t;s_0)} = \frac{D(s_0) - D(s_n)e^{-[h(s_n)-h(s_0)]x}}{D(s_0)R_s} \tag{A.14b}$$

determines the level in terms of the swap rate. This model then yields:

$$G(R_s) = \frac{Z(t;t_p)}{L(t)} \approx \frac{R_s e^{-[h(t_p)-h(s_0)]x}}{1 - \dfrac{D(s_n)}{D(s_0)} e^{-[h(s_n)-h(s_0)]x}} \tag{A.15a}$$

where x is determined implicitly in terms of R_s by:

$$R_s \sum_{j=1}^{n} \alpha_j D(s_j) e^{-[h(s_j)-h(s_0)]x} + D(s_n) e^{-[h(s_n)-h(s_0)]x} = D(s_0) \tag{A.15b}$$

To continue further requires selecting the function $h(s_j)$ which determines the shape of the non-parallel shift. This is often done by postulating a constant mean reversion:

$$h(s) - h(s_0) = \frac{1}{\kappa}[1 - e^{-\kappa(s-s_0)}] \tag{A.16}$$

Alternatively, one can choose $h(s_j)$ by calibrating the vanilla swaptions which have the same start date s_0 and varying end dates as their market prices.

FOOTNOTE

1. We follow the standard (if bad) practice of referring to both the physical instrument and its value as the "numeraire".

21
Mind the Cap

Peter Jäckel

Wilmott magazine, September 2003

Few of the readers will have missed the recent proliferation of articles on various aspects of the increasingly popular market models of interest rates. The reasons for this trend are easy to see: market models allow traders to design the risk-neutral volatility functions and correlations for their exotic pricing models as close as they wish to the real-world structure of uncertainty they can see in the market-observables. Despite the fact that risk-neutral model parameters are, from a purely theoretical point of view, not really required to be similar to the real world behaviour, it is intuitively clear that the capturing of fundamental features of real-world dynamics in any given model process will lead to more realistic and thus stable hedge ratios. A poignant example for this is the characteristically unimodal evolution of both realised[1] and implied volatility of any given caplet: having undergone a long and slow rise, just before the caplet's expiry those volatility figures tend to decrease noticeably. Not surprisingly, any trader hedging some exposure to caplet volatilities using the underlying futures contract would like his quants to design the modelling framework to take this into account. Another reason for the plethora of work lately published on market models is the progress made both by computer hardware manufacturers and by practitioners' Monte Carlo techniques. The framework for distributed calculations of simulations using additional variance reduction techniques is more and more readily implemented in all the major derivatives houses, and specifically for the Libor market model, fast drift approximations that obviate the need for short-stepped Euler schemes are available (Hunter, Jäckel and Joshi, 2001; Pietersz, Pelsser and van Regenmortel, 2002). What's more, with the recent developments of algorithms that allow for the approximate pricing of products that depend on the exercise strategy of the investor such as Bermudan swaptions (Longstaff and Schwartz, 1998; Andersen, 1999; Andersen and Andreasen, 2000; Jäckel, 2002), market models have now become the method of choice for the pricing of complex interest rate derivatives.

All of these developments created an ever more urgent need for fast calibration procedures for the Libor market model that are viable in a production environment. At the heart of any fast calibration procedure is an analytical or semi-analytical pricing formula for the given calibration instruments. Since the Libor market model reprices the canonical caplets by

Contact-address: ABN AMRO, 250 Bishopsgate, London ECZM 4AA, UK.
E-mail: p@jaeckel.org

[1] Measured over a suitable window in time.

construction, it is only natural that so far most of the attention for analytical approximations of other market instruments has been on swaptions, and some very impressive formulae have been found (Rebonato, 1999; Jäckel and Rebonato, 2000; Hull and White, 2000; Schoenmakers, 2000). However, in practice it is also important to be able to calculate (semi-)analytical prices for all the possible caplets, not just those that coincide in expiry and payment date as well as accrual period with the abstract discretisation of the yield curve used within the model's own discretised framework. For instance, a given Libor market model implementation may be based on 1-monthly discrete forward rates, but we may wish to calibrate to caplet prices for contracts on 3-month Libor rates. All of the existing swaption approximations work well whenever there is a significant averaging effect due to the swap rate being effectively a weighted sum of all of the discrete forward rates. In contrast, when a non-canonical rate depends only on a small number of the discrete forward rates, or when the payment frequency of the fixed side of a swap does not match the floating side exactly, the known low-order approximations start to break down, and higher order corrections are required. A 3-month caplet that is to be composed from 1-monthly forward rates is such an example. Another, and probably more important case is the value of a caplet on a 6-month forward Libor rate in a model framework of 3-monthly discrete forward rates, or even an option on a 12-month rate. These situations require some kind of basket approximation, and, ideally, the method should allow for some sort of implied volatility skew to be embedded in the model, and it should take into account the possible funding spread difference between 3-monthly and 6-monthly (or 12-monthly, respectively) Libor rates a firm may be subject to. The obvious application of the latter two features is the Japanese interest rate market where not only the caplet skew is far too pronounced to be ignored, but where in addition to all other complications the fact that most US and European investment houses fund at a rate that is *lower* than Yen-Libor leads to significant pricing implications.

1 A simple Libor market model with a skew

There are many methods to incorporate a skew in a Libor market model. Examples include the constant elasticity of variance model (Andersen and Andreasen, 2000), quadratic volatility specifications (Zühlsdorff, 2002), and jump-diffusion processes (Glassermann and Merener, 2003). For reasons of simplicity, I choose the *displaced diffusion* setup (Rubinstein, 1983) which is also known as *affine volatility* (Zühlsdorff, 2002). In this framework, the discrete forward rates evolve according to the stochastic differential equation:

$$\frac{d(f_i + s_i)}{f_i + s_i} = \mu_i(\boldsymbol{f}, \boldsymbol{s}, t)\, dt + \sigma_i(t)\, dW_i \tag{1}$$

with some constant shift s_i associated with the forward rate f_i. Equation (1) describes the stochastic evolution of geometric Brownian motion for the quantity $(f_i + s_i)$ with instantaneous deterministic volatility $\sigma_i(t)$ and instantaneous indirectly stochastic drift $\mu_i(\boldsymbol{f}(t), \boldsymbol{s}, t)$. This feature will be important later when we approximate the drift as a constant and thus render the expression $(f_i + s_i)$ as a lognormal variate.

1.1 The skew parametrisation

Any given forward rate is drift-free in its own natural measure, i.e:

$$d(f + s) = df = \sigma(t)(f + s)\, dW \tag{2}$$

Since it has become more and more common practice to express the volatility-rate dependence as some equivalent constant elasticity-of-variance parameter (Karasinski, 2002; Hagan, Kumar and Lesniewski, 2002), it is desirable to find a scaling for the skew that allows us to specify the proximity to the lognormal or normal volatility setting directly in a similar fashion, albeit in a somewhat approximate way. One such possible parametrisation is to replace the term $(f + s)$ on the right hand side of equation (2) by $(f \cdot q + f(0)(1 - q))$ for some constant q, i.e:

$$df = \sigma_q \cdot [q \cdot f + (1 - q) \cdot f(0)] \cdot dW \tag{3}$$

This approach allows the continuous transition from the lognormal framework for $q = 1$ to the normal model first introduced by Bachelier (1900) at $q = 0$. However, this kind of parametrisation has two practical drawbacks. Firstly, the end-users of any model tend to explore the available parameter scales in a rather indiscriminate fashion in order to achieve the skew they desire to model. For $q < 0$, the above parametrisation unfortunately results in a shifted lognormal distribution *with inverted asymmetry* stretching from $-\infty$ to $f_{\max} = f(0) \cdot (\frac{q-1}{q})$. In other words, it predicts that the forward rate will at expiry not exceed a certain positive threshold $f_{\max} > f(0)$, but may take on any negative value with potentially quite considerable probability. One of the shortcomings of the extended Vasicek model, in comparison, that the Libor market model is frequently used to remedy for, is that it allows for negative forward rates with an approximately normal distribution. It therefore seems natural to impose a skew limitation at the point where the q parametrisation meets the Bachelier model, or possibly even before. This leads to an alternative parametrisation in a new skew parameter Q that gently approaches the normal model as $Q \to 0$ but requires $Q \in (0, 2)$ by virtue of the following definition:

$$Q := 2^{-\frac{s}{f(0)}} \tag{4}$$

or, equivalently:

$$s := -f(0) \log_2 Q \tag{5}$$

which leads to:

$$df = \sigma_Q \cdot [f - \log_2 Q \cdot f(0)] \cdot dW \tag{6}$$

The transformation from the (q, σ_q) to the (Q, σ_Q) parametrisation is given by:

$$\begin{aligned} Q &= 2^{\frac{q-1}{q}} \\ \sigma_Q &= q \cdot \sigma_q \end{aligned} \tag{7}$$

It is easy to see that this parametrisation is equivalent to the former at the three most important points: the lognormal model is in both cases given by $q = Q = 1$, the approximation for the square root model is given at $q = Q = \frac{1}{2}$, and the normal (Bachelier) model is given by $q = 0$ and approximated in the limit of $Q \to 0$ but never quite reached. The feature of the Q parametrisation only being able to approach the Bachelier model in the limit of $Q \to$ in a very explicit fashion is, rather subtly, shared by but somewhat disguised in the q parametrisation. In

fact, for the q parametrisation and the Q encoding of the skew alike, analytical formulæ as well as Monte Carlo schemes based on approximations to the transfer densities need a distinctive switch to the Bachelier framework as q, or Q, for that matter, vanish, since the transition from displaced lognormal to normal distributions undergoes a singular change at $q = Q = 0$ in a similar fashion as $\int x^{q-1}\,dx$ switches structurally from $\frac{x^q}{q}$ to $\ln x$ at $q \equiv 0$.

It should be pointed out that the discussion in the following sections equally holds regardless of whether one prefers the q or the Q parametrisations outlined above since both of them result in volatility specifications of *affine* nature and are thus equivalent to the displaced diffusion equation (1). The choice of parametrisation does, in practice, though make a difference to the user-friendliness of a given model, and, in my experience, the limitation of a control parameter such as the skew coefficient Q, or q, respectively, to a *finite* interval tends to be more intuitive. The constraints of the skew control coefficient are directly related to the range of the skew that we want to allow for, and this is elaborated in the next section.

1.2 The skew range

In order to establish whether the restriction $Q > 0$ poses in practice any noticeable limitation, let us define the skew χ as the change in implied volatility incurred at the money as the strike is varied by one $\frac{1}{10}$-th of the forward, i.e:

$$\chi := \left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=f} \cdot \frac{f}{10} \tag{8}$$

Since the implied volatility $\hat{\sigma}$ relates to the price given in the limit of $Q \to 0$ via the Black and the Bachelier pricing formulæ, respectively, we have:

$$V_{\text{Black}}(f, K, \hat{\sigma}, T) = V_{\text{Bachelier}}(f, K, \hat{\sigma}_{\text{Bachelier}}, T) \tag{9}$$

and thus:

$$\begin{aligned} &\left.\frac{\partial V_{\text{Black}}(f, K, \hat{\sigma}, T)}{\partial K}\right|_{K=f} + \left.\frac{\partial V_{\text{Black}}(f, K, \hat{\sigma}, T)}{\partial \hat{\sigma}}\right|_{K=f} \cdot \left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=f} \\ &\quad = \left.\frac{\partial V_{\text{Bachelier}}(f, K, \hat{\sigma}_{\text{Bachelier}}, T)}{\partial K}\right|_{K=f}. \end{aligned} \tag{10}$$

Since the right hand side of equation (10) is exactly given by $\frac{-1}{2}$ times the discount factor to the payment date, this leads to:

$$\left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=f} = \frac{\mathrm{N}\left(\frac{-\hat{\sigma}\sqrt{T}}{2}\right) - \frac{1}{2}}{f\sqrt{T}\varphi\left(\frac{\hat{\sigma}\sqrt{T}}{2}\right)} \tag{11}$$

with $\varphi(x) = dN(x)\,/dx = \mathrm{e}^{-\frac{1}{2}x^2} \Big/ \sqrt{2\pi}$. A first-order expansion of the numerator and denominator in $\hat{\sigma}\sqrt{T}$ gives us the approximate rule:

$$\chi_{\text{Bachelier}} \simeq -\frac{\hat{\sigma}}{20} \tag{12}$$

In other words, for implied volatilities around 20%, the maximum (negative) attainable skew of the displaced diffusion model is approximately 1% which is usually more than sufficient for caplets. What's more, the markets that require a stronger skew calibration such as Japan tend to have significantly higher volatilities and this means that the displaced diffusion approach can be calibrated to the skew prevailing there, too.

Of course, it is arguable whether one should allow for negative interest rates at all in any Libor market model. In this context it is helpful to note that the at-the-money-forward skew required in most major interest rate markets for most maturities is considerably less strong than predicted by the normal, or equivalently, extended Vasicek model (Vasicek, 1977; Hull and White, 1990). The probabilities of negative interest rates are thus even smaller than in the Hull–White or extended Vasicek model, and should therefore in practice be of no concern. The negativity of rates would be completely suppressed in a CEV modelling framework as suggested in Andersen and Andreasen (2000). However, the CEV framework suffers from one major drawback: for most market-calibrated parameters especially long-dated forward rates incur a rather too large probability of absorption at zero. This may seem innocuous in comparison to the possibility of stochastic paths to spend some time in the negative domain. However, the absorption feature makes the whole concept of pricing in a risk-neutral measure rather questionable since it jeopardises the existence of an equivalent martingale measure (Platen, 2002). On the other hand, some people might argue that effective Libor rates should actually be allowed to become temporarily slightly negative, although this line of reasoning almost inevitably leads to a debate based on economic grounds that is of no particular relevance here.

Nevertheless, should one desire to adjust the displaced diffusion framework not to allow for negative rates at the expense of an absorbing boundary at zero, it is indeed possible to include such a boundary condition for the affine volatility specification of the displaced diffusion model, and still obtain a very simple closed form solution for options on canonical caplets (Zühlsdorff, 2002). Also, the incorporation of an absorbing boundary at zero poses no problem to any Monte Carlo simulation whatsoever. However, as shown in Zühlsdorff (2002), unless we are concerned with calibration to extremely far out-of-the-money floorlets, the distinction between the displaced diffusion setup with and without absorbing barrier at zero makes no practical difference for implied volatilities whence I neglect this issue in the following.

One of the attractive features of the stochastic differential equation (1) is that it not only allows for a negative skew, but also for a positive dependence of implied volatilities on the strike. In particular for calibration at high strikes, the positive skew observed in the market poses frequently a severe problem for HJM models based on a quasi-Gaussian evolution of the forward rates such as the extended Vasicek or Hull-White model. Whilst it is usually still possible to calibrate those models at any given strike, the implied risk-neutral distribution as given by a quasi-Gaussian forward rate evolution differs significantly from the distribution as implied by the market's smile (Breeden and Litzenberger, 1978) which can give rise to substantial pricing differences if, for instance, an exotic contract is valued that contains any form of forward rate related digital features.

In analogy to the analysis and expansion that led to the expression for the skew in the Bachelier model given by equation (12), we arrive at the following approximation for the Q skew:

$$\chi_Q \simeq \frac{\hat{\sigma}}{20} \cdot \frac{\log_2 Q}{1 - \log_2 Q} \simeq \frac{\hat{\sigma}}{20} \cdot (q - 1) \tag{13}$$

This formula requires that the parameter Q must be in the interval (0, 2). In fact, for $Q \to 2$, the skew expression diverges. This effect can be understood better if we have a look at the risk-neutral densities shown in Figure 1. As we can see, for $Q \gtrsim 3/2$, the density becomes more and more peaked. In fact, in the limit of $Q \to 2$, the density approaches a Dirac distribution. This feature of the displaced diffusion equations for $Q > 1$ bears consequences for any Monte Carlo simulation: when the density is strongly peaked and has a very long but thin tail, the simulation converges rather poorly since most variates are drawn in the area of the peak, and only very few fall in the tail. In this case, it may be advisable to employ importance sampling or *sampler density* (Jäckel, 2002) techniques that lay much more emphasis on the long tail and thus improve convergence considerably. In practice, I would recommend not to use values of Q greater than $\frac{3}{2}$.

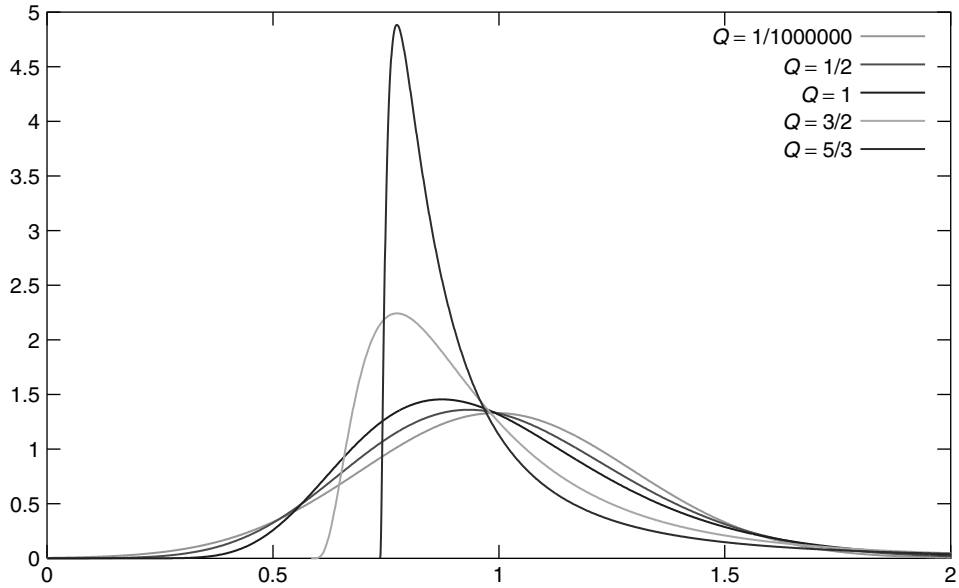


Figure 1: The forward rate density for different levels of the skew coefficient Q with $T = 1$, $\sigma_Q = 30\%/(1 - \log_2 Q)$, and $f_0 = 1$

1.3 The drift conditions in the displaced diffusion framework

Following the convention that the canonical discrete forward rate f_i with associated accrual factor τ_i fixes at time t_i, and that the chosen numéraire is given by a zero coupon maturing at

t_N, the drift conditions for the forward rates subject to the stochastic differential equation (1) are:

$$\mu_i(\boldsymbol{f}(t), t) = -\sigma_i \underbrace{\sum_{k=i+1}^{N-1} \frac{(f_k(t) + s_k)\tau_k}{1 + f_k(t)\tau_k} \sigma_k \rho_{ik}}_{\text{non-zero for } i<N-1} + \sigma_i \underbrace{\sum_{k=N}^{i} \frac{(f_k(t) + s_k)\tau_k}{1 + f_k(t)\tau_k} \sigma_k \rho_{ik}}_{\text{non-zero for } i \geq N} . \tag{14}$$

1.4 Interpolating Libors from canonical discrete forward rates

It is common to highlight the fundamental features of Libor market models using the example of interest rate products that depend only on cashflows occurring precisely on dates coinciding with the model's yield curve discretisation. In practice, however, a Libor market model implementation has to cope with many intermediate cashflows, with settlement delays, fixing conventions, and many other idiosyncracies of the fixed income market. This means that it may be necessary to compute discount factors that span several canonical periods, potentially with a stub discount factor covering only part of the associated discrete forward rate's accrual period. An example for this is given in Figure 2. It is difficult to construct non-canonical discount factors from a given set of discrete forward rates in a completely arbitrage-free manner. However, in practice, it is usually sufficient to choose an approximate interpolation rule such that the residual error is well below the levels where arbitrage could be enforced. It is also important to remember that the numerical evaluation of any complex deal with a Libor market model is ultimately still subject to inevitable errors resulting from the calculation scheme: Monte Carlo simulations, non-recombining trees, or recombining trees with their own drift approximation problems. In this context it may not be surprising that the following discount factor interpolation approach is highly accurate for practical purposes.

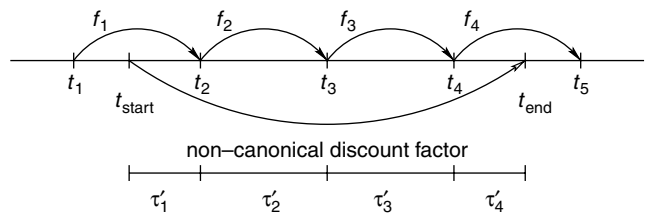


Figure 2: A non-canonical discount factor and its decomposition into canonical forward rates

Given any forward discount factor $P(t; t_{\text{start}}, t_{\text{end}})$ at time $t \leq t_{\text{start}} < t_{\text{end}}$ that represents the forward funding cost of borrowing one currency unit at time t_{start} and paying back $1/P(t; t_{\text{start}}, t_{\text{end}})$ at time t_{end}, we compute $P(t; t_{\text{start}}, t_{\text{end}})$ from the discrete forward rates according to:

$$P(t; t_{\text{start}}, t_{\text{end}}) = \prod_i (1 + f_i(t)\tau'_i)^{-1} \tag{15}$$

The product on the right hand side of equation (15) is hereby over all the forward rates that are partly or completely spanned by the discount factor period $(t_{\text{start}}, t_{\text{end}})$. The modified accrual

factors τ_i' reflect the potentially partial coverage at either end of the period as depicted in Figure 2 where both τ_1' and τ_4' are smaller than τ_1 and τ_4, respectively. When a firm's funding cost happens to be given directly by (forward) Libor rates for a given period τ as they are observed in the market, the relationship between the Libor rate $L(t; t_{\text{start}}, t_{\text{start}} + \tau)$ and the discount factor over the associated accrual period is:

$$L(t; t_{\text{start}}, t_{\text{start}} + \tau) = \left(\frac{1}{P(0; t_{\text{start}}, t_{\text{start}} + \tau)} - 1 \right) \Big/ \tau \tag{16}$$

In other words, using the decomposition (15) into canonical forward rates, we have:

$$1 + L_\tau \cdot \tau = \prod_i (1 + f_i(t)\tau_i') \tag{17}$$

where I have dropped the explicit mentioning of the dependence on t and t_{start}. In the following, I shall assume that the yield curve is sufficiently smooth in between canonical forward rate dates to justify the simple accrual factor adjustment of the stub periods at either end of the Libor rate accrual interval akin to discrete rate interpolations customary in the short dated money markets. However, it is straightforward to add an additional Libor rate correction factor γ_τ by setting:

$$L_\tau \cdot \tau = \gamma_\tau \cdot \left[\prod_i (1 + f_i(t)\tau_i') - 1 \right] \tag{18}$$

with:

$$\gamma_\tau = \frac{L_\tau(0) \cdot \tau}{\prod_i (1 + f_i(0)\tau_i') - 1} \tag{19}$$

which would correct the Libor rate exactly in the limit of vanishing volatilities. Equation (17) will form the basis of the analytical valuation of non-canonical caplets. First, however, let us have a look at yet another interesting feature of the fixed income market: the spread between funding and interbank offered rates.

1.5 Spread differentials

Most of the major investment houses fund their cash requirements in the Euro, Dollar, and Sterling markets at rates that are very close to the official interbank offered rates. After all, it is precisely this interbank borrowing and lending for funding purposes that originally gave rise to the introduction of the Interbank Offered Rates (IBOR) quotation averages. Some financial institutions, however, have the privilege of higher-than-average credit ratings, and fund themselves accordingly somewhat more cheaply, and others can only borrow at less advantageous rates. In the Yen market, for instance, this phenomenon is particularly pronounced where most of the Western investment banks fund significantly more cheaply than IBOR. There are several ways to incorporate such a spread between funding and IBOR rates into a market model. In the following, I present a simple procedure based on adjustment factors that are structurally similar to discount factors.

Let us assume that we are building a Libor market model that is based on a 3-monthly canonical forward rate discretisation of the yield curve. In this framework, it may be desirable

to be able to price options on forward rate agreements that happen to fall precisely on the canonical dates by a straightforward application of Black's formula and a multiplication by a funding discount factor. In other words, for all forward rates' displacement coefficients Q_i being exactly unity, we may wish to see no skew for such canonical caplet prices struck at different levels. In order to accomplish a setup that allows for spreads, and indeed for spread differentials since the spread between funding rates and 3 month Libor rates may be different than the spread between funding and 6 month rates, I define the (forward) *Libor equivalent discount factor*:

$$\tilde{P}(t; t_{\text{start}}, t_{\text{start}} + \tau) = \frac{1}{1 + L_\tau(t; t_{\text{start}}) \cdot \tau} \tag{20}$$

Funding discount factors P and Libor equivalent discount factors $\tilde{P}$ are related by virtue of a deterministic *spread factor*, i.e:

$$\tilde{P}(t; t_{\text{start}}, t_{\text{start}} + \tau) = P(t; t_{\text{start}}, t_{\text{start}} + \tau) \cdot \zeta_\tau(t_{\text{start}}, t_{\text{start}} + \tau) \tag{21}$$

The spread factor $\zeta_\tau(t_{\text{start}}, t_{\text{start}} + \tau)$ is less than unity whenever funding can be done at a more favourable rate than Libor. Since the spread factor is effectively a credit spread discount factor that represents a simplified amalgamation of default hazard rates into a single number, it is decreasing in the accrual period τ. The decomposition of (forward) funding discount factors now becomes:

$$\zeta_{\tau^*}(t_{\text{start}}, t_{\text{end}}) \cdot P(t; t_{\text{start}}, t_{\text{end}}) = \prod_i (1 + f_i(t)\tau_i')^{-1} \tag{22}$$

where τ^* stands for the model's canonical discretisation period. All Libor rates that are not for a period that is equal to τ^* can then be computed indirectly via the funding discount factors. This yields:

$$L_\tau \cdot \tau = \frac{\zeta_{\tau^*}}{\zeta_\tau} \cdot \prod_i (1 + f_i(t)\tau_i') - 1 \tag{23}$$

wherein both ζ_{τ^*} and ζ_τ are of course to be taken over the accrual period of the Libor rate L_τ. For $\tau \neq \tau^*$, i.e. when we are interested in a Libor rate that is based on an accrual period different from the model's intrinsic discretisation period, in the presence of a spread differential of the spread between funding and τ-Libor versus the spread between funding and τ^*-Libor, the multiplicative spread ratio term $\frac{\zeta_{\tau^*}}{\zeta_\tau}$ on the right hand side of equation (23) gives rise to a *spread differential induced skew*, as we will see in the following.

2 Analytical caplet valuation

The analytical valuation of a caplet[2] is based on the evaluation of the expectation:

$$\mathrm{E}[(L \cdot \tau - \mathrm{K} \cdot \tau)_+] \tag{24}$$

[2] I restrict the discussion to caplets. The translation to floorlets is, naturally, straightforward, and should not pose a problem to the reader if I succeed in my attempt to make the exposition of the case of a caplet sufficiently clear.

2.1 First order approximation ignoring the drift

For τ^* not too large, and for moderate interest rates, a Taylor expansion of the product on the right hand side of equation (23) is an obvious approach:

$$L \cdot \tau = (\delta - 1) + \delta \cdot \sum_i f_i \tau_i' + \mathcal{O}\left((f_i \tau_i')^2\right) \quad \text{with} \quad \delta := \frac{\zeta_{\tau^*}}{\zeta_\tau} \tag{25}$$

The right hand side of expansion (25) is $(\delta - 1)$ plus a sum of displaced lognormals. In other words, we have a constant term plus a sum of correlated lognormal variates. Now, taking into account the displacements s_i, let us define:

$$\gamma := \frac{L(0) \cdot \tau + 1 - \delta}{\delta \cdot \sum_i f_i(0) \cdot \tau_i'} \tag{26}$$

$$x_i := \gamma \cdot \delta \cdot (f_i + s_i) \cdot \tau_i' \tag{27}$$

$$\kappa := K \cdot \tau + 1 - \delta + \gamma \cdot \delta \cdot \sum_i s_i \cdot \tau_i' \tag{28}$$

This enables us to write the first order approximation for (24) as:

$$\mathrm{E}\left[\left(\sum_i x_i - \kappa\right)_+\right] \tag{29}$$

Note that the scaling factor γ was introduced to ensure that the (undiscounted) forward contract $\mathrm{E}[(\sum_i x_i - \kappa)]$ is priced exactly.

Ignoring the fact that most of the involved forward rates are not drift-free in the terminal payment measure of the caplet, we can evaluate (29) as a basket option on a linear combination of lognormal variates x_i with individual expectations $x_i(0)$ struck at κ. This means we have now reduced the first order caplet approximation to the calculation of the expectation (29) where the x_i are lognormal variates with expectations:

$$\mathrm{E}[x_i] = x_i(0) = \gamma \cdot \delta \cdot (f_i(0) + s_i) \cdot \tau_i' \tag{30}$$

and log-covariances:

$$\mathrm{E}[\ln x_i \cdot \ln x_j] - \mathrm{E}[\ln x_i] \cdot \mathrm{E}[\ln x_j] = c_{ij} = \int_0^{T_{\text{expiry}}} \sigma_i(t)\sigma_j(t)\rho_{ij}(t)\, dt \tag{31}$$

There are many methods for the approximation of basket options such as Mike Curran's excellent geometric conditioning approach (Curran, 1994), the matching of two moments to a lognormal distribution (Levy, 1992), the matching of three moments to a Johnson distribution (which, incidentally, is the distribution resulting from a displaced diffusion), the method by Turnbull and Wakeman (1991), or Taylor expansion approaches (Ju, 2001; Reiner, Davidov

and Kumanduri, 2001). For the specific case here, however, the particularly fast *rank reduction* method lends itself readily since we can take advantage of the fact that all of the involved forward rates are typically very strongly positively correlated. This method is based on an analysis of the pricing of options on baskets of perfectly correlated lognormally distributed coupons that arises in a single factor extended Vasicek modelling environment (Jamshidian, 1989) and is detailed in the Appendix. The rank reduction method works extremely well when correlations are moderate to high, volatilities are at similar levels, and the expectations of the constituents of the basket are also of comparative magnitude. All of these criteria are satisfied by the basket option problem at hand in equation (29). In addition, the rank reduction method is very fast indeed and particularly easy to implement, and all of this is why it is the designated method of choice for the caplet approximation.

2.2 Second order approximation with drift estimate

Let us denote the number of forward rates that contribute to the value of the caplet based on the non-canonical Libor rate L as m. Extending the expansion of the Libor decomposition (23) to second order, we obtain:

$$L \cdot \tau = (\delta - 1) + \delta \cdot \sum_{i=1}^{m} f_i \tau_i' + \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i=1} f_i \tau_i' f_j \tau_j' + \mathcal{O}\left((f_i \tau_i')^3\right) \tag{32}$$

This time, it is not immediately obvious how we can substitute the expansion (32) into the caplet pricing formula (24) and treat the resulting expectation as a basket option on a sum of correlated lognormal variates. However, rewriting the second order expansion (32) as:

$$\begin{aligned}
L \cdot \tau \simeq\ & (\delta - 1) + \delta \cdot \sum_{i=1}^{m} \eta(f_i + s_i)\tau_i' - \delta \cdot \sum_{i=1}^{m} \eta s_i \tau_i' + \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta(f_i + s_i)\tau_i' \\
& \times \eta(f_j + s_j)\tau_j' - \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta f_i \tau_i' \eta s_j \tau_j' - \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta s_i \tau_i' \eta f_j \tau_j' \\
& - \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta s_i \tau_i' \eta s_j \tau_j' \\
=\ & (\delta - 1) - \delta \cdot \sum_{i=1}^{m} \eta S_i \tau_i' + \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta S_i \tau_i' \eta S_j \tau_j' \\
& + \delta \cdot \sum_{i=1}^{m} \left(1 + \eta S_i \tau_i' - \sum_{j=1}^{m} \eta S_j \tau_j'\right) \eta(f_i + S_i)\tau_i' \\
& + \delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{i-1} \eta(f_i + S_i)\tau_i' \eta(f_j + S_j)\tau_j'
\end{aligned} \tag{33, 34}$$

with some constant scaling coefficient η (that is to be determined later) provides some insight. The terms on the right hand side of equation (34) form three groups. The first group consists of all the constant terms on the first line of the right hand side. If we approximate the drift conditions (14) for the forward rates by a constant expression, we can treat the second group as a sum of lognormal variates as it comprises only terms of the form *constant* $\cdot (f_i + s_i)$. The last group is then a sum of bilinear combinations of lognormal variates, and this is where we can take advantage of a feature of the lognormal distribution: products of lognormals are again lognormally distributed, and we can compute their expectations and covariances with the original set of lognormals analytically!

Before we proceed to the calculation of the covariances of all the linear and bilinear terms, though, we ought to remember that particularly for caplets on accrual periods that are significantly longer than the model's intrinsic discretisation period, the risk-neutral drift of the involved discrete forward rates is no longer entirely negligible. Choosing the numéraire given by a zero coupon bond that pays one currency unit at the end of the (potentially truncated) accrual period of the last involved discrete forward rate, i.e. at $t_m + \tau'_m$ in our previous notation, we arrive at the following constant drift approximation:

$$\mathrm{E}_{T_{\text{expiry}}}[(f_i + s_i)] \approx (f_i(0) + s_i) \cdot \prod_{j=i+1}^{m} e^{-\frac{(f_j(0)+s_j)\tau'_j}{1+f_j(0)\tau'_j} c_{ij}} \tag{35}$$

There are, of course, a whole series of rather ad-hoc assumptions in equation (35). As we know, the drift of the discrete forward rates is neither constant nor deterministic[3] due to its instantaneous dependence on the forward rates that bridge the gap between the payment time of any one forward rate and the numéraire asset. This means wherever we have used the initial values for the forward rates in equation (35) we are both using the wrong value to represent the path-average for the evolution of the forward rates (since we are using the initial value), and we are ignoring the indirect stochasticity of the drift since we are using a constant value for each and every forward rate. In my experience, the suppression of the variance of the drift term due to the stochasticity of the forward rates is typically the dominant error in the constant drift expression. As the drift term is in the exponent, it is Jensen's inequality that is raising its head here. Ignoring the variability of the forward rates in the expression $\frac{(f_j(0)+s_j)\tau'_j}{1+f_j(0)\tau'_j}$ leads to a much bigger discrepancy than the fact that we are ignoring the drift or path-average for f_j when we replace it by a constant value. This phenomenon is reasonably well understood and has led to the development of highly accurate stepwise drift approximations that enable us to construct Monte Carlo schemes that do not need short time steps as we would with the Euler method (Hunter, Jäckel and Joshi, 2001; Pietersz, Pelsser and van Regenmortel, 2002). For our caplet calculations, however, this effect is fortunately quite small. Still, we can try to correct for it to some extent by the approximation that each of the terms $(f_j + s_j)$ is almost lognormally distributed, i.e:

$$(f_j + s_j) \approx (f_j(0) + s_j) e^{-\frac{1}{2}c_{jj} + \sqrt{c_{jj}} \cdot z_j} \qquad \text{with} \qquad z_j \sim \mathcal{N}(0, 1) \tag{36}$$

In this way, we can expand each of the terms in the product of the right hand side of equation (35) individually in c_{jj} and integrate over an independent normal standard normal

[3] The only exception is, of course, the one forward rate that pays at the same time as the numéraire.

distribution for z_j, i.e:

$$\begin{aligned}&\frac{-(f_j+s_j)\mathrm{e}^{-\frac{1}{2}c_{jj}+\sqrt{c_{jj}}\cdot z_j}}{\mathrm{e}^{1+(f_j+s_j)\mathrm{e}^{-\frac{1}{2}c_{jj}+\sqrt{c_{jj}}\cdot z_j}-s_j}}c_{ij}\\ &\quad\approx \mathrm{e}^{\frac{-(f_j+s_j)}{1+f_j}c_{ij}}\cdot\left(1+\frac{(f_j+s_j)^2(1-s_j)((1-s_j)c_{ij}+2(1+f_j))c_{ij}}{2(1+f_j)^4}c_{jj}\right)+\mathcal{O}(c_{jj}^2)\end{aligned} \tag{37}$$

where I have suppressed the modified accrual factors and dropped all initial value $\cdot(0)$ notation for clarity. Let us now define the approximate expectation for the displaced forward rate using the above expansions as:

$$\mathrm{E}_{T_{\text{expiry}}}[(f_i+s_i)]\approx e_i \tag{38}$$

with:

$$\begin{aligned}e_i &:= (f_i(0)+s_i)\cdot\prod_{j=i+1}^{m}\mathrm{e}^{-\frac{(f_j(0)+s_j)\tau_j'}{1+f_j(0)\tau_j'}c_{ij}}\\ &\quad\cdot\left(1+\frac{(f_j\tau_j'+s_j\tau_j')^2(1-s_j\tau_j')((1-s_j\tau_j')c_{ij}+2(1+f_j\tau_j'))c_{ij}}{2(1+f_j\tau_j')^4}c_{jj}\right)\end{aligned} \tag{39}$$

I now turn the attention to the earlier introduced scaling coefficient η. In analogy to the scaling coefficient γ that we used in the lower order approximation, η is supposed to ensure that our analytical approximation will return the correct expectation of forward rate agreements exactly. To compute η, we need the expectation of all the terms on the right hand side of equation (34). Rearranging the resulting terms as coefficients of a quadratic expression in η, we obtain:

$$\mathrm{E}[L\cdot\tau]=(\delta-1)+\alpha_1\cdot\eta+\alpha_2\cdot\eta^2 \tag{40}$$

with

$$\alpha_1=\delta\cdot\sum_{i=1}^{m}(e_i\tau_i'-s_i\tau_i') \tag{41}$$

$$\alpha_2=\delta\cdot\sum_{i=1}^{m}\sum_{j=1}^{i-1}(s_i\tau_i's_j\tau_j'+e_i\tau_i'e_j\tau_j'e^{c_{ij}})+\delta\cdot\sum_{i=1}^{m}e_i\tau_i'\left(s_i\tau_i'-\sum_{j=1}^{ms}s_j\tau_j'\right) \tag{42}$$

Naturally, the solution for η that will ensure the correct value for forward rate agreements within our analytical approximations is:

$$\eta=\begin{cases}\dfrac{\alpha_1}{2\alpha_2}\left(\sqrt{1+\dfrac{4\alpha_2}{\alpha_1^2}[L(0)\cdot\tau+1-\delta]}-1\right) & \text{for}\quad \alpha_2\neq 0\\ \dfrac{1}{\alpha_1}[L(0)\cdot\tau+1-\delta] & \text{for}\quad \alpha_2=0\end{cases} \tag{43}$$

We now have almost all the components that we need to put together an approximate caplet valuation formula based on the rank reduction method applied to an option on the basket of lognormal variates. Since we have a second order expansion of equation (23), the vector of lognormal variates with expectation $\boldsymbol{\xi}$ will in total have:

$$N := \frac{m(m+1)}{2} \tag{44}$$

elements of which the first m account for the first order terms, and the remaining $\frac{m(m-1)}{2}$ result from the bilinear combinations. The individual expectations are given by:

$$\xi_k = \begin{cases} \delta\eta e_k \tau_k' \left(1 - \sum_{j=1, j\neq k}^{m} \eta s_j \tau_j'\right) & \text{for} \quad k \leq m \\ \delta\eta^2 e_i \tau_i' e_j \tau_j' \cdot \mathrm{e}^{c_{ij}} \quad \text{with} \quad k = m + \dfrac{(i-1)(i-2)}{2} + j, \quad i = 2..m,\ j = 1..(i-1) & \text{for} \quad k > m \end{cases} \tag{45}$$

The extended log-covariance matrix C' has N^2 entries. Its elements c'_{kl} can be expressed as sums of elements of the original matrix $C \in \mathbb{R}^{m\times m}$. They are:

$$c'_{kl} = \begin{cases} c_{kl} & \text{for} \quad \mathrm{k} \leq m \text{ and } l \leq m \\ c_{il} + c_{jl} \quad \text{with} \quad k = m + \dfrac{(i-1)(i-2)}{2} + j, \quad i = 2..m,\ j = 1..(i-1) & \\ & \text{for} \quad k > m \text{ and } l \leq m \\ c_{kp} + c_{kq} \quad \text{with} \quad l = m + \dfrac{(p-1)(p-2)}{2} + q, \quad p = 2..m, q = 1..(p-1) & \\ & \text{for} \quad k \leq m \text{ and } l > m \\ c_{ip} + c_{iq} + c_{jp} + c_{jq} \quad \text{with} \quad \left\{ \begin{array}{c} k = m + \dfrac{(i-1)(i-2)}{2} + j, \\ i = 2..m,\ j = 1..(i-1) \\ \text{and} \\ l = m + \dfrac{(p-1)(p-2)}{2} + q, \\ p = 2..m, q = 1..(p-1) \end{array} \right\} & \\ & \text{for} \quad k > m \text{ and } l > m \end{cases} \tag{46}$$

Finally, we need to know the effective strike that is to be used in the basket formula. It is given by:

$$\lambda := K \cdot \tau + 1 - \delta + \sum_{i=1}^{m} s_i \tau_i' - \sum_{i=1}^{m} \sum_{j=1}^{i-1} s_i \tau_i' s_j \tau_j' \tag{47}$$

Using all of the above definitions, the non-canonical caplet approximation is finally given by the expectation:

$$\mathrm{E}\left[\left(\sum_{k=1}^{N} x_k - \lambda\right)_+\right] \tag{48}$$

for lognormal variates x_k with expectations:

$$\mathrm{E}[x_k] = \xi_k \tag{49}$$

and log-covariances:

$$\mathrm{E}[\ln x_k \cdot \ln x_l] - \mathrm{E}[\ln x_k] \cdot \mathrm{E}[\ln x_l] = c'_{kl} \tag{50}$$

which can be computed with any basket approximation such as the rank reduction method given in the Appendix.

3 Analysis of the skew resulting from the approximation formulæ

There are various effects that contribute to the skew that we can observe in the implied volatilities of caplets as given by the prices we obtain from Monte Carlo simulations with a Libor market model. First of all there is, of course, the skew that was deliberately put into the model by virtue of, for instance, a displaced diffusion evolution of the canonical forward rates. In addition to that, though, non-canonical caplets incur other effects leading to a skew just by themselves, even if the underlying canonical forward rates were designed to be as lognormally distributed as possible (for instance, by setting $q = Q = 1$).

3.1 The basket effect

The first effect, albeit that it is the smaller out of the two addressed in this article, is due to the fact that a non-canonical caplet bears some similarity to an option on a basket. To analyse this feature, I shall assume that a caplet can indeed be priced very accurately using an expansion of the Libor calculation formula (17) in conjunction with the rank reduction method. To simplify matters, I will also assume that an expansion as presented in Sections 2.1 and 2.2 is sufficiently precise not to taint the results significantly. Let the basket pricing formula given by the rank reduction method be denoted by:

$$v(\boldsymbol{x}, K, R, C) \tag{51}$$

where $\mathbf{x}$ stands for a vector of expectations of displaced forward rates (or products thereof), K is the strike, R is a strike displacement, and C is the effective log-covariance matrix of lognormally distributed variates whose sum comprises the basket. All mentioning of the modified accrual

factors τ_i' has been suppressed since they can be absorbed into the entries of the vector $\mathbf{x}$, the strike K, the strike displacement R, and the Libor rate L, respectively. The skew as defined in equation (8) is then implicitly given by the equation:

$$\begin{aligned}&\left.\frac{\partial V_{\text{Black}}(L, K, \hat{\sigma}, T)\cdot\tau}{\partial K}\right|_{K=L} + \left.\frac{\partial V_{\text{Black}}(L, K, \hat{\sigma}, T)\cdot\tau}{\partial \hat{\sigma}}\right|_{K=L} \cdot \left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=L} \\ &\quad = \left.\frac{\partial v(\boldsymbol{x}, K, R, C)}{\partial K}\right|_{K=L}\end{aligned} \tag{52}$$

The rank reduction approximation involves a modification of the covariance matrix such that its rank is reduced to one, and the calculation of the expectation:

$$v(\boldsymbol{x}, K, R, C) = \mathrm{E}\left[\left(\sum_{i=1}^{n} x_i \mathrm{e}^{-\frac{1}{2}\tilde{\sigma}_i^2 T + \tilde{\sigma}_i\sqrt{T}y} - (K+R)\right)_+\right] \tag{53}$$

where the $\tilde{\sigma}_i$ stand for the square root of the diagonal elements of the modified and rank reduced time-to-expiry-averaged covariance matrix, i.e. $\tilde{\sigma}_i = a_i/\sqrt{T}$ with a_i defined in equations (83) and (84) in the Appendix. By virtue of the condition $Q > 0$, all of the elements of the vector $\boldsymbol{x}$ are positive, and since we assume positive correlation between all forward rates, the expectation in equation (53) can be expressed as:

$$v(\boldsymbol{x}, K, R, C) = \sum_{i=1}^{n} x_i \cdot \mathrm{N}(\tilde{\sigma}_i\sqrt{T} - y^*) - (K+R)\cdot\mathrm{N}(-y^*) \tag{54}$$

with $y^* = y^*(K)$ being the solution of:

$$\sum_{i=1}^{n} x_i \mathrm{e}^{-\frac{1}{2}\tilde{\sigma}_i^2 T + \tilde{\sigma}_i\sqrt{T}y^*} = (K+R), \tag{55}$$

as shown in the Appendix. Equations (54) and (55) can be used to compute the unknown quantity on the right hand side of equation (52). This yields:

$$\begin{aligned}\frac{\partial v}{\partial K} = &\left[(K+R)\cdot\varphi(y^*(K)) - \sum_{i=1}^{n} x_i \cdot \varphi(\tilde{\sigma}_i\sqrt{T} - y^*(K))\right]\cdot\frac{\partial y^*(K)}{\partial K} \\ &- \mathrm{N}(-y^*(K))\end{aligned} \tag{56}$$

Thanks to the fact that equation (55) can be rewritten as:

$$(K+R) = \sum_{i=1}^{n} x_i \frac{\varphi(\tilde{\sigma}\sqrt{T} - y^*(K))}{\varphi(y^*(K))} \tag{57}$$

equation (56) can be simplified to:

$$\frac{\partial v}{\partial K} = -\mathrm{N}(-y^*(K)) \tag{58}$$

As a consequence, for $K = L$, the skew is governed by:

$$\left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=L} = \frac{\mathrm{N}(-\frac{1}{2}\hat{\sigma}\sqrt{T}) - \mathrm{N}(-y^*(L))}{\varphi(\frac{1}{2}\hat{\sigma}\sqrt{T})L\sqrt{T}} \tag{59}$$

where $\hat{\sigma}$ stands for the implied Black volatility consistent with the caplet price.

At this point, in order to make some more progress on our understanding of the skew resulting from the basket effect on the skew, I resort to Taylor expansions. First, let us remember that for small ε, we have:

$$\mathrm{N}(\varepsilon) \simeq \frac{1}{2} + \frac{\varepsilon}{\sqrt{2\pi}} - \frac{\varepsilon^3}{\sqrt{2\pi}} + \mathcal{O}(\varepsilon^5) \tag{60}$$

Also, let us recall that *at-the-money* means that the unconditional expectation of the basket is equal to the displaced Libor rate:

$$(L + R) = \sum_{i=1}^{n} x_i \tag{61}$$

Combining equations (61) and (55), and expanding the exponentials in equation (55) to first order, we can approximate y^* as:

$$y^* \simeq \frac{1}{2} \cdot \frac{\sum_i x_i \tilde{\sigma}_i^2 T}{\sum_i x_i \tilde{\sigma}_i \sqrt{T}} \tag{62}$$

Equally, expanding the at-the-money Black formula:

$$V_{\mathrm{Black}}(L, L, \hat{\sigma}, T) = L \cdot [\mathrm{N}(\tfrac{1}{2}\hat{\sigma}\sqrt{T}) - \mathrm{N}(-\tfrac{1}{2}\hat{\sigma}\sqrt{T})] \tag{63}$$

and the rank reduction basket pricing formula (54) at the money for small T using (60), we arrive at:

$$\frac{L\hat{\sigma}\sqrt{T}}{\sqrt{2\pi}} \simeq \frac{\sum_i x_i \tilde{\sigma}_i \sqrt{T}}{\sqrt{2\pi}} \qquad \text{i.e.} \qquad L\hat{\sigma} \simeq \sum_i x_i \tilde{\sigma}_i \tag{64}$$

Now, substituting (64) and (62) into (59), expanding according to (60), and using (61), we obtain:

$$\begin{aligned} \left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=L} &= \frac{1}{2}\mathrm{e}^{\frac{1}{2}\hat{\sigma}^2 T}\frac{1}{L}\left[\frac{\sum_i x_i \tilde{\sigma}_i^2}{\sum_i x_i \tilde{\sigma}_i} - \hat{\sigma}\right] \\ &= \frac{1}{2}\mathrm{e}^{\frac{1}{2}\hat{\sigma}^2 T}\frac{1}{L\hat{\sigma}}\left[\sum_i \nu_i \tilde{\sigma}_i^2 - \left(\sum_i \nu_i \tilde{\sigma}_i\right)^2\right] \end{aligned} \tag{65}$$

where I have used the definition:

$$\nu_i := \frac{x_i}{\sum_i x_i - R} \simeq \frac{x_i}{L} \tag{66}$$

A closer look at the terms in the square brackets on the right hand side of equation (65) reveals that, within the scope of the used approximations, the equation can be rewritten as:

$$\left.\frac{d\hat{\sigma}(K)}{dK}\right|_{K=L} = \frac{1}{e}\mathrm{e}^{\frac{1}{2}\hat{\sigma}^2 T}\frac{1}{L\hat{\sigma}}\left[\sum_i \nu_i(\tilde{\sigma}_i - \hat{\sigma})^2 - \frac{R}{L}\hat{\sigma}^2\right] \tag{67}$$

Obviously, equation (67) implies that the skew is positive for $R \leq 0$, i.e. for $Q \geq 1$. What is interesting about this equation is that it predicts that even for $R = 0$, i.e. for instance for $Q = 1$ (which means that all the canonical forward rates are lognormally distributed in their own natural measure) and in the absence of any spread differential, a non-canonical caplet would display a very small, but *positive* skew, unless all the involved forward rates have identical modified average volatility $\tilde{\sigma}_i$. This means, even when we keep the effective at-the-money volatility of a non-canonical caplet fixed, and even when we keep the effective implied volatility of all canonical caplets fixed or virtually unchanged, it is possible to increase the skew of the given non-canonical caplet ever so slightly by a simple change to the term structure of the instantaneous volatility of the canonical forward rates. This is because, out of all the discrete forward rates that contribute to the value of the non-canonical caplet, at most one of them expires naturally on the same date as the caplet. The values of all the remaining canonical discrete forward rates that eventually contribute to the fixing value of the non-canonical rate that determines the payoff of the caplet are taken as a snapshot *before* their natural expiry. This means that the root-mean-square volatility they realise until the fixing date of the non-canonical caplet is not given by their canonical implied volatility, but misses out on the last part of instantaneous volatility between expiry of the non-canonical caplet and the natural fixing date of the individual contributing discrete forward rates. Since we are free to tailor term structures of instantaneous volatilities of canonical forward rates at will in the Libor market model framework, we can change the shape of the volatility curve, and thus the value of the partially averaged root-mean-square volatility to expiry of the non-canonical caplet, whilst keeping the implied volatility of each canonical caplet unchanged.

Fortunately, the basket effect for non-canonical caplets is very small as long as the non-canonical accrual period doesn't span too many canonical periods and thus proves to be of no practical importance. It is, however, from a theoretical point of view astounding to observe a noticeable effect of the shape of term structure of the canonical forward rates on the skew of non-canonical caplets. It remains to be seen if this kind of effect is also observable in other financial modelling environments, and to what extent it can be detected in the skew of the implied volatilities associated with European swaptions.

As a side note, it may be worth mentioning that the positive sign of the skew effect resulting from the summation of lognormally distributed assets is fairly well known for Asian and basket options when they are approximated by a Johnson distribution. The Johnson distribution is identical to a displaced lognormal distribution. For Asian and basket options, it is fairly straightforward to write down the equations for the matching of the first three moments, and to show that the displacement is negative, thus giving rise to a positive skew when all the underlying constituents are strongly positively correlated.

3.2 The spread differential induced skew

The spread induced displacement is negative if the spread incurred by any one Libor rate is larger than the spread of the Libor that determines the dynamics of the model. For example, if we build the model from a 3 m Libor rate with a spread of 10 bp (i.e. our funding is 10 bp cheaper than 3 m Libor), and we have a 20 bp Libor spread, then we will end up with a negative spread induced displacement resulting in a positive skew for options on the 6 m-Libor rate.

Since I am at this point at serious risk of stretching the readers' patience beyond redemption, I shall only outline the analysis of the spread differential induced skew. As a starting point, we can approximate the non-canonical Libor rate as a single lognormal variate with relative volatility σ subject to a spread differential induced skew as given in equation (25). We assume $\delta \gtrsim 1$ since we place ourselves in the position of a financial institution that funds approximately at the 3 m Libor cost but writes a caplet on a longer accrual period for which the equivalent Libor rates are higher than the simple compounding effect for the longer period could justify. The equivalent Black volatility at the money is implicitly (approximately) specified by the leading terms in equation (25), i.e.:

$$V_{\text{Black}}(L, K, \hat{\sigma}, T)|_{K=L} = V_{\text{Black}}((1 + h\tau) \cdot L^*, K - h, \sigma, T)|_{K=L} \tag{68}$$

where I have used the abbreviation $h := (\delta - 1)/\tau$ and assumed that we can model the basket of canonical forward rates as a single lognormally distributed L^* with volatility $\hat{\sigma}^*$. Straightforward expansions of equation (68) lead to:

$$\hat{\sigma} = \hat{\sigma}^* \cdot \left(1 - \frac{h}{L} + \frac{1}{2} h\tau\right) + \mathcal{O}(h^2) \tag{69}$$

The next step is then to differentiate (68) with respect to K, and carry out some further Taylor expansions and simplifications. We finally arrive at a dependence of the spread differential induced skew as defined in equation (8) on the at-the-money implied volatility of the non-canonical Libor rate, the spread differential h, and the non-canonical forward rate L itself given by:

$$\chi \approx \frac{\hat{\sigma}^*}{20} \cdot \frac{h}{L} + \mathcal{O}(h^2) \tag{70}$$

The interesting fact here is that the spread differential induced skew diverges as Libor rates approach zero, and that it grows linearly with h (as long as implied volatilities or times to maturity are small since I used first order expansions in $\hat{\sigma}\sqrt{T}$ and $\hat{\sigma}^*\sqrt{T}$). For small values of the actual spread and the assumption that spread discount factors are given by:

$$\zeta_\tau(t_{\text{start}}, t_{\text{start}} + \tau) = \mathrm{e}^{-\varepsilon_\tau \cdot \tau} \tag{71}$$

with ε_τ representing the cumulatively compounded spread rate for τ-period Libor rates, we obtain:

$$h \cdot \tau = (\varepsilon_\tau - \varepsilon_{\tau^*}) \cdot \tau + \mathcal{O}(((\varepsilon_\tau - \varepsilon_{\tau^*}) \cdot \tau)^2) \tag{72}$$

which means that the skew is approximately linear in the spread differential $h = (\varepsilon_\tau - \varepsilon_{\tau^*})$. If we recall that spread differentials are currently noticeably pronounced in Japan, where rates are low and volatilities high, we may expect the spread differential induced skew to be of non-negligible size in that market.

4 Numerical examples

The first example I give to demonstrate the accuracy of the presented approximations is an option on a 12 m Libor rate, expiring in 12 months from inception. All the discrete forward rates that contribute to this caplet are initially set to values near 4%, and are assumed to be perfectly lognormal in their natural measure, i.e. $Q = 1$. I used the same instantaneous term structure of volatility for all of the canonical forward rates given by:

$$\sigma_i(t) = [a + b(T_i - t)] \cdot \mathrm{e}^{-c\cdot(T_i - t)} + d \tag{73}$$

with $a = 0.1, b = 1, c = 2, d = 0.1$, and a time-constant correlation structure given by:

$$\rho_{ij} = \mathrm{e}^{-\beta\cdot(T_i - T_j)} \tag{74}$$

with $\beta = 0.1$. In Figure 3, I show the results from numerical simulations using 2^{20} Sobol' vector draws and analytical expansions for the given term structure (labelled as "peaked volatility") in comparison with the numbers we would obtain if we had set the volatility of all canonical forward rates to 26.05% (denoted as "flat volatility"). The description "first order" refers to the expansion outlined in Section 2.1, whereas "second order" is the implied volatility curve resulting from the method explained in Section 2.2. The skew as defined in equation (8) associated with the curves is given in Table 1. As we can see, the agreement of the first order expansion with the numerical results is for practical purposes just about at the edge of being useful, whereas the agreement of the second order expansion with the numerical data is rather excellent indeed.

TABLE 1: THE SKEW NUMBERS ASSOCIATED WITH THE CURVES IN FIGURE 3

volatility type	first order expansion χ	second order expansion χ	numerically χ
peaked	0.017%	0.035%	0.037%
flat	0	0.018%	0.020%

The figure and table highlight several features that we had already identified in the analytical discussion in Section 2. First, there is clear evidence of the small but positive skew as a consequence of the basket effect explained in Section 3.1. In accordance with the analysis given in that section, the skew increases as we switch from equal and flat volatility of the forward rates to a peaked term structure of volatility. The term structure of instantaneous volatility gives rise to the effective variances of all the contributing discrete forward rates to differ, and as we can tell from equation (67), this in turn causes the skew to increase.

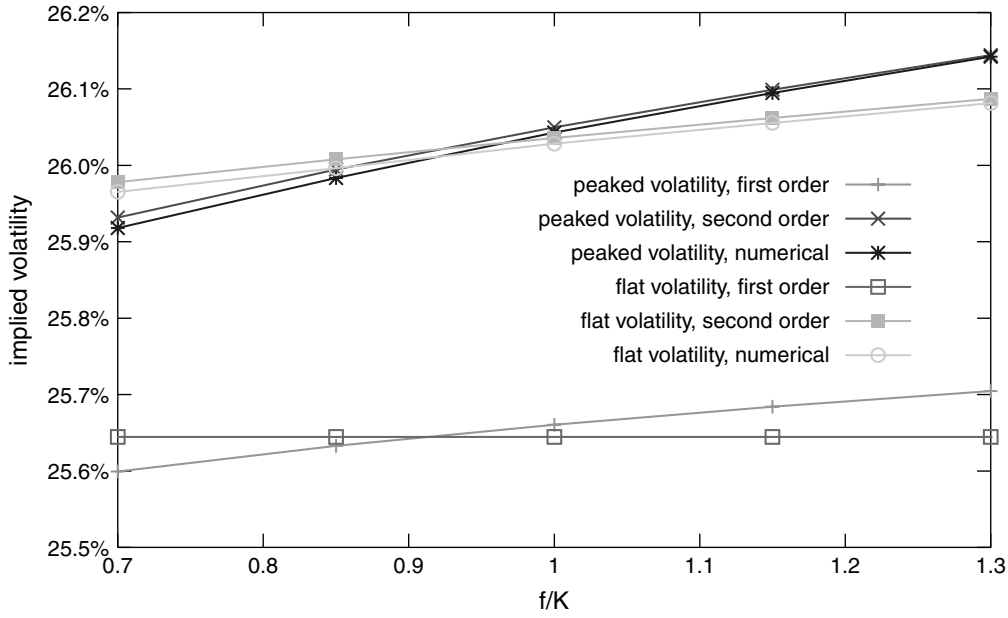


Figure 3: The implied volatilities of a 12 m caplet on a 12 m Libor rate

The fact that there is still some residual skew even for flat volatilities can also be explained if we compare the first and second order expansion results. Since the second order expansion takes into account the effect of the (nearly) lognormally distributed products of forward rates which have a larger variance than the first order terms, it effectively values a basket of differing constituents, and that in turn causes a slight skew, as discussed at great length by now.

The next example I give is to show the effect of the spread differential for a caplet on a 6 m Libor rate with 12 months to expiry, as analysed in Section 3.2, using lower levels of interest rates and somewhat higher volatilities than before, albeit not quite as extreme as those prevailing in the Japanese market. Again, I set $Q = 1$ for all the canonical forward rates, and I choose the parametrisation $a = 0.2$, $b = 1$, $c = 1$, $d = 0.2$, and $\beta = 0.1$ for the correlation coefficient. A total of 12 curves are displayed in Figure 4, representing the implied volatilities computed numerically and analytically (using the second order expansion) from different rate and spread differential settings. In the legend of the figure, the level of the canonical forward rates is indicated by either L = 60 bp or L = 30 bp, which is to mean that the Libor rates are just slightly lower than the given numbers. The spread differential between the 3 m canonical rates and the 6 m Libor rate is given by h = 0 bp, h = 10 bp or h = 20 bp. In all four cases, the analytical approximation matches the numerically computed results extremely well. This good agreement between numerical and analytical figures for a 12 m × 6 m caplet is not that surprising if we consider that the 6 m rate in question is composed of two canonical 3 m Libor rates which in turn means that there are no third order terms in equation (3.2) that would be neglected by the approximation given in Section 2.2. What's more, just as one would expect from the relationship (70), the implied volatilities for $L \approx 60$ bp and the spread differential at

20 bp coincide with the values for $L \approx 30$ bp and around 10 bp spread differential. I should also explain why the point at $f/K = 0.7$ is missing for $L \approx 30$ bp and around 20 bp spread differential. The reason is that this is where the effective negative displacement of the Libor rate results in a floorlet struck at $0.7 \cdot L$ being perfectly worthless, which is why no equivalent Black volatility can be implied.

To summarise the results on the skew, I give in Table 2 the skew figures that were computed from the results shown in Figure 4. Clearly, the significant magnitude of the skew that is induced by spread differentials emphasises how important it is that the forward rates that are evolved in a Libor market model are directly linked to interbank offered rates, and not immediately to funding rates, since this would cause an unintended skew to be built into the model. This is to say that even when we correct the volatility levels such that the effective implied volatilities at the money are calibrated to the market, we still have to bear in mind that there may be a significant skew for non-canonical caplets when spread differentials are present.

TABLE 2: THE SKEW NUMBERS ASSOCIATED WITH THE CURVES IN FIGURE 3

L	h	χ (numerically)	χ (from analytical prices)	χ from approximation (70)
60 bp	0 bp	0.0044%	0.0047%	0
30 bp	0 bp	0.0029%	0.0034%	0
60 bp	10 bp	0.5357%	0.5361%	0.51%
60 bp	20 bp	1.0709%	1.0714%	1.01%
30 bp	10 bp	1.0690%	1.0693%	1.01%
30 bp	20 bp	2.2158%	2.2160%	2.03%

Finally, I present an example of the accuracy of the approximations for a user-controlled skew. In order to show how strong the given higher order approximations are, I have chosen the scenario of a non-canonical 3 m caplet with 49 months and 2 weeks to expiry in a 3 m Libor market model. This means that the non-canonical rate is almost exactly split between two canonical discrete forward rates which makes it a particularly hard test. The volatility parameters are $a = 0.1$, $b = 1$, $c = 2$, $d = 0.1$, and this time I use a term structure of instantaneous volatility given by:

$$\rho_{ij}(t) == \mathrm{e}^{-\beta\left||T_i - t|^\kappa - |T_j - t|^\kappa\right|} \tag{75}$$

with $\beta = 0.8$ and $\kappa = 0.2$. This term structure of instantaneous volatility and correlation allows for quite a considerable decorrelation of the forward rates. In addition to that, I used forward rates near 9%. As you can see in Figure 5, the numerical and analytical results agree very well for different levels of the skew, even for options on the Libor rate that are considerably far away from the money.

In summary, I would like to say that I was surprised how complicated it turned out to find a sufficiently accurate caplet approximation in the framework of a Libor market model

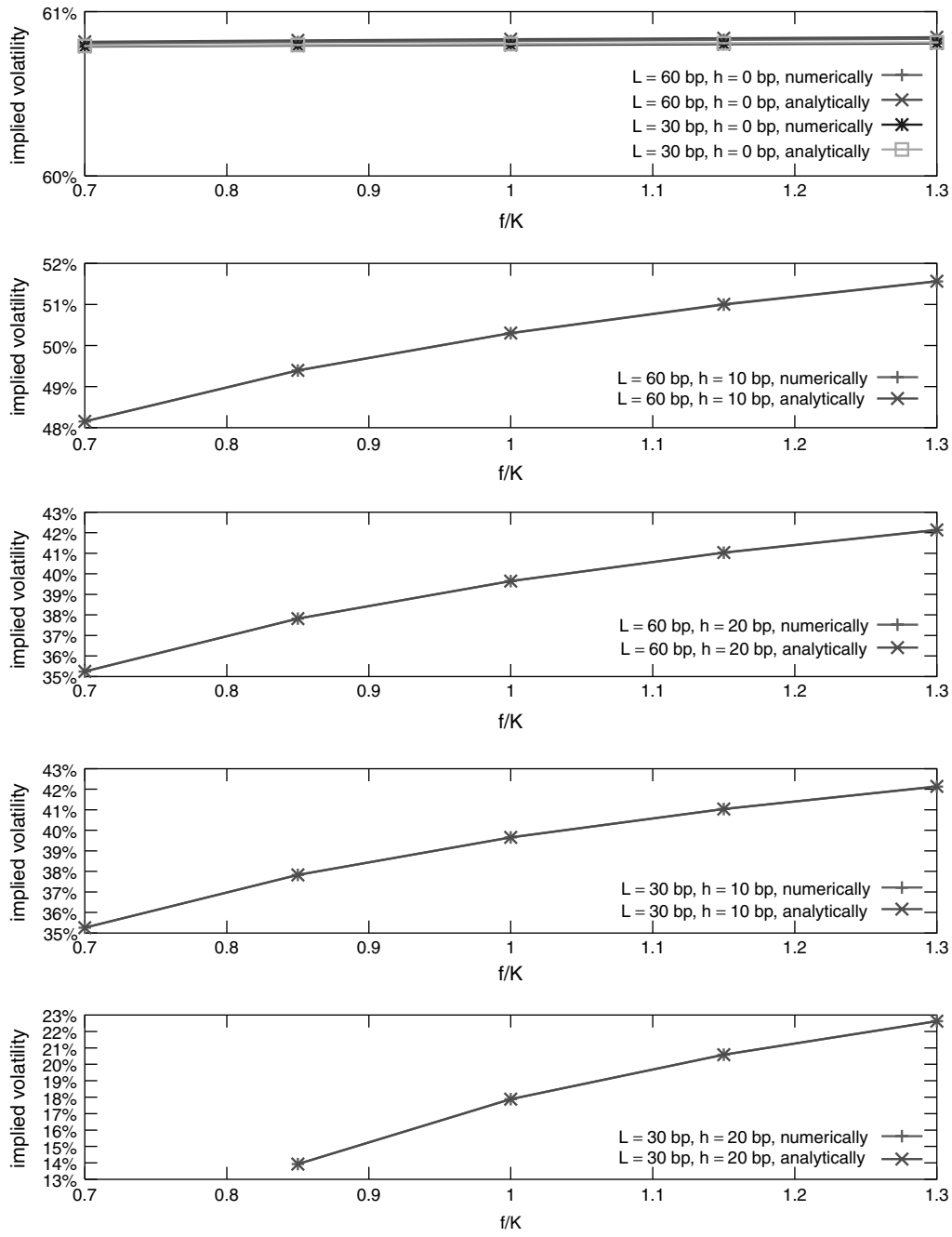


Figure 4: The implied volatility ($\hat{\sigma}$) of a caplet on a 6 m Libor rate expiring in 12 months for different levels of the non-canonical forward rate L and different spread differentials h

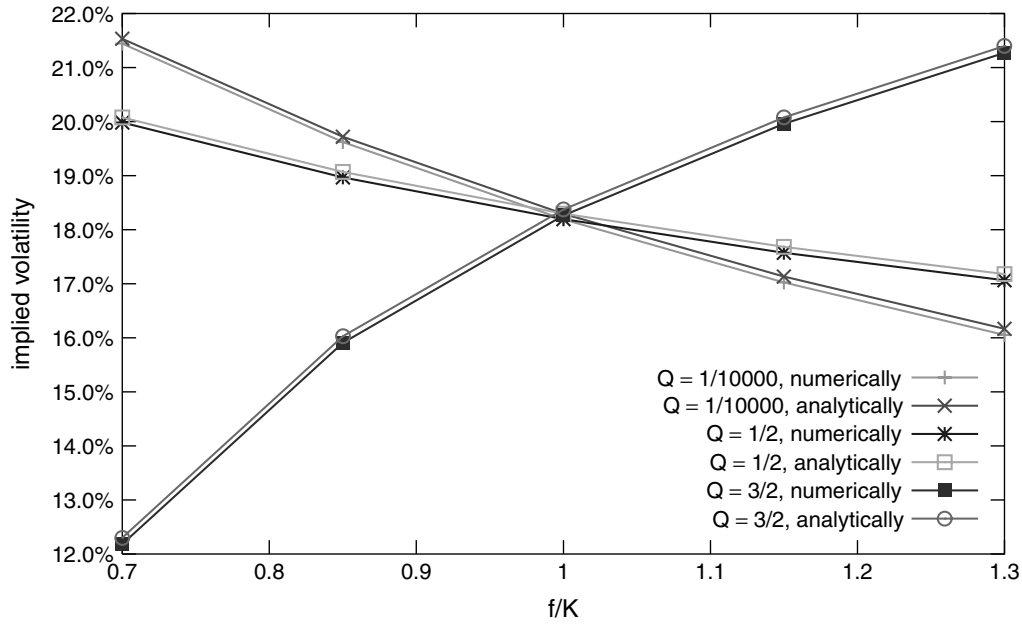


Figure 5: The implied volatility ($\hat{\sigma}$) of a non-canonical caplet on a 3 m Libor rate expiring in 49 months and two weeks for different values of the skew parameter Q

with a simple user-controlled skew such as given by the stochastic differential equation (1). After all, we are talking here about an interest rate model *that is designed to meet the market features of options on Libor rates by design*, and the pricing of caplets is rarely what the model is originally implemented for. However, since the trading of exotic derivatives valued with a Libor market model requires the model to be reasonably calibrated to market instruments (which sometimes includes options on 6 m Libor rates where they are sufficiently liquid, and always includes many non-canonical caplets), and since the handling of many different instruments in a consistent framework requires not only the ability to value all exotics using Monte Carlo simulations, but also the much larger numbers of simpler derivatives such as caps and floors (that are typically in any interest rate option book) in a timely fashion, analytical approximations for caplets and floorlets for the given model may be a very desirable thing to have.

Appendix. The rank reduction method for options on baskets of positively correlated lognormals

The problem at hand is the pricing of a call or put option on a weighted average of correlated lognormal variates with expectation f_i. In general, there is no requirement for the fixing of the associated correlated assets to occur simultaneously, which means we could also allow for the pricing of Asian and Asian basket options. All we need for the pricing of the basket option is the covariance matrix C of the logarithmic returns and the weights w_i. When the fixing of all

of the involved assets is to be simultaneous at time T, we would have:

$$c_{ii} = \sigma_i^2 T \tag{76}$$

and:

$$c_{ij} = \sigma_i \sigma_j \rho_{ij} T \quad \text{for} \quad i \neq j \tag{77}$$

using the usual notation for implied volatility and correlation. The basket, or weighted average, of the involved n lognormal variates is given by:

$$B = \sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}c_{ii}+z_i} \tag{78}$$

with the modified weights:

$$\omega_i = w_i f_i \tag{79}$$

and the normal variates z_i satisfying the covariance conditions:

$$\mathrm{E}[z_i] = 0 \quad \text{and} \quad \mathrm{E}[z_i z_j] = c_{ij} \tag{80}$$

The pricing of an option on the geometric average of lognormal variates can be done without any difficulty since the geometric average is itself lognormally distributed. However, for an arithmetic average, this can only be done if the covariance matrix is of rank 1, subject to an additional criterion that is elaborated in the following.

The key idea of the rank reduction method is to substitute the original covariance matrix C with a matrix C' of rank one such that the log-variance of a geometric basket with the same modified weighting coefficients as B is preserved. In other words, we need to find a covariance matrix C' such that:

$$\sum_{i,j=1}^{n} \omega_i \omega_j c_{ij} = \sum_{i,j=1}^{n} \omega_i \omega_j c'_{ij} \tag{81}$$

Any symmetric positive semi-definite matrix C' of rank one can be written as the dyadic product of a vector $\mathbf{a}$ with itself:

$$C' = \mathbf{a} \cdot \mathbf{a}^{\mathrm{T}} \tag{82}$$

In order to retain the ratios of the standard deviations of all of the constituents, we set:

$$a_i := s \cdot \sqrt{c_{ii}} \tag{83}$$

with some common scaling factor s. This factor can be determined from the geometric basket log-variance preserving condition (81):

$$s := \sqrt{\frac{\sum_{i,j=1}^{n} \omega_i \omega_j c_{ij}}{\sum_{i,j=1}^{n} \omega_i \omega_j \sqrt{c_{ii} c_{jj}}}} \tag{84}$$

Once we have computed the coefficients a_i, the approximate (undiscounted) price of a call option on the arithmetically weighted basket struck at K is given by:

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}a_i^2 + a_i y} - K\right)_+\right] \tag{85}$$

where y is a standard normal variate. As long as the function:

$$g(y) = \sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}a_i^2 + a_i y} \tag{86}$$

is monotonic in y, we can compute expectation (85) comparatively easily. A sufficient condition for the monotonicity of the function $g(y)$ is given if all of the weighting coefficients ω_i are positive. For general basket options such as the option on a bond that not only pays coupons but also demands repayments (which would involve negative weights), this requirement may be too strict. Even when there are some slightly negative weighting coefficients, the function $g(y)$ may still remain monotonic in y. However, for simplicity, we demand at this point that:

$$\omega_i \cdot a_i \geq 0. \tag{87}$$

In practice, this restriction rarely poses a problem. Given (87), we can price the call option on the basket by first identifying the critical value y^* where:

$$g(y^*) - K = 0 \tag{88}$$

The value y^* can be found by the use of the standard Newton method, and converges very rapidly due to the smoothness of the function g. A good initial guess is usually given by the second order expansion of $g(y)$ in y around zero. Given the definitions:

$$b := \sum_{i=1}^{n} \frac{1}{2} a_i^2 \omega_i \mathrm{e}^{-\frac{1}{2}a_i^2}$$

$$c := \sum_{i=1}^{n} a_i \omega_i \mathrm{e}^{-\frac{1}{2}a_i^2}$$

$$d := \sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}a_i^2} - K$$

calculate the discriminant $\delta := c^2 - 4bd$. Then, if the discriminant δ is positive, use:

$$y_{\text{initial guess from second order expansion}} := \frac{\sqrt{\delta} - c}{2b} \tag{89}$$

as your initial guess; else use:

$$y_{\text{initial guess from first order expansion}} := -\frac{d}{c} \tag{90}$$

The second order expansion is usually already within a relative accuracy of 10^{-5} and may thus be a sufficiently precise approximation for y^* for certain applications. Nonetheless, due to the availability of an extremely good initial guess, any subsequent Newton iterations typically converge to sufficient precision within a single step. Having established the critical value y^*, the approximate value of the call option is given by:

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}c_{ii}+z_i} - K\right)_+\right] \simeq \left(\sum_{i=1}^{n} \omega_i \mathrm{N}(-y^* + a_i) - K\mathrm{N}(-y^*)\right) \tag{91}$$

wherein $\mathrm{N}(\cdot)$ is the cumulative normal distribution function. Equally, the approximation for the value of a put option can be computed as:

$$\mathrm{E}\left[\left(K - \sum_{i=1}^{n} \omega_i \mathrm{e}^{-\frac{1}{2}c_{ii}+z_i}\right)_+\right] \simeq \left(K\mathrm{N}(y^*) - \sum_{i=1}^{n} \omega_i \mathrm{N}(y^* - a_i)\right) \tag{92}$$

REFERENCES

- Andersen, L. (1999) A simple approach to the pricing of Bermudan swaptions in the multifactor LIBOR market model. *The Journal of Computational Finance*, 3(2): 5–32.
- Andersen, L. and Andreasen, J. (2000) Volatility skews and extensions of the libor market model. *Applied Mathematical Finance*, 7(1).
- Bachelier (1900) Théorie de la Spéculation. PhD thesis, Université de Paris.
- Breeden, D. T. and Litzenberger, R. H. (1978) Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4): 621–651.
- Borodin, A. N. and Salminen, P. (1996) *Handbook of Brownian Motion – Facts and Formulae*. Birkhäuser. ISBN 3-7643-5463-1.
- Curran, M. (1994) Valuing Asian and portfolio options by conditioning on the geometric mean price. *Management Science*, 40: 1705–1711.
- Glassermann, P. and Merener, N. (2003) Numerical solution of jump-diffusion libor market models. *Finance and Stochastics,* 7(1): 1–27.
- Hagan, P., Kumar, D., and Lesniewski, A. S. (2002) Managing smile risk. *Wilmott*, 2(1): 84–108.

■ Hull, J. and White, A. (1990) Pricing interest rate derivative securities. *Review of Financial Studies*, 3(4): 573–592.
■ Hull, J. and White, A. (2000) Forward rate volatilities, swap rate volatilities, and the implementation of the LIBOR market model. *Journal of Fixed Income*, 10(2): 46–62, September 2000.
■ Hunter, C., Jäckel, P., and Joshi, M. (2001) Getting the drift. *RISK Magazine*, July 2001. http://www.rebonato.com/MarketModelPredictorCorrector.pdf.
■ Jäckel, P. (2002) *Monte Carlo Methods in Finance*. John Wiley and Sons, Chichester, February 2002.
■ Jäckel, P. and Rebonato, R. (2003) The link between caplet and swaption volatilities in a Brace-Gatarek-Musiela/Jamshidian framework: approximate solutions and empirical evidence. *The Journal of Computational Finance*, 6(4): 41–59 (submitted in 2000). www.rebonato.com/capletswaption.pdf.
■ Jamshidian, F. (1989) An exact bond option pricing formula. *Journal of Finance*, 44(9): 205–209.
■ Ju, N. (2001) Pricing Asian and basket options via taylor expansion of the underlying volatility. Working paper, The Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, +1 301 405 2934, nju@rhsmith.umd.edu, March 2 2001.
■ Karasinski, P. (2002) Do we need a new market volatility standard? Exploring rate-level dependence of volatility presented at the Quantitative Finance London 2002 conference. Fixed Income Research Schroder Salomon Smith Barney, November 25 2002.
■ Levy, E. (1992) Pricing European average rate currency options. *Journal of International Money and Finance*, 11: 474–491.
■ Longstaff, F. A. and Schwartz, E. S. (1998) Valuing American options by simulation: a simple least squares approach. Working paper, The Anderson school, UCLA.
■ Platen, E. (2002) Consistent pricing and hedging for a modified constant elasticity of variance model presented at the Quantitative Finance London 2002 conference, University of Technology, Sydney, November 25 2002.
■ Pietersz, R., Pelsser, A., and van Regenmortel, M. (2002) Fast drift approximated pricing in the BGM model Working paper, Erasmus University Rotterdam. http://www.few.eur.nl/few/people/pelsser/papers/Pietersz_Pelsser_vanRegenmortel-FastDriftApproxBGM-02.pdf.
■ Rebonato, R. (1999) Calibration of the BGM model. *RISK Magazine*, March: 74–79.
■ Reiner, E., Davidov, D., and Kumanduri, R. (2001) A rapidly convergent expansion method for Asian and basket options. In *Risk Europe*, Paris, April 11th 2001.
■ Rubinstein, M. (1983) Displaced diffusion option pricing. *Journal of Finance*, 38: 213–217, March 1983.
■ Schoenmakers, J. G. M. (2000) Calibration of LIBOR models to caps and swaptions: a way around intrinsic instabilities via parsimonious structures and a collateral market criterion. Working paper, Weierstrass Institute, Berlin. http://www.wias-berlin.de/publications/preprints/740.
■ Turnbull, S. and Wakeman, L. (1991) A quick algorithm for pricing european average options. *Journal of Financial and Quantitative Analysis*, 26(3): 377–390.
■ Vasicek, O. A. (1977) An equilibrium characterisation of the term structure. *Journal of Financial Economics*, 5: 177–188.

■ Zühlsdorff, C. (2002) Extended libor market models with affine and quadratic volatility. Working paper, University of Bonn. ftp://ftp.wiwi.uni-bonn.de/papers/bgse/2002/bgse6_2002.pdf.
■ Zühlsdorff, C. (2002) The pricing of derivatives on assets with quadratic volatility. Working paper, University of Bonn. ftp://ftp.wiwi.uni-bonn.de/papers/bgse/2002/bgse5_2002.pdf.

22
The Art and Science of Curve Building

Owen Walsh

Wilmott magazine, November 2003

At first glance, the mechanics of building a discount factor curve would appear to be a fairly mundane subject. At the very least, one would expect that it must be trivial, completely understood and there must be a market standard way of doing it. After all, a good interest rate curve is the most basic requirement for pricing and hedging interest rate derivatives. If the input curve is bad, no matter the model, it is guaranteed that the prices and hedging parameters it returns will be bad. As the saying goes: garbage in, garbage out. In fact, it turns out, that there is quite a bit of subtlety and flexibility in the curve building process. There is room for some (well thought-out) artistic license. Moreover, the subject is certainly not as well understood as it should be.

To illustrate, we will focus on the swap market. In this market, generally, the discount factor curve is built based on a combination of quoted money market and par swap rates. For the purposes of our discussion, the money market rates are not important. Our example is based on the following swap rate data:

Term	Par swap rate	Term	Par swap rate
1Y	2.335%	10Y	4.600%
3Y	3.120%	15Y	5.000%
5Y	3.600%	25Y	5.400%

The swaps pay annually and accrue 30/360. This means that a 4.5% swap pays a fixed coupon of 4.5 each year. For the sake of simplicity, we have further assumed that the swaps are spot settled.

Contact address: VP Analytics, FinancialCAD Corp., 7455 132nd Street, Suite 100, Survey, BC, Canada V3W 1J8. E-mail: walsh@fincad.com Telephone: 604 507 2763 www.financialCAD.com

Consistency is not enough

The process of building a discount factor curve from market quotes is known as *bootstrapping*. The basic requirement of the discount factor curve is that it be *consistent* with all input market rates. Consistent means that after we have built the curve, if we perform a *round-trip* calculation of, for example, the 10Y par swap rate, we indeed obtain 4.6% and similarly for all other rates.

It turns out that the consistency requirement is not enough. The bootstrapping process is (usually) an under-determined problem and we have a fair amount of flexibility in determining how the bootstrapping proceeds. We can take advantage of this flexibility to build better interest rate curves, and we will use the structure of implied forward rates to guide our intuition on how to do it. Before we get ahead of ourselves, let's first understand some basic concepts.

Back to basics

For a given date, a discount factor, also known as a zero coupon price, is the present value of one unit paid on that date. In our notation, Df_{1Y} is the present value of one unit paid one year from today. A discount factor curve is a set of dates and discount factors. Given a discount factor curve, we can present value any future cash flow. A discount factor curve also contains other implied information, like the structure of forward rates. Given the one and two year discount factors, the one year implied forward rate, F, effective one year from today can be calculated from the formula, $F = Df_{1Y}/Df_{2Y} - 1$.

Let's now look at the bootstrapping process and, to illustrate, we will focus on the 10Y par swap rate in our example. As we add (or bootstrap) this rate to the curve, it follows, by definition, that:

$$4.6 \sum_{i=1}^{i=10} Df_{iY} + 100 Df_{10Y} = 100$$

Now the first five of these discount factors are known because we have already built a curve that includes the 5Y rate. Hence:

$$4.6 \sum_{i=6}^{i=10} Df_{iY} + 100 Df_{10Y} = 100 - 4.6 \sum_{i=1}^{i=5} Df_{iY}$$

The right-hand side is known and, in this case, we have an under-determined system with one equation and five unknowns. At this point, in order to obtain a unique solution, constraints are added to the structure of the Df_{iY}. Different constraints correspond to different *bootstrapping methods.*

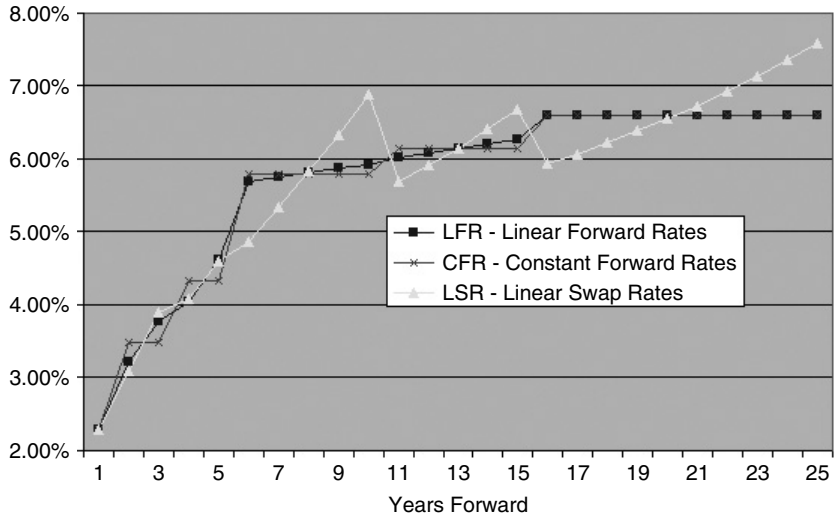
Bootstrapping method 1: linear swap rates (LSR)

The first bootstrapping method we consider, *Linear Swap Rates (LSR)*, assumes that the par swap rate at each intermediate coupon date lies on a straight line between the 5 and 10Y rates. In

our example 6Y = 3.8%, 7Y = 4.0%, 8Y = 4.2% and the 9Y = 4.4%. With these constraints, we now solve for the discount factors. First, we solve for Df_{6Y}:

$$103.8Df_{6Y} = 100 - 3.8\sum_{i=1}^{i=5} Df_{iY}$$

Continue in the obvious way to calculate the 7, 8, 9 and 10-years discount factors. The results are shown in Graph 1 and labeled LSR.



Graph 1: 1-year implied forward rates

Bootstrapping method 2: constant forward rates (CFR)

The second method, *Constant Forward Rates (CFR)*, constrains the problem by enforcing that all one year forward rates, effective at 5, 6, 7, 8 and 9 years, be equal. Let F be this rate. This implies $Df_{6Y} = Df_{5Y}/(1+F)$, $Df_{7Y} = Df_{5Y}/(1+F)^2$ and so on. In our example:

$$4.6\sum_{i=1}^{i=4}(1+F)^{-i} + 104.6(1+F)^{-5} = Df_{5Y}^{-1}\left(100 - 4.6\sum_{i=1}^{i=5} Df_{iY}\right)$$

and it is straightforward to solve for F. The results are labeled CFR in Graph 1.

These two bootstrapping methods are fairly standard and we will go out on a limb and say they are "*market standard*" methodologies (whether they both should be is another question). For a reference on the LSR method, see, for example, Miron and Swannell (1991).

Let's take a step back and analyze the two methods. In Graph 1, we have plotted the one year implied forward rates. Let's focus on the LSR curve. At the short-end of the curve, the

forward rate profile looks fairly good. After 10 years we see a disturbing pattern of "overshoots" where high rates are followed by much lower rates. It is a fact that the LSR methodology can often lead to this type of nasty behavior. On the other hand, the CFR curve does not have any of these overshoots. We see a pleasant looking step function, which in our cleverly chosen example, happens to be increasing. If we only had the choice between these two methods, we would certainly choose to use the CFR. On the other hand, the resulting CFR curve is not completely satisfying. In this case, it seems reasonable to expect that the actual forward rates would be a little *smoother* (at the very least our artistic side says so).

Bootstrapping method 3: linear forward rates (LFR)

We now consider another bootstrapping methodology that we have developed. We call it *Linear Forward Rates (LFR).* The first step in the LFR method is to calculate the CFR curve. The next step involves another bootstrapping pass to "*smooth-out*" the forward rates. During this step, as we splice each point to the curve, rather than looking for a constant forward rate F, we look for linear forward rates that lie on a line of the form $F_0 + KT$, where K is the slope and F_0 is chosen so as to best "fit" the current, previous and next forward rates. It turns out that the discount factors are best described in the following recursive way where $Df_{i+1} = Df_i/(1 + F_0 + KT_i)$ where T_i is the time in years of the i^{th} coupon period. In our example, $T_i = 1$ and if we again consider the 10Y point, it follows that:

$$4.6\sum_{i=6}^{i=9} Df_{iY} + 104.6Df_{10Y} = Df_{5Y}^{-1}\left(100 - 4.6\sum_{i=1}^{i=5} Df_{iY}\right),$$

and we can solve for the slope K. Though this example does not demonstrate it, when we are in a situation where the curve increases and then decreases (or vice versa), we do not modify the forward rates at these points. It turns out, that in this case, the best solution is to leave the forward rates in this region constant. The results for our example are labeled LFR in Graph 1.

Is interpolation good enough?

Combining this LFR discount factor with some sort of interpolation method is probably good enough for most applications. But, is it good enough for all applications? And, for that matter, what interpolation should we use?

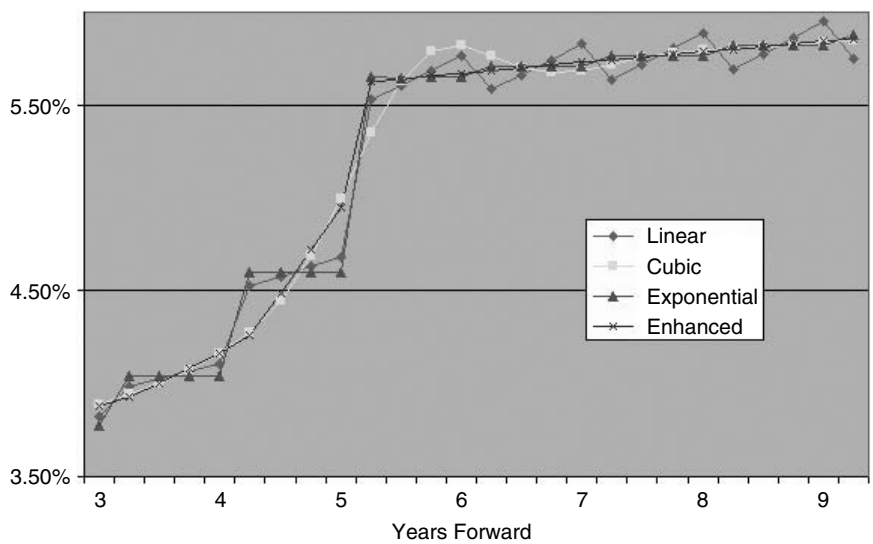
Suppose, for example, we wanted to value a strip of three-month caps. Clearly, this derivative is extremely sensitive to the value of the underlying forward rates. We choose to use the LFR curve in combination with one of the following three interpolation methods; for more details see, for example, Mathews (1987) and Press *et al* (1996).

1. Linear-discount factors between points are assumed to lie on a straight line.
2. Cubic Spline – The whole curve of discount factors lie on a "smooth" cubic spline (piecewise cubic, continuous and differentiable).

3. Exponential – discount factors between two points lie on an exponential curve of the form Ce^{RT}, where C and R are constants.

Looking at 3-month forward rates we see that linear interpolation has regions where a disturbing "saw-toothed" pattern appears where one would expect fairly constant forward rates. Not surprisingly, exponential interpolation returns a "step-function" pattern of forward rates. In this particular example, the results obtained using the cubic spline looks fairly good, though there is a slightly disturbing "bulge" at around 5.5 years. It turns out that in some situations, when the underlying curve has jumps, the cubic spline, because it needs to be smooth and differentiable (while passing through all of the points), will bulge in regions leading to undesirable results like bad forward rates. The unfortunate conclusion is that none of these interpolation methods is perfect.

What is required is a method that combines the smoothness of cubic spline interpolation with the stability of exponential interpolation (in regions where the splines bulge). We know of no generic interpolation method that offers this. The good news is that we don't need a generic interpolation method, what we need is a method that is specific to discount factor curves. The method we have developed depends on applying a *post-smoothing* process to the curve. In this process, we *enhance* the discount factor curve (e.g. LFR) by adding points at 3-month intervals (actually at any user desired frequency). We do this in such a way as to obtain smooth forward rates, in regions where the curve is smooth, while avoiding bulges in regions where the curve has "jumps". We stress that none of the original discount factors are altered and the enhanced curve remains consistent with the original input rates. We then use this enhanced discount factor curve as the basic input to our pricing models and likely choose exponential interpolation (or even linear interpolation with this enhanced curve would be fine). The results of this curve are labeled "enhanced" in Graph 2.



Graph 2: 3-month implied forward rates

To view an article on how you could use Fincad XL and other Fincad products to build better curves, go to *www.fincad.com/curves.html.* Or, if you have any questions or feedback on this article please email Owen at *o.walsh@fincad.com*.

About FinancialCAD

FinancialCAD provides software and online services for financial risk measurement and analysis. Our industry standard financial analytics cover all asset classes and are already used in 60 countries by over 25,000 business users. For more info, visit our website at *www.fincad.com*, or download a free trial version of our software at www.fincad.com/innovate.asp.

FOOTNOTES

Mathews, H. (1987) *Numerical Methods for Computer Science, Engineering and Mathematics*. Prentice Hall Inc, Toronto.
Miron and Swannell, (1991) *Pricing and Hedging Swaps*, Euromoney Publications PLC, London.
Press, W. H., *et al.*, (1996) *Numerical Recipes in C*. Cambridge University Press, New York.

23
Stochastic Volatility Models: Past, Present and Future

Peter Jäckel

Quantitative Financial Review 2003

There are many models for the uncertainty in future instantaneous volatility. When it comes to an actual implementation of a stochastic volatility model for the purpose of the management of *exotic* derivatives, the choice of model is rarely made to capture the particular dynamical features relevant for the specific contract structure at hand. Instead, more often than not, the model is chosen that provides the greatest ease with respect to market calibration by virtue of (semi-)closed form solutions for the prices of *plain vanilla* options. In this article, the further implications of various stochastic volatility models are reviewed with particular emphasis on both the dynamic replication of exotic derivatives and on the implementation of the model. Also, a new class of models is suggested that not only allows for the level of volatility, but also for the observed *skew* to vary stochastically over time.

1 Why stochastic volatility?

- Realised volatility of traded assets displays significant variability. It would only seem natural that any model used for the hedging of derivative contracts on such assets should take into account that volatility is subject to fluctuations.
- More and more derivatives are explicitly sensitive to future (both implied and instantaneous) volatility levels. Examples are globally floored and/or capped cliquets, and many more.
- Some (apparently) comparatively straightforward exotic derivatives such as double barrier options are being re-examined for their sensitivity to uncertainty in volatility.

Contact address: ABN AMRO, 250 Bishopsgate, London EC2M 4AA, UK.
E-mail: p@jaeckel.org

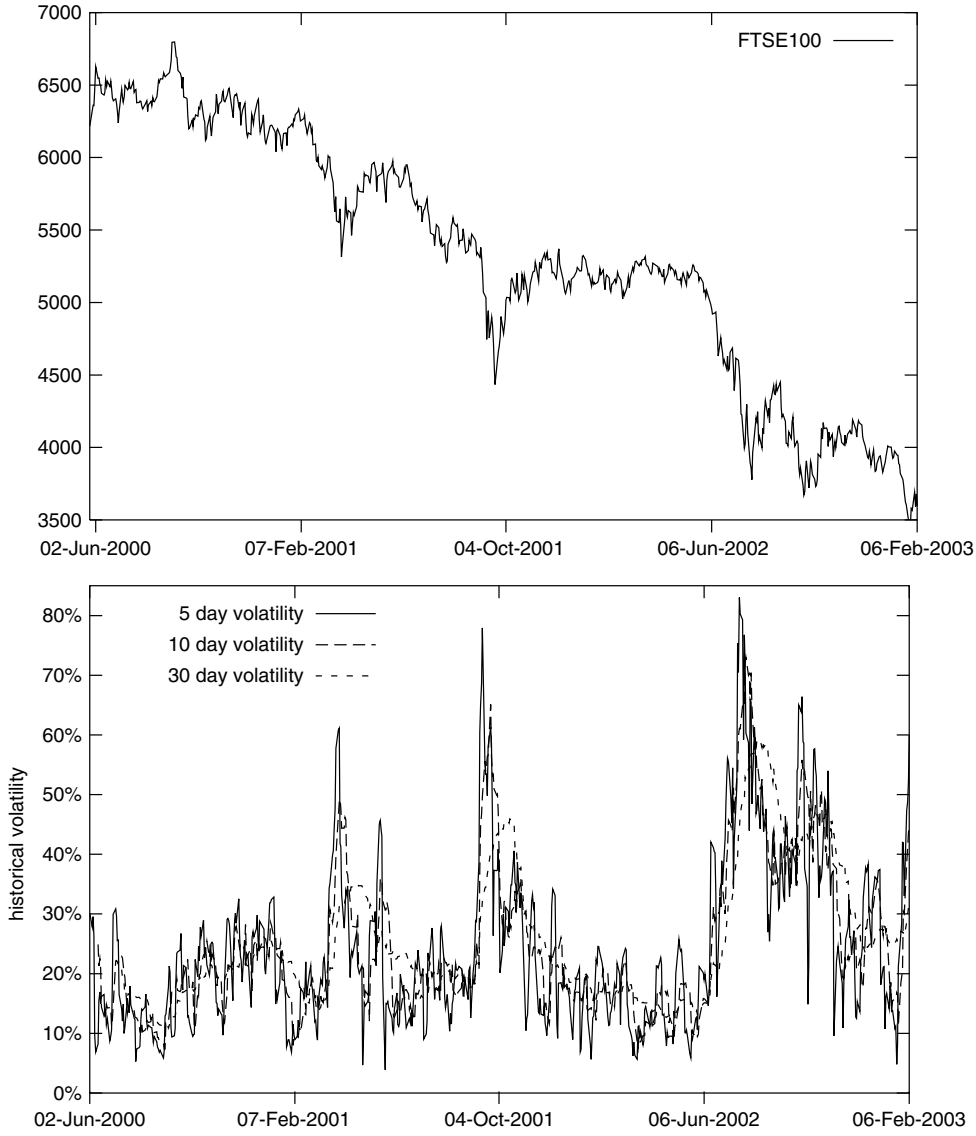


Figure 1: FTSE 100 performance and realised volatility between June 2000 and February 2003

- New trading ideas such as *exotic volatility options* and *skew swaps*, however, give rise to the need for a new kind of stochastic volatility model: the stochastic skew model.

2 What stochastic volatility?

The concept of stochastic volatility, or rather the idea of a second source of risk affecting the level of instantaneous volatility, should not be seen in isolation from the nature of the

underlying asset or deliverable contract. In particular, for the three most developed modelling domains of equity, FX, and interest rate derivatives, different effects are considered to be at least partially responsible for the smile or skew observed in the associated option markets.

Economic effects giving rise to an equity skew

- Leverage effects (Geske, 1977; Geske and Johnson, 1984; Rubenstein, 1983). A firm's value of equity can be seen as the net present value of all its future income plus its assets minus its debt. These constituents have very different relative volatilities which gives rise to a leverage related skew.
- Supply and demand. Equivalently, downwards risk insurance is more desired due to the intrinsic asymmetry of positions in equity: by their financial purpose it is more natural for equity to be held long than short, which makes downwards protection more important.
- Declining stock prices are more likely to give rise to massive portfolio rebalancing (and thus volatility) than increasing stock prices. This asymmetry arises naturally from the existence of thresholds below which positions must be cut unconditionally for regulatory reasons.

Economic effects giving rise to an FX skew and smile

- Anticipated government intervention to stabilise FX rates.
- Government changes that are expected to change policy on trade deficits, interest rates, and other economic factors that would give rise to a market bias.
- Foreign investor FX rate protection.

Economic effects giving rise to an interest rate skew and smile

- Elasticity of variance and/or mean reversion. In other words, interest rates are for economic reasons linked to a certain band. Unlike equity or FX, interest rates cannot be *split, bought back* or *re-valued* and it is this intrinsic difference that connects volatilities to absolute levels of interest rates.
- Anticipated central bank action.

None of these effects are well described by strong correlation between the asset's own driving factor and a second factor governing the uncertainty in volatility since they are all based on deterministic relationships.

Still, most stochastic volatility models incorporate a skew by virtue of strong correlation of volatility and stock. The strong correlation is usually needed to match the pronounced skew of short-dated plain vanilla options.

In this context, one might wonder if it wouldn't be more appropriate to let the stochasticity of volatility explain the market-observed features related to or associated with uncertainty in volatility, and use other mechanisms to account for the skew.

3 One model to rule them all?

An important question that must be asked when a stochastic volatility model is considered is: what is it to be used for?

- Single underlying moderate exotics with strong dependence on forward volatility? Forward fixing options such as cliquets with global floor and/or cap?
- Single underlying exotics with strong dependence on forward skew or smile? Options on variance or skew?
- Single underlying exotics with strong path dependence? Barriers of all natures (single, double, layered, range accruals).
- Pseudo-single underlying options with exposure to forward volatility of different traded contracts? Captions? Capped/floored volatility bonds? Total redemption notes with exotic coupons?
- Multiple underlying moderate exotics with strong dependence on forward volatility? Options on baskets. Cliquets on baskets.
- Multiple underlying moderate exotics with strong dependence on forward skew? Mountain range or rainbow options.
- Multiple underlying moderate exotics with strong dependence on correlation? Mountain range options.

Not all of these applications would necessarily suggest the use of the same model. In fact, a stochastic volatility model that can be perfectly adequate to capture the risk in one of the above categories may completely miss the exposures in other products. As an example, consider the use of a conventional stochastic volatility model for the management of options on variance swaps versus the use of the same model for options on future market skew in the plain vanilla option market.

4 Mathematical features of stochastic volatility models

Heston [Hes93]: $V[\sigma_S^2] \sim \mathcal{O}(\sigma_S^2)$ (mean reverting)

$$\mathrm{d}S = \mu S\,\mathrm{d}t + \sqrt{v}S\,\mathrm{d}W_S \tag{1}$$

$$\mathrm{d}v = \kappa(\theta - v)\,\mathrm{d}t + \alpha\sqrt{v}\,\mathrm{d}W_v \tag{2}$$

$$\mathrm{E}[\mathrm{d}W_S \cdot \mathrm{d}W_v] = \rho\,\mathrm{d}t \tag{3}$$

In order to achieve calibration to the market given skew, almost always one needs to have:

- $0.7 < |\varrho| \lesssim 1$ is required.
- κ must be very small (*kappa kills the skew*).
- α must be sizeable.
- θ is by order of magnitude not too far away from the implied volatility of the longest dated option calibrated to.

The volatility process can reach zero unless [Fel51, RW00]:

$$\kappa\theta > \tfrac{1}{2}\alpha^2 \tag{4}$$

which is hardly ever given in a set of parameters calibrated to market! This means the Heston model achieves calibration to today's observed plain vanilla option prices by balancing the probabilities of very high volatility scenarios against those where future instantaneous volatility drops to very low levels. The average time volatility stays at high or low levels and is measured by the mean reversion scale $1/\kappa$. Even when $\kappa\theta > \frac{1}{2}\alpha^2$, the long-term distribution of $\int_t^{t+\tau} \sigma(t)^2 \mathrm{d}t$ is sharply peaked at low values of volatility as a result of calibration.[1]

The dynamics of the calibrated Heston model predict that volatility can reach zero, stay at zero for some time, or stay extremely low or very high for long periods of time.

Stein and Stein/Schöbl and Zhu [SS91, SZ99]: $V[\sigma_S] \sim \mathrm{O}(1)$ (mean reverting)

$$\mathrm{d}S = \mu S\,\mathrm{d}t + \sigma S\,\mathrm{d}W_S \tag{5}$$

$$\mathrm{d}\sigma = \kappa(\theta - \sigma)\,\mathrm{d}t + \alpha\,\mathrm{d}W_\sigma \tag{6}$$

$$\mathrm{E}[\mathrm{d}W_S \cdot \mathrm{d}W_\sigma] = \rho\,\mathrm{d}t \tag{7}$$

The distribution of volatility converges to a Gaussian distribution with mean θ and variance $\frac{\alpha^2}{2\kappa}$. Since the sign of σ bears meaning only as a sign modifier of the correlation, we have the following two consequences:

- The sign of correlation between movements of the underlying and volatility can suddenly switch.
- The level of volatility has its most likely value at zero.

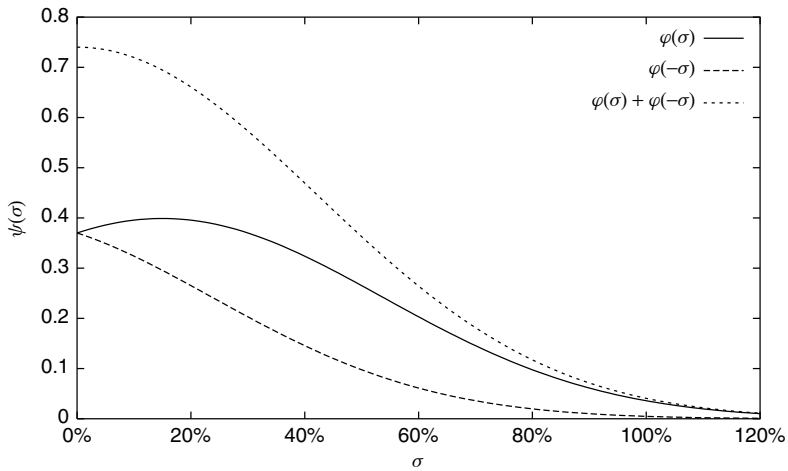


Figure 2: Stationary Stein and Stein volatility distribution for $\alpha = 0.3$, $\kappa = 0.3$, and $\theta = 0.25$

[1] See http://www.dbconvertibles.com/dbquant/Presentations/LondonDec2002RiskTraining Volatility.pdf, slides 33–35, for diagrams on this feature.

The dynamics of the Stein and Stein/Schöbl and Zhu model predict that volatility is very likely to be near zero and that the sign of correlation with the spot movement driver can switch.

Hull–White [HW87]: $V[\sigma_S^2] \sim \mathrm{O}(\sigma_S^4)$ (zero reverting for $\mu_v < 0$)

$$\mathrm{d}S = \mu_S S\,\mathrm{d}t + \sqrt{v} S\,\mathrm{d}W_S \tag{8}$$

$$\mathrm{d}v = \mu_v v\,\mathrm{d}t + \xi v\,\mathrm{d}W_\sigma \tag{9}$$

$$\mathrm{E}[\mathrm{d}W_S \cdot \mathrm{d}W_\sigma] = \rho\,\mathrm{d}t \tag{10}$$

Since v is lognormally distributed in this model, and since $\sigma = \sqrt{v}$, we have:

$$\mathrm{E}[\sigma(t)] = \sigma(0) \cdot \mathrm{e}^{\frac{1}{2}\mu_v t - \frac{1}{8}\xi^2 t} \tag{11}$$

$$\mathrm{V}[\sigma(t)] = \sigma(0)^2 \cdot \mathrm{e}^{\mu_v t} \cdot (1 - \mathrm{e}^{-\frac{1}{4}\xi^2 t}) \tag{12}$$

$$\mathrm{M}[\sigma(T)] = \sigma(0) \cdot \mathrm{e}^{\frac{1}{2}(\mu_v - \xi^2)t} \tag{13}$$

where $\mathrm{M}[\cdot]$ is defined as the most likely value. This means, for $\mu_v < \frac{1}{4}\xi^2$, the expectation of volatility converges to the mean-reversion level at zero. For $\mu_v > \frac{1}{4}\xi^2$, the expectation diverges. Further, unless $\mu_v < 0$, the variance of volatility grows unbounded. In contrast to that, if $\mu_v < 0$, the variance of variance diminishes over time. And finally, the most likely value for volatility converges to zero unless $\mu_v > \xi^2$. For the particular case of $\mu_v = 0$, we have the special combination of features that the expectation and most likely value of volatility converges to zero, whilst the variance of volatility converges to σ^2.

Any choice of parameters that provides a reasonable match of market given implied volatilities is extremely likely to lead to $\mu_v < 0$ in which case we have:

The dynamics of the Hull–White stochastic volatility model predict that both expectation and most likely value of instantaneous volatility converge to zero.

Hagan [HKL02]: $V[\sigma_S] \sim \mathrm{O}(\sigma_S^2)$ (*not* mean reverting)

$$\mathrm{d}S = \mu S\,\mathrm{d}t + \sigma S\,\mathrm{d}W_S \tag{14}$$

$$\mathrm{d}\sigma = \alpha\sigma\,\mathrm{d}W_\sigma \tag{15}$$

$$\mathrm{E}[\mathrm{d}W_S \cdot \mathrm{d}W_\sigma] = \rho\,\mathrm{d}t \tag{16}$$

This model is equivalent to the Hull–White stochastic volatility model for the special case of $\mu_v = \alpha^2$ and $\xi = 2\alpha$. In this model, instantaneous volatility is a martingale but the variance of volatility grows unbounded. At the same time, the *most likely* value for volatility converges to zero.

The dynamics of the Hagan model predict that the expectation of volatility is constant over time, that variance of instantaneous volatility grows without limit and that the most likely value of instantaneous volatility converges to zero.

Scott and Scott–Chesney [Sco87, CS89]: $V[\sigma_S] \sim \mathrm{O}(\sigma_S^2)$ (mean reverting)

$$\mathrm{d}S = \mu S\,\mathrm{d}t + \mathrm{e}^{y} S\,\mathrm{d}W_S \tag{17}$$

$$\mathrm{d}y = \kappa(\theta - y)\mathrm{d}t + \alpha\,\mathrm{d}W_y \tag{18}$$

$$\mathrm{E}[\mathrm{d}W_S \cdot \mathrm{d}W_y] = \rho\,\mathrm{d}t \tag{19}$$

Volatility cannot reach zero, nor does its most likely value converge there. The market-observable skew of implied volatilities would require a strong negative correlation for this model to be calibrated.

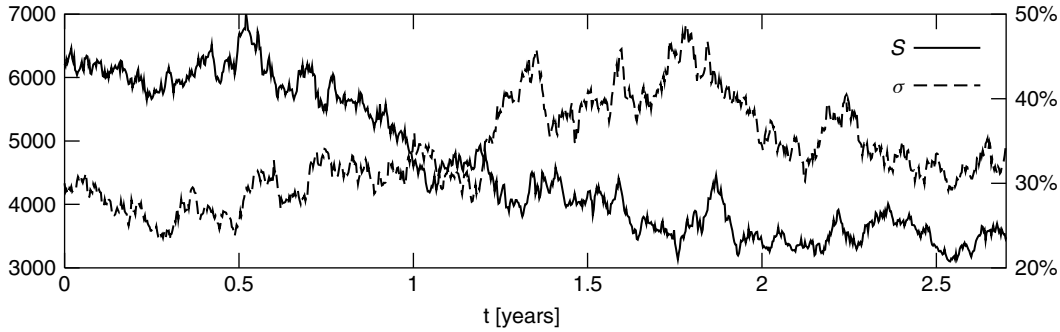


Figure 3: Sample path for Scott–Chesney model with $S_0 = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = 30\%$, $\theta = \ln 30\%$, $\kappa = 0.1$, $\alpha = 40\%$, $\frac{\alpha^2}{2\kappa} = 0.8$, $\rho = 0$. Euler integration with $\Delta t = 1/365$

However, the required strong correlation between volatility and spot is not supported by any econometric analysis. Nonetheless, it is possible to reproduce the burstiness of real volatility returns by increasing the mean reversion. Fouquet *et al*. (2000) compare strong mean reversion dynamics with real data and find that it captures the apparent burstiness of realised volatilities very well [FPS00]:

- The larger κ, the more rapidly the volatility distribution converges to its stationary state.
- $1/\kappa$ is the time scale for volatility auto-decorrelation.
- The right measure for uncertainty in volatility is:

$$\frac{\alpha^2}{2\kappa}$$

not α on its own.

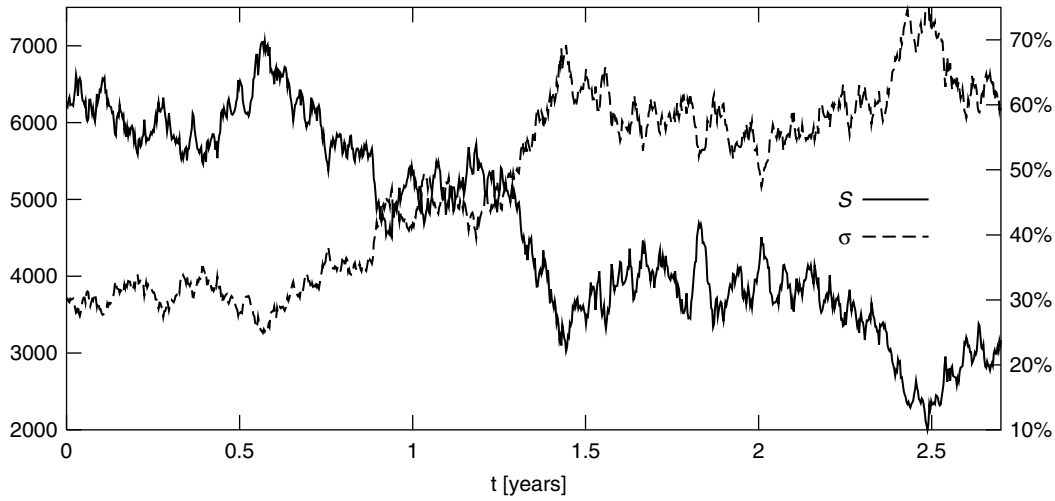


Figure 4: Sample path for Scott–Chesney model with $S_0 = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = 30\%$, $\theta = \ln 30\%$, $\kappa = 0.1$, $\alpha = 40\%$, $\frac{\alpha^2}{2\kappa} = 0.8$, $\rho = -0.9$. Euler integration with $\Delta t = 1/365$

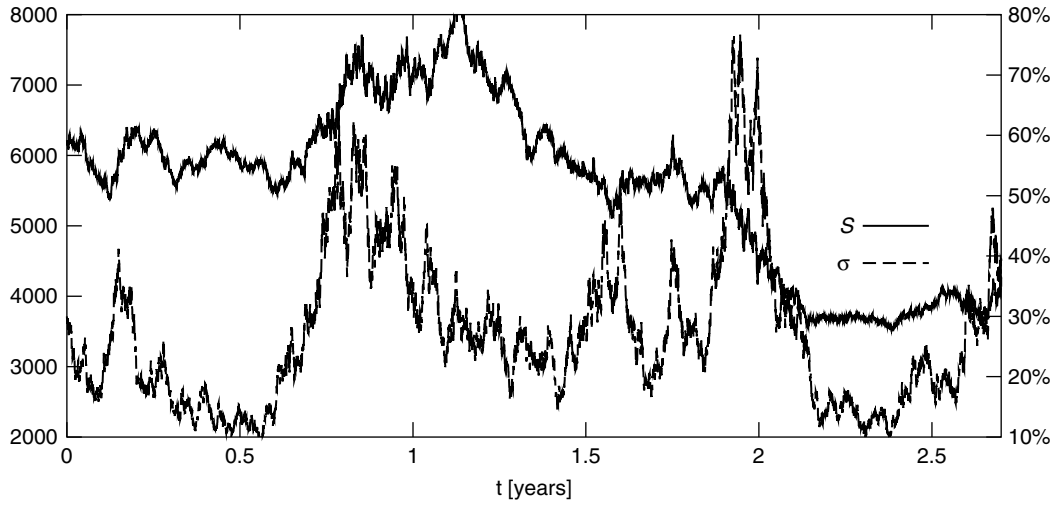


Figure 5: Sample path for Scott–Chesney model with $S_0 = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = 30\%$, $\theta = \ln 30\%$, $\kappa = 6$, $\alpha = 1.5$, $\frac{\alpha^2}{2\kappa} = 0.1875$, $\rho = 0$. Euler integration with $\Delta t = 1/2920$

Large mean reversion causes volatility to approach its stationary distribution quickly. The problem with future volatility being likely to hover near zero for models such as the Heston and the Stein and Stein model goes away when mean reversion is strong. However, if mean reversion is large, correlation between volatility and spot does not suffice to generate a significant skew. To achieve market calibration, a different mechanism is needed. This could be independent jumps of the stock itself, or a stock-dependent volatility scaling function.

The main drawback of the Scott–Chesney model is that it requires very high correlation between the spot and the volatility process to calibrate to a pronounced skew, and that the skew is fully deterministic. These features are also shared by all of the above-discussed models.

5 A stochastic skew model

$$\mathrm{d}S = \mu S\,\mathrm{d}t + \sigma f(S;\gamma) S\,\mathrm{d}W_S \tag{20}$$

$$\mathrm{d}\ln\sigma = \kappa_\sigma(\ln\sigma_\infty - \ln\sigma)\,\mathrm{d}t + \alpha_\sigma\,\mathrm{d}W_\sigma \tag{21}$$

$$\mathrm{d}\gamma = \kappa_\gamma(\gamma_\infty - \gamma)\,\mathrm{d}t + \alpha_\gamma\,\mathrm{d}W_\gamma \tag{22}$$

with:

$$f(S;\gamma) = \mathrm{e}^{\gamma\cdot(\frac{S}{H}-1)} \tag{23}$$

and:

$$\mathrm{E}[\mathrm{d}W_\sigma \mathrm{d}W_\gamma] = \mathrm{E}[\mathrm{d}W_\sigma \mathrm{d}W_S] = \mathrm{E}[\mathrm{d}W_\gamma \mathrm{d}W_S] = 0 \tag{24}$$

This scaling ensures that:

- For negative γ, the local volatility scaling factor decays from $\mathrm{e}^{-\gamma}$ for $S \to 0$ to 0 for $S \to \infty$.
- The local volatility scaling factor f at spot level H is exactly 1.
- The local volatility scaling factor f change for a spot move of $\delta \cdot H$ near H is given by:

$$\Delta f = \left.\frac{\partial f}{\partial S}\right|_{S=H} \cdot \delta \cdot H = \frac{\gamma}{H} \cdot \delta \cdot H = \delta \cdot \gamma \tag{25}$$

In other words, γ is a measure for the local volatility skew at H.

Maintenance of correlation matrices is greatly simplified by the assumption of independence of the individual factors. The associated partial differential equation governing the boundary value problem of derivatives prices is:

$$V_t + \underbrace{\left(\mu - \tfrac{1}{2}\mathrm{e}^{2y} f^2(\mathrm{e}^x;\gamma)\right)}_{\widehat{\mu}_x} V_x + \underbrace{\kappa_\sigma\,(\ln\sigma_\infty - y)}_{\widehat{\mu}_y} V_y + \underbrace{\kappa_\gamma\,(\gamma_\infty - \gamma)}_{\widehat{\mu}_\gamma} V_\gamma + \tfrac{1}{2}\mathrm{e}^{2y} f^2(\mathrm{e}^x;\gamma) V_{xx} + \tfrac{1}{2}\alpha_\sigma^2 V_{yy} + \tfrac{1}{2}\alpha_\gamma^2 V_{\gamma\gamma} = r \cdot V \tag{26}$$

with

$$x = \ln S \qquad \text{and} \qquad y = \ln\sigma \tag{27}$$

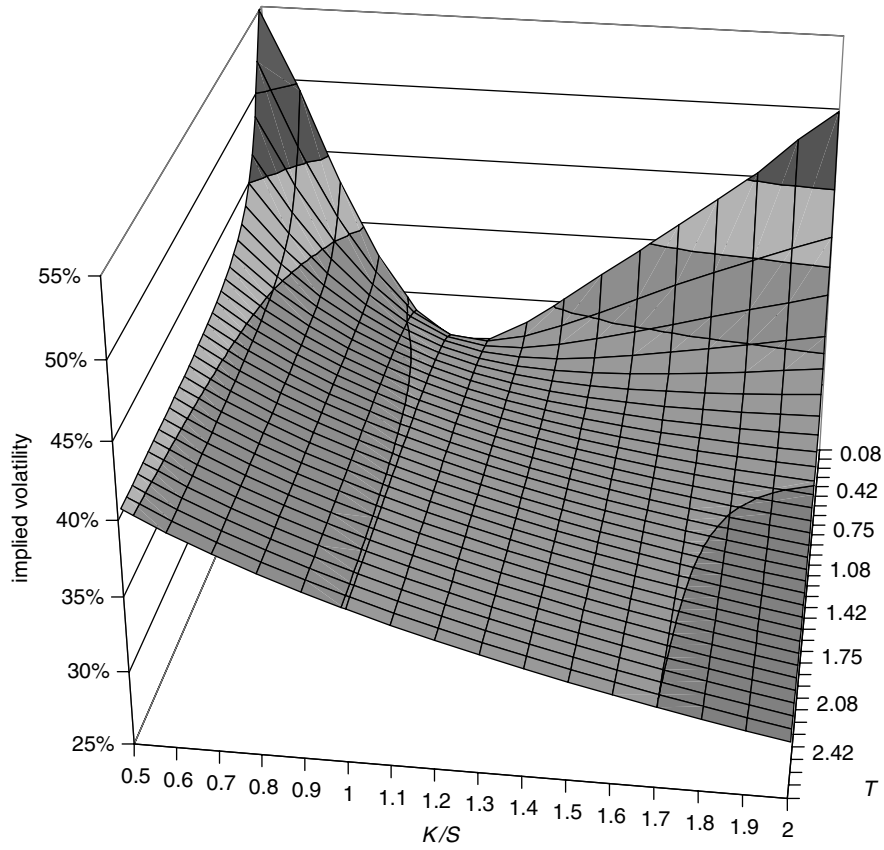


Figure 6: Implied volatility surface for stochastic exponential skew model with $S_0 = H = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = \sigma_\infty = 30\%$, $\kappa_\sigma = 12$, $\alpha_\sigma = 2$, $\sqrt{\frac{\alpha_\sigma^2}{2\kappa_\sigma}} = 41\%$, $\gamma_0 = \gamma_\infty = -0.5$, $\kappa_\gamma = 4$, $\alpha_\gamma = 0.5$, $\sqrt{\frac{\alpha_\gamma^2}{2\kappa_\gamma}} = 0.18$

Jumps without jumps

The exponential dependence of the volatility scaling function f on the spot level S can lead to jump-like upward (for $\gamma > 0$) or downward (for $\gamma < 0$) rallies when $|\gamma|$ is of significant size. A sample path showing this behaviour is given in Figure 7. This can happen due to the exponential nature of the scaling function f, especially during periods of increased $|\gamma|$. These events only occur when the skew is very pronounced as shown, for example, in Figure 8.

A hyperbolic alternative

The shown implosions of the spot are caused by the exponential form of the scaling function f and are technically akin to process explosions seen also for the short rate in a lognormal HJM setting and other equations involving a locally exponential scaling of volatility. Naturally, it is straightforward to use other scaling functions that avoid the spot implosions, should they be undesirable.

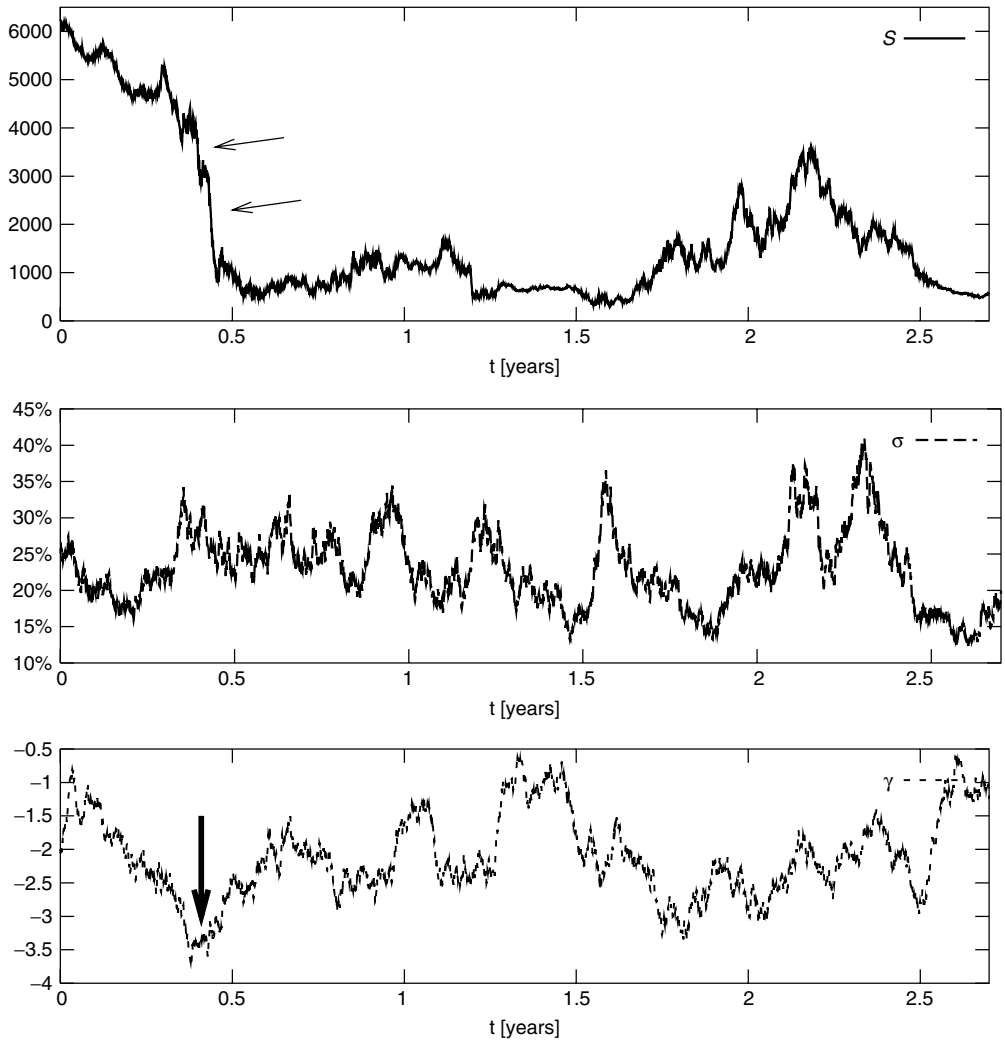


Figure 7: Jump-like almost instantaneous downwards corrections of the spot for $S_0 = H = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = \sigma_\infty = 25\%$, $\kappa_\sigma = 6$, $\alpha_\sigma = 1$, $\sqrt{\frac{\alpha_\sigma^2}{2\kappa_\sigma}} = 29\%$, $\gamma_0 = \gamma_\infty = -2$, $\kappa_\gamma = 3$, $\alpha_\gamma = 2$, $\sqrt{\frac{\alpha_\gamma^2}{2\kappa_\gamma}} = 0.82$

An alternative to the exponential scaling is the hyperbolic function:

$$f = \gamma\left(\frac{S}{H} - 1\right) + \sqrt{\gamma^2\left(\frac{S}{H} - 1\right)^2 + (1-\eta)^2} + \eta \tag{28}$$

This model also allows for a wide variety of shapes of the implied volatility surface.

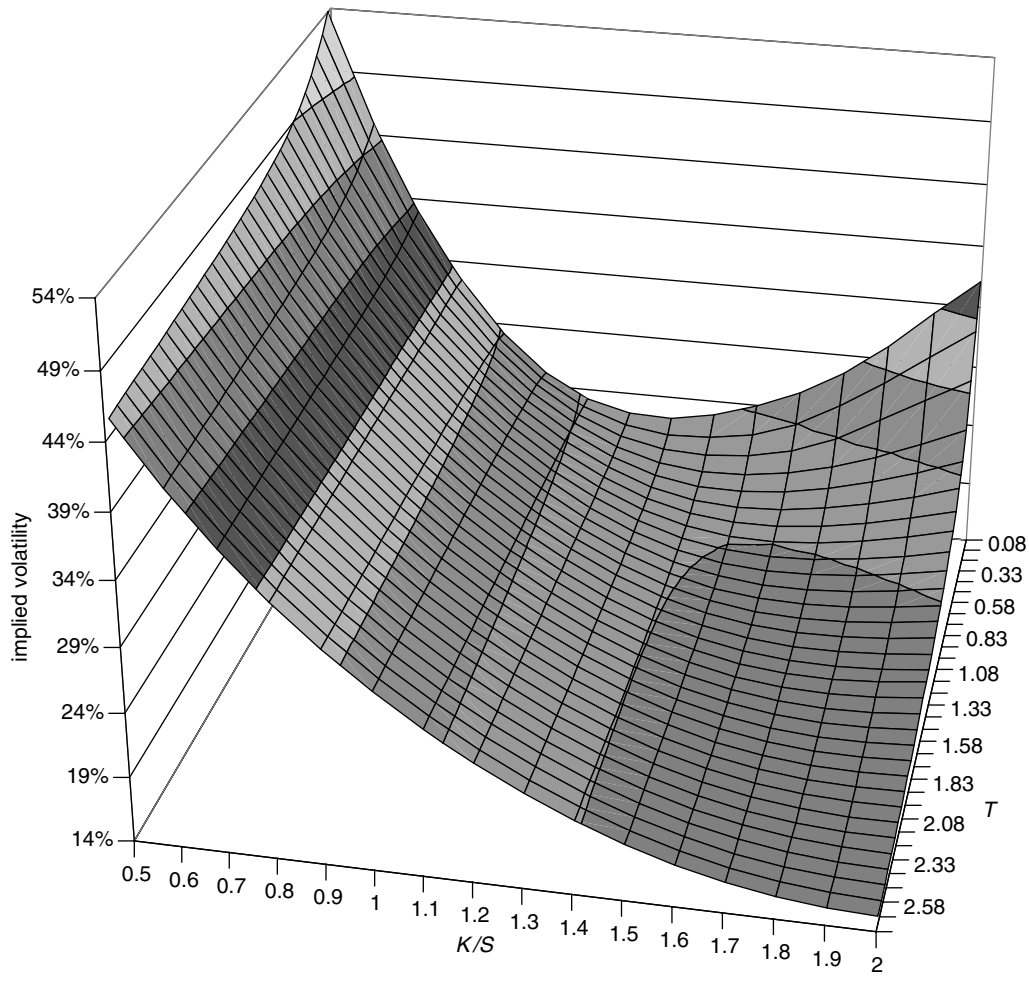


Figure 8: Implied volatility surface for stochastic exponential skew model with $S_0 = H = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = \sigma_\infty = 25\%$, $\kappa_\sigma = 6$, $\alpha_\sigma = 1$, $\sqrt{\frac{\alpha_\sigma^2}{2\kappa_\sigma}} = 29\%$, $\gamma_0 = \gamma_\infty = -2$, $\kappa_\gamma = 3$, $\alpha_\gamma = 2$, $\sqrt{\frac{\alpha_\gamma^2}{2\kappa_\gamma}} = 0.82$

6 Monte Carlo methods and stochastic volatility models

The Heston model is often used to parametrise the observed market volatilities since there are semi-analytical solutions for plain vanilla options under this model. However, when multi-asset derivatives are priced, we often need to resort to numerical integration of the governing stochastic differential equations.

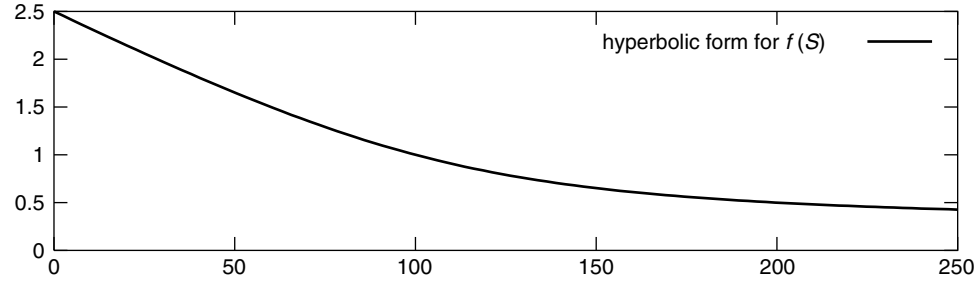


Figure 9: Hyperbolic example for the scaling function $f(S)$ with $\gamma = -1$, $H = 100$, and $\eta = 1/4$

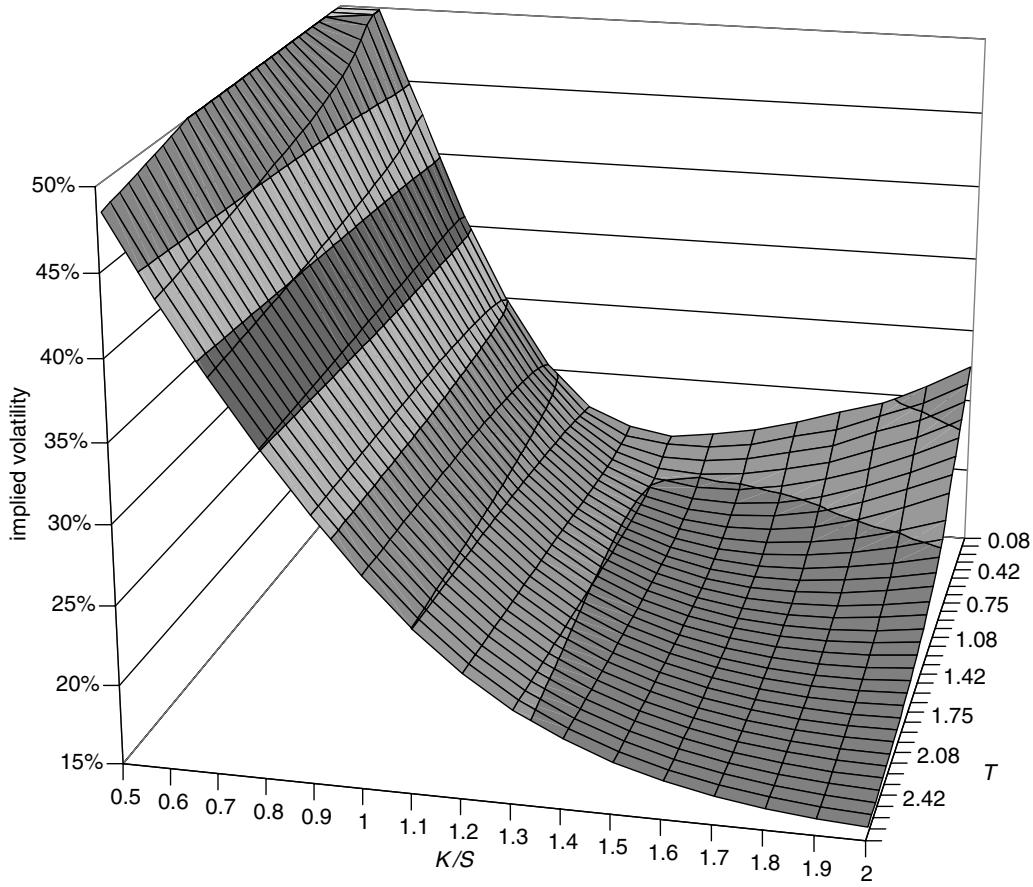


Figure 10: Implied volatility surface for stochastic hyperbolic skew model with a hyperbolic scaling function f and $S_0 = H = 6216$, $r = 5\%$, $d = 1\%$, $\sigma_0 = \sigma_\infty = 25\%$, $\kappa_\sigma = 6$, $\alpha_\sigma = 1$, $\sqrt{\frac{\alpha_\sigma^2}{2\kappa_\sigma}} = 28.87\%$, $\gamma_0 = \gamma_\infty = -3$, $\kappa_\gamma = 3$, $\alpha_\gamma = 3$, $\sqrt{\frac{\alpha_\gamma^2}{2\kappa_\gamma}} = 1.22$, and $\eta = 1/4$

The Euler discretisation of the Heston variance process is given by:

$$\Delta v = \kappa(\theta - v)\Delta t + \alpha\sqrt{v}\sqrt{\Delta t}\cdot z \tag{29}$$

with $z \sim \mathcal{N}(0, 1)$. This means for $z < z^*$ with:

$$z^* = -\frac{v + \kappa(\theta - v)\Delta t}{\alpha\sqrt{v\Delta t}} \tag{30}$$

the Euler step causes variance to cross over to the negative domain!

A popular method of choice to avoid this artifact of Euler integration is to use Itô's lemma to transform to coordinates where the Euler step remains in the domain of validity for all possibly drawn Gaussian variates. For the Heston variance process, the coordinate we have to transform to is volatility itself:

$$\mathrm{d}\sigma = \frac{\kappa}{2}\left[\frac{1}{\sigma}\left(\theta - \frac{\alpha^2}{4\kappa}\right) - \sigma\right]\mathrm{d}t + \frac{1}{2}\alpha\,\mathrm{d}W \tag{31}$$

Alas, it seems we have transformed ourselves from the pan into the fire: whilst equation (2), for $v \to 0$, would always show a positive drift term for all $\theta > 0$ no matter how close v came to zero, and only the diffusion component could make it reach zero, the drift term in equation (31) diverges to negative infinity if $\theta < \frac{\alpha^2}{4\kappa}$ irrespective of the path taken by the diffusion component. This means that the transformed equation shows strong (drift-dominated) absorption into zero near zero, whilst the original stochastic differential equation for the variance only exhibits zero as an attainable boundary due to the diffusion component being able to overcome the mean reversion effect (i.e. the positive drift) for $2\theta\kappa < \alpha^2$.

The apparently contradictory behaviour near zero has a simple explanation:

In an infinitesimal neighbourhood of zero, Itô's lemma cannot be applied to the variance process (2). The transformation of the variance process to a volatility formulation results in a structurally different process!

Naturally, this feature raises its ugly head in any numerical implementation where we may prefer to use a transformed version of the original equations!

An alternative, when suitable transformations are not available, is to use *implicit* or *mixed* Euler schemes [KP99] in order to ensure that the stepping algorithm does not cause the state variable to leave the domain of the governing equations, possibly in conjunction with Doss's method[2] of constructing pathwise solutions.

An example for such an approach is as follows. First, let us assume that we have discretised the evolution of time into a sequence of time intervals $[t_n, t_{n+1}]$, and that we have drawn an independent Wiener path over those points in time, i.e. that we know $W(t_n)$ for all n for one specific path. Then, approximate $W(t)$ as a piecewise linear function in between the known values at t_n and t_{n+1}, i.e.:

$$W(t) \simeq \gamma_n + \delta_n t \quad \text{for} \quad t \in [t_n, t_{n+1}] \tag{32}$$

[2] See [Dos77] or [KS91] (pages 295–296).

with

$$\gamma_n = W(t_n) - \delta_n t_n \quad \text{and} \quad \delta_n = \frac{W(t_n) - W(t_{n+1})}{t_n - t_{n+1}}$$

Using the resulting dependency $\mathrm{d}W = \delta_n \mathrm{d}t$, this gives us the approximate ordinary differential equation:

$$\frac{\mathrm{d}v}{\mathrm{d}t} \simeq \varkappa(\theta - v) + \alpha\delta_n\sqrt{v} \tag{33}$$

which has the implicit solution:

$$t - t_n = T(v(t)) - T(v(t_n)) \tag{34}$$

with:

$$T(v) = \frac{2\alpha\delta_n}{\kappa\sqrt{\alpha^2\delta_n^2 + 4\theta\kappa^2}}\operatorname{atanh}\left(\frac{2\kappa\sqrt{v} - \alpha\delta_n}{\sqrt{\alpha^2\delta_n^2 + 4\theta\kappa^2}}\right) - \frac{1}{\kappa}\ln(\kappa(v-\theta) - \alpha\delta_n\sqrt{v}) \tag{35}$$

The above equation can be solved numerically comparatively readily since we know that, given δ_n, over the time step from t_n to t_{n+1}, v will move monotonically, and that in the limit of $\Delta t_n := (t_{n+1} - t_n) \to \infty$, for fixed δ_n, we have:

$$\lim_{\Delta t_n \to \infty} v_{n+1} = \left(\frac{\alpha\delta}{2\kappa} + \sqrt{\left(\frac{\alpha\delta}{2\kappa}\right)^2 + \theta}\right)^2 \tag{36}$$

which can be computed by setting the argument of the logarithm in the right hand side of equation (35) to zero. An example for the paths of $\sqrt{v}$ over a single unit time step for different draws of δ is shown in Figure 11. Putting all of the above together enables us to construct paths for the stochastic variance without the need for very small time steps.

The explicit knowledge of the functional form of the volatility, or variance, path has another advantage. Fouquet *et alii* [FPS00] explain how we can directly draw the logarithm of the spot level at the end of a large time step $(t_{n+1} - t_n)$ if we can explicitly compute, for the given volatility or variance path, the quantities:

$$\Delta\hat{v}_n := \int_{t_n}^{t_{n+1}} \sigma^2(t)\,\mathrm{d}t \tag{37}$$

$$\Delta\hat{\omega}_n := \int_{t_n}^{t_{n+1}} \sigma(t)\,\mathrm{d}W(t) \tag{38}$$

The first term poses no difficulty since the primitive of $T(v)$ can be computed analytically and:

$$\int_{t_n}^{t_{n+1}} \sigma^2(t)\,\mathrm{d}t = v(t_{n+1}) \cdot [T(v(t_n)) + \Delta t_n] - v(t_n) \cdot T(v(t_n)) - \int_{v_n}^{v_{n+1}} T(v)\,\mathrm{d}v \tag{39}$$

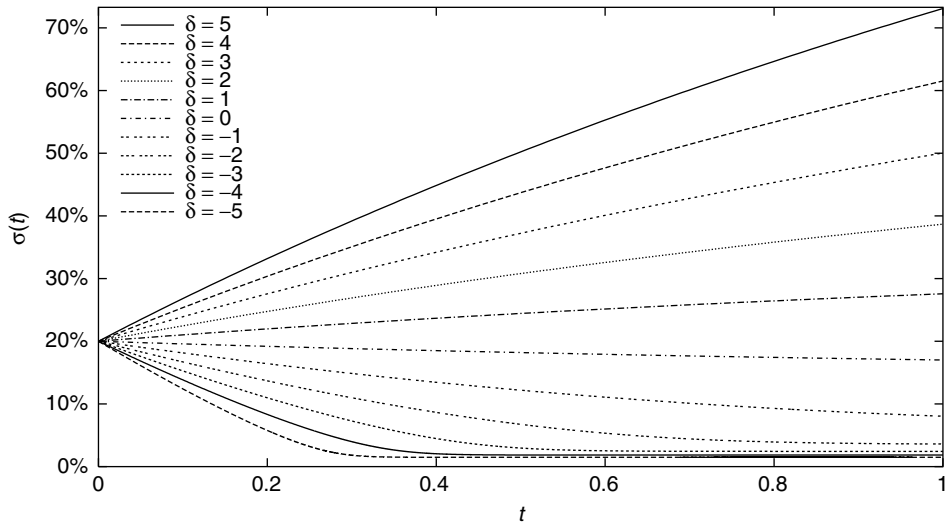


Figure 11: Sample paths for the Heston volatility process for $\sigma(0) = \sqrt{v(0)} = 20\%$, $\sqrt{\theta} = 15\%$, $\alpha = 30\%$, $\kappa = 1$ over a unit time step for different levels of the variate $\delta = W_v(1) - W_v(0)$ when $W_v(t)$ is approximated linearly over the time step

The second term requires another approximate numerical scheme which will also be no major obstacle since:

$$\int_{t_n}^{t_{n+1}} \sigma(t)\,\mathrm{d}W(t) \simeq \delta_n \cdot \int_{t_n}^{t_{n+1}} \sigma(t)\,\mathrm{d}t$$

$$= \delta_n \cdot \left(\sigma(t_{n+1}) \cdot [T(v(t_n)) + \Delta t_n] - \sigma(t_n) \cdot T(v(t_n)) - \int_{v_n}^{v_{n+1}} T(v)/(2\sqrt{v})\,\mathrm{d}v \right) \tag{40}$$

The numerical approximation is needed for the calculation of the integral on the right hand side of equation (40). A simple Simpson scheme or Legendre quadrature is likely to produce excellent results given that the function is guaranteed to be monotonic and smooth. Using all of the above quantities, the draw for the logarithm of the spot level can be constructed as:

$$\ln S_{n+1} = \ln S_n + \mu \Delta t_n - \tfrac{1}{2}\Delta \hat{v}_n + \rho \Delta \hat{\omega}_n + \sqrt{1-\rho^2} \cdot \sqrt{\Delta \hat{v}_n} \cdot z \tag{41}$$

where z is a standard normal variate that is independent from the variate used to construct the variance step $v_n \to v_{n+1}$. The above scheme is essentially an extension of the root-mean-square volatility lemma given in [HW87] beyond the case of $\rho = 0$.

In comparison, the Stein and Stein/Schöbl and Zhu, Hull–White, Hagan and Scott/Scott–Chesney model can be simulated much more easily since the stochastic differential equation for the volatility component has simple analytical solutions. Naturally, similar techniques to the one elaborated above for the Heston model can be used to obviate the need for very small time steps.

7 Finite differencing methods and stochastic volatility models

Whenever we have non-zero correlation between the different factors, we cannot use Alternating Direction Implicit methods (unless we transform away the correlation term which is usually very bad for the handling of boundary conditions, or we combine it with an explicit method for the cross terms which makes the scheme effectively explicit). Explicit methods, however, require rather small time steps in order to avoid explosions due to numerical instabilities.

The multi-dimensional equivalent of the Crank–Nicolson method[3] can be implemented efficiently using iterative solver algorithms such as the stabilised biconjugate gradient method [GL96] that don't require the explicit specification of the discretised differential operator matrix at all. Making the need for an explicit representation of the matrix redundant, sparsely encoded or otherwise, is a major advantage since the main part of the solving algorithm does thus not depend on the explicit form of the matrix. For the use of iterative methods, all that is needed is a function that carries out the same calculations that would be done in an explicit method. A useful collection of utilities for this purpose is the Iterative Template Library [LLS].

However, when there are no correlation terms, such as in the case of the stochastic skew model, the independence of the three factors makes it possible to use an *operator split* algorithm. The simplest versions of operator-splitting algorithms in two diffusion dimensions are also known as *alternating direction implicit* methods. The use of these methods allows us to propagate over large time steps in a very fast finite differencing scheme. The main reason for operator splitting schemes to be still slower in more than one dimension is simply the fact that, typically, the total number of nodes in any time-(hyper)slice grows like the product of spatial levels in each of the diffusion factors. For the stochastic skew model, since we have zero correlation, the number of discretisation layers in both the volatility and the skew factor can be kept small ($\sim$ 20–30). Also, boundary conditions can be kept simple in all directions and in the corners: $V_{ii} = 0$ for $i = x, y, \gamma$. All in all, the speed of a three factor operator-splitting implementation when two diffusion dimensions can be discretised rather coarsely, i.e. with few numbers of spatial levels, is actually well compatible with that of any *safe* implementation involving numerical contour integrals or Fourier inversions of characteristic functions etc.

The generalisation of alternating direction implicit (or alternating direction Crank–Nicolson) to multiple spatial dimensions is based on the idea of an *operator split* [PR55, DR56, Mar89]. Take the equation:

$$V_t + \sum_i \widehat{\mu}_i(t, \boldsymbol{x}) V_{x_i} + \tfrac{1}{2} \sum_i \sigma_i^2(t, \boldsymbol{x}) V_{x_i x_i} = r \cdot V \tag{42}$$

transform away the source term by setting $u := V \mathrm{e}^{-rt}$ (which may change your boundary conditions):

$$u_t + \sum_i \underbrace{\left(\widehat{\mu}_i \partial_{x_i} + \tfrac{1}{2}\sigma_i^2 \partial_{x_i}^2\right)}_{L_i} \cdot u = 0 \tag{43}$$

[3] Also denoted as Peaceman–Rachford–Douglas method [PR55, DR56].

i.e.

$$(\partial_t + L) \cdot u = 0 \qquad \text{with} \qquad L = \sum_i L_i \tag{44}$$

Finite differencing of accuracy order $\mathcal{O}(\Delta t^2)$ and $\mathcal{O}(\Delta t^2)$ amounts to:

$$\partial_t \cdot u \to \frac{1}{\Delta t}[u(t, \boldsymbol{x}) - u(t - \Delta t, \boldsymbol{x})] \tag{45}$$

$$\partial_{x_i} \cdot u \to \frac{1}{2\Delta x_i}[u(t, \ldots, x_i + \Delta x_i, \ldots) - u(t, \ldots, x_i - \Delta x_i, \ldots)] \tag{46}$$

$$\partial_{x_i}^2 \cdot u \to \frac{1}{\Delta x_i^2}[u(t, \ldots, x_i + \Delta x_i, \ldots) - 2u(t, \ldots, x_i, \ldots) + u(t, \ldots, x_i - \Delta x_i, \ldots)] \tag{47}$$

Discretisation of the differential operators yields $L_i \to D_i$ with:

$$\begin{aligned} D_i \cdot u(t, \boldsymbol{x}) = {} & \widehat{\mu}_i(t, \boldsymbol{x}) \frac{1}{2\Delta x_i}[u(t, \ldots, x_i + \Delta x_i, \ldots) - u(t, \ldots, x_i - \Delta x_i, \ldots)] \\ & + \frac{1}{2}\sigma_i^2(t, \boldsymbol{x}) \frac{1}{\Delta x_i^2}[u(t, \ldots, x_i + \Delta x_i, \ldots) \\ & - 2u(t, \boldsymbol{x}) + u(t, \ldots, x_i - \Delta x_i, \ldots)] \end{aligned} \tag{48}$$

$$\begin{aligned} D_i \cdot u(\boldsymbol{x}) = {} & \frac{1}{2\Delta x_i^2}[(\sigma_i^2(t, \boldsymbol{x}) + \widehat{\mu}_i(t, \boldsymbol{x})\Delta x_i)u(\ldots, x_i + \Delta x_i, \ldots) - 2\sigma_i^2(t, \boldsymbol{x})u(\boldsymbol{x}) \\ & + (\sigma_i^2(t, \boldsymbol{x}) - \widehat{\mu}_i(t, \boldsymbol{x})\Delta x_i)u(\ldots, x_i + \Delta x_i, \ldots)] \end{aligned} \tag{49}$$

The Crank–Nicolson algorithm means:

$$(\partial_t + L) \cdot u(t, \boldsymbol{x}) = 0 \tag{50}$$

is to be approximated by:

$$\frac{1}{\Delta t}[u(t, \boldsymbol{x}) - u(t - \Delta t, \boldsymbol{x})] + \frac{1}{2} D \cdot [u(t, \boldsymbol{x}) + u(t - \Delta t, \boldsymbol{x})] = 0 \tag{51}$$

This means that a single step in the Crank–Nicolson scheme is given by solving

$$(\mathbf{1} - \tfrac{1}{2}\Delta t D) \cdot u(t - \Delta t, \boldsymbol{x}) = (\mathbf{1} + \tfrac{1}{2}\Delta t D) \cdot u(t, \boldsymbol{x}) \tag{52}$$

for $u(t - \Delta t, \boldsymbol{x})$.

The *operator split* of the discretised operator $D = \sum_i D_i$ is to split D into its commuting components $\{D_i\}$, and to solve (52) for each of the D_i individually in sequence. A single time step in the n-dimensional operator-split finite differencing scheme is thus given by a sequence of n one-dimensional finite differencing steps. Solve:

$$\begin{aligned} (\mathbf{1} - \tfrac{1}{2}\Delta t D_1) \cdot \tilde{u}^{(1)}(\boldsymbol{x}) &= (\mathbf{1} + \tfrac{1}{2}\Delta t D_1) \cdot u(t, \boldsymbol{x}) \\ (\mathbf{1} - \tfrac{1}{2}\Delta t D_2) \cdot \tilde{u}^{(2)}(\boldsymbol{x}) &= (\mathbf{1} + \tfrac{1}{2}\Delta t D_2) \cdot \tilde{u}^{(1)}(\boldsymbol{x}) \\ & \vdots \quad \vdots \quad \vdots \\ (\mathbf{1} - \tfrac{1}{2}\Delta t D_n) \cdot \tilde{u}^{(n)}(\boldsymbol{x}) &= (\mathbf{1} + \tfrac{1}{2}\Delta t D_n) \cdot \tilde{u}^{(n-1)}(\boldsymbol{x}) \end{aligned}$$

and set:

$$u(t-\Delta t, \boldsymbol{x}) := \tilde{u}^{(n)}(\boldsymbol{x}).$$

For commuting D_i and D_j, i.e. $D_i D_j = D_j D_i$, this scheme is, like the one-dimensional Crank–Nicolson method, of convergence order $\mathcal{O}(\Delta t^2)$:

$$\begin{aligned}\tilde{u}^{(j)}(\boldsymbol{x}) &= (\mathbf{1} - \tfrac{1}{2}\Delta t D_i)^{-1} \cdot (\mathbf{1} + \tfrac{1}{2}\Delta t D_i) \cdot \tilde{u}^{(j-1)}(\boldsymbol{x}) \\ &= (\mathbf{1} + \tfrac{1}{2}\Delta t D_i + \tfrac{1}{4}\Delta t^2 D_i^2) \cdot (\mathbf{1} + \tfrac{1}{2}\Delta t D_i) \cdot \tilde{u}^{(j-1)}(\boldsymbol{x}) + \mathcal{O}(\Delta t^3) \\ &= (\mathbf{1} + \Delta t D_i + \tfrac{1}{2}\Delta t^2 D_i^2) \cdot \tilde{u}^{(j-1)}(\boldsymbol{x}) + \mathcal{O}(\Delta t^3) \end{aligned} \tag{53}$$

$\Longrightarrow$

$$\begin{aligned} u(t-\Delta t) &= \left[\prod_i \left(\mathbf{1} + \Delta t D_i + \frac{1}{2}\Delta t^2 D_i^2\right)\right] \cdot u(t) + \mathcal{O}(\Delta t^3) \\ &= \left[\mathbf{1} + \Delta t \sum_i D_i + \frac{1}{2}\Delta t^2 \sum_{i,j} D_i D_j\right] \cdot u(t) + \mathcal{O}(\Delta t^3) \end{aligned} \tag{54}$$

Equation (54) is of precisely the same form as the one we obtain for u in t from the continuous equation $(\partial_t + L) \cdot u = 0$:

$$u(t-\Delta t) = \left[\mathbf{1} + \Delta t \sum_i L_i + \frac{1}{2}\Delta t^2 \sum_{i,j} L_i L_j\right] \cdot u(t) + \mathcal{O}(\Delta t^3) \tag{55}$$

In order to avoid a building up of lower order error terms due to the fact that D_i and D_j don't always commute perfectly (primarily due to the boundary conditions, but also due to round-off), the ordering of the scheme can be permuted. For a three-factor model, this means there are $3! = 6$ permutations that we can cycle through, as shown in the Appendix.

Appendix

Code schematic for alternating permutation operator split method for three-dimensional diffusions with zero correlation

```
//
// Schematic sample code for the control block and main loop of a three-dimensional operator split
// Crank-Nicolson method.
//
// This code does not contain examples for the implementation of the actual Crank-Nicolson steps
// that need to be carried out for each of the three components, nor the incorporation of the
// lateral boundary conditions.
//

const unsigned long n1 = 200, n2 = 30, n3 = 30; // Sample values for the number of spatial levels
                                                // in each direction.

//
```

```
// There are 6 possible permutations of a sequence of three elements. We therefore adjust the number
// of steps to be a multiple of 6. When product related event dates are to be considered, this ought
// to be done for each time interval.
//
const unsigned long numberOfSteps = 200, adjustedNumberOfSteps = ((numberOfSteps+5)/6)*6;

//
// Each scheme consists of three steps. The set of all possible schemes is given by all possible
// permutations. We sort them such that the last step of any one scheme is different from the first
// step of the next scheme in the sequence.
//
const unsigned long schemes[6][3] = {
    { 0, 1, 2 },  // D1, D2, D3
    { 0, 2, 1 },  // D1, D3, D2
    { 2, 0, 1 },  // D3, D1, D2
    { 2, 1, 0 },  // D3, D2, D1
    { 1, 2, 0 },  // D2, D3, D1
    { 1, 0, 2 },  // D2, D1, D3
};

//
// The class ThreeDimensionalContainer is a user-written container for the solution values at the
// grid nodes. Keep it simple and fast.
//
ThreeDimensionalContainer terminalBoundaryCondition(n1,n2,n3), workspace;

//
// Here, the terminal boundary conditions should be evaluated to populate the known lattice values
// at the final point in time which is the starting point for the backwards induction algorithm.
// The evaluation of the terminal boundary conditions will normally involve the layout of the grid
// in all three coordinates taking into account potential discontinuities of the terminal boundary
// condition (effectively the initial values) or its derivative (you should always have a grid
// level at the strike of plain vanilla options), the precomputation of any coefficient
// combinations that will be constant for each spatial node through time, etc.
//
ThreeDimensionalContainer * threeDimensionalContainers[2] = { &terminalBoundaryCondition, &workspace };
ThreeDimensionalContainer * knownValues = &terminalBoundaryCondition, * unknownValues;
unsigned long i, j, k, schemeindex=5, stepInSchemeIndex, containerIndicator=0;

////////////////////////////////////////////////////////////////////////////////
//                                                                            //
// The main loop of backward induction.                                       //
//                                                                            //
for (i=0;i<adjustedNumberOfSteps;++i){                                        //
  ++schemeindex  %= 6;                                                        //
  for (stepInSchemeIndex=0;stepInSchemeIndex<3;++stepInSchemeIndex){          //
    ++containerIndicator  %= 2;                                              //
    unknownValues = threeDimensionalContainers[containerIndicator];           //
    switch (schemes[schemeindex][stepInSchemeIndex]){                         //
      case 0 :          //                                                    //
                        // Crank-Nicolson step in D1 to be placed here.       //
                        //                                                    //
        break;                                                                //
      case 1 :          //                                                    //
                        // Crank-Nicolson step in D2 to be placed here.       //
                        //                                                    //
        break;                                                                //
      case 2 :          //                                                    //
                        // Crank-Nicolson step in D3 to be placed here.       //
                        //                                                    //
        break;                                                                //
    }                                                                         //
    knownValues = threeDimensionalContainers[containerIndicator];             //
  }                                                                           //
}                                                                             //
////////////////////////////////////////////////////////////////////////////////
//
```

```
// Assuming that the grid levels are stored in the three one-dimensional vectors x1Values[],
// x2Values[], and x3Values[], and that the spot coordinates are given by x1, x2, and x3,
// and that we have already asserted that (x1,x2,x3) is inside the grid, we interpolate the
// solution at (x1,x2,x3) from the grid values.
//


//
// Find the right coordinate indices i, j, and k.
//
for (i=0;x1Values[i]<x1;++i);
for (j=0;x2Values[j]<x2;++j);
for (k=0;x3Values[k]<x3;++k);


//
// Compute weights.
//
const double p1 = (x1-x1Values[i-1])/(x1Values[i]-x1Values[i-1]), q1 = 1 - p1;
const double p2 = (x2-x2Values[j-1])/(x2Values[j]-x2Values[j-1]), q2 = 1 - p2;
const double p3 = (x3-x3Values[k-1])/(x3Values[k]-x3Values[k-1]), q3 = 1 - p3;


//
// Below, we assume that an object v of class ThreeDimensionalContainer allows you to retrieve
// the value at the (i,j,k) grid coordinates by the use of the notation v(i,j,k).
//
const ThreeDimensionalContainer &v = *knownValues;

//
// Trilinear interpolation (consistent with the original operator discretisation).
//
const double solution  =   p1*p2*p3*v(i,j,k)     + p2*q1*p3*v(i-1,j,k)
                         + p1*q2*p3*v(i,j-1,k)   + q1*q2*p3*v(i-1,j-1,k)
                         + p1*p2*q3*v(i,j,k-1)   + p2*q1*q3*v(i-1,j,k-1)
                         + p1*q2*q3*v(i,j-1,k-1) + q1*q2*q3*v(i-1,j-1,k-1);
```

REFERENCES

■ Baaquie, B. E. (1997) A path integral approach to option pricing with stochastic volatility: some exact results. *Journal de Physique*, 1(7): 1733–1753.

■ Chesney, M. and Scott, L. (1989) Pricing European currency options: a comparison of the modified Black–Scholes model and a random variance model. *Journal of Financial and Quantitative Analysis*, 24: 267–284, September 1989.

■ Crank, J. and Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proceedings of the Cambridge Philosophical Society*, 43: 50–67.

■ Doss, H. (1977) Liens entre équations différentielles stochastiques ordinaires. *Annales de l'Institut Henrie Poincaré. Probabilités et Statistiques*, 13: 99–125.

■ Douglas J. and Rachford, H. H. (1956) On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82: 421–439.

■ Feller W. (1951) Two singular diffusion problems. *Annals of Mathematics*, 54: 173–182.

■ Fouque, J.-P., Papanicolaou, G. and Sircar, K. R. *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press, September 2000. ISBN 0521791634.

■ Geske, R. (1977) The valuation of corporate liabilities as compound options. *Journal of Financial and Quantitative Analysis*, 12: 541–552.

■ Geske, R. and Johnson,H. E.H. E. (1984) The valuation of corporate liabilities as compound options: a correction. *Journal of Financial and Quantitative Analysis*, 19: 231–232.

■ Golub, G. H. and Van Loan, C. F. (1983, 1989, 1996) *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD.

■ Hagan, P., Kumar, D. and Lesniewski, A. S. (2002) Managing smile risk. *Wilmott*, 2(1): 84–108.

■ Heston, S. L. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6: 327–343.

■ Hull, J. and White, A. (1987) The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2): 281–300, June. http://faculty.baruch.cuny.edu/lwu/890/HullWhite87.pdf

■ Hull, J. and White, A. (1988) An analysis of the bias in option pricing caused by a stochastic volatility. *Advances in Futures and Options Research*, 3: 27–61.

■ Jäckel, P. *Monte Carlo Methods in Finance*. Wiley, Chichester, February 2002.

■ Karatzas, I. and Shreve, S. E. (1991) *Brownian Motion and Stochastic Calculus*. Springer-Verlag, Berlin.

■ Kloeden, P. E. and Platen, E. (1992, 1995, 1999) *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin.

■ Lumsdaine, A., Lee, L. and Siek, J. The iterative template library. http://www.osl.iu.edu/research/itl/

■ Marchuk, G. I. (1989) Splitting and alternating direction methods. In Lions, J. and Ciarlet, P. (eds), *Handbook of Numerical Analysis*, vol. I, pages 197–462. Elsevier Science Publishers.

■ Merton, R. C. (1973) Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4: 141–183, Spring.

■ Merton, R. C. (1976) Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3: 125–144.

■ Merton, R. C. (1990) *Continuous-Time Finance*. Blackwell, Oxford.

■ Morton, K. W. and Mayers, D. F. (1994) *Numerical Solution of Partial Differential Equations*. Cambridge University Press, Cambridge. ISBN 0521429226.

■ Peaceman, D. W. and Rachford, H. H. (1955) The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3: 28–41.

■ Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in C*. Cambridge University Press, Cambridge.

■ Rogers, L. C. G. and Williams, D. *Diffusions, Markov Processes and Martingales: Volume 2, Ito Calculus*. Cambridge University Press, Cambridge, September 2000.

■ Rubinstein, M. (1983) Displaced diffusion option pricing. *Journal of Finance*, 38: 213–217, March 1983.

■ Schöbel, R. and Zhu, J. (1999) Stochastic volatility with an Ornstein Uhlenbeck process: an extension. *European Finance Review*, 3: 23–46.

■ Scott, L. Option pricing when the variance changes randomly: theory, estimation and an application. *Journal of Financial and Quantitative Analysis*, 22: 419–438, December 1987.

■ Stein, E. M. and Stein, J. C. (1991) Stock price distribution with stochastic volatility: an analytic approach. *Review of Financial Studies*, 4: 727–752.

■ Tavella, D. and Randall, C. (2000) *Pricing Financial Instruments: The Finite Difference Method*. Wiley, Chichester, April 2000. ISBN 0471197602.

■ Wilmott, P. (2000) *Quantitative Finance*. Wiley, Chichester.

24
Cliquet Options and Volatility Models

Paul Wilmott

Wilmott magazine, December 2002

Cliquet options are at present the height of fashion in the world of equity derivatives. These contracts, illustrated by the term sheet below, are appealing to the investor because of their protection against downside risk, yet with significant upside potential. Capping the maximum, as in this globally floored, locally capped example, ensures that the payoff is never too extreme and therefore that the value of the contract is not too outrageous.

Five-year Minimum Coupon Cliquet on ABC Index

Option Buyer	XXXX
Option Seller	YYYY
Notional Amount	EUR 25MM
Start Date	dd/mm/yyyy
Maturity Date	Start Date + Five years
Option Seller Pays at Maturity	$\text{Notional}^{*}\max\left\{\sum_{i=1}^{5}\max\left[0,\min\left(\text{Cap},\frac{S_i-S_{i-1}}{S_{i-1}}\right)\right],\text{Floor}\right\}$
Index	ABC Index
Cap	8%
Floor	16%
Option Premium	???
Index Levels	S_i = Closing Level of Index on Start Date + i years

This indicative term sheet is neither an offer to buy or sell securities nor an OTC derivative product which includes options, swaps, forwards and structured notes having similar features to OTC derivative transactions, nor a solicitation to buy or sell securities or an OTC derivative product. The proposal contained in the foregoing is not a complete description of the terms of a particular transaction and is subject to change without limitation.

E-mail: paul@wilmott.com

From the point of view of the sell side, aiming to minimize market risk by delta hedging, their main exposure is to volatility risk. However, the contract is very subtle in its dependence on the assumed model for volatility.

In this brief note, I will show how the contract value depends on the treatment of volatility. In particular, I shall show results for constant volatility and volatility ranges.

The subtle nature of the cliquet option

Traditionally one measures sensitivity to volatility via the vega. This is defined as the derivative of the option value with respect to a (usually constant) volatility. This number is then used to determine how accurate a price might be should volatility change. As part of one's risk management, perhaps one will vega hedge to reduce such sensitivity.

This is entirely reasonable when the contract in question is an exchange-traded vanilla contract and one is measuring sensitivity to the market's (implied) volatility.

However, when it comes to the risk management of exotic options the sensitivity to a constant volatility is at best irrelevant and at worst totally misleading. By now this is common knowledge and I don't need to dwell on the details. It suffices to say that whenever a contract has a Gamma which changes sign (as does any 'interesting' exotic) vega may be small at precisely those places where sensitivity to actual volatility is very large.

Confused? As a rule of thumb if you increase volatility when Gamma is positive you will increase a contract's value. At points of inflection in the option value (where Gamma is zero) the option value may hardly move. But this is sensitivity to a parameter that takes the same value everywhere. What if you increase volatility when Gamma is positive and decrease it when Gamma is negative? The net effect is an increase in option value even at points of inflection.

Skews and smiles can make matters even worse, unless you are fortunate enough that your skew/smile model forecasts *actual* volatility behavior accurately.

The classical references to this phenomenon are Avellaneda, Levy and Parás (1995) and Lyons (1995) but also see Wilmott (2000).

And the relevance to cliquet options? To see this you just need to plot the formula:

$$\max\left[0, \min\left(\text{Cap}, \frac{S_i - S_{i-1}}{S_{i-1}}\right)\right]$$

against S to see the non convex nature of the option price; Gamma changes sign.

Now comes the subtle part. *The point at which Gamma changes sign depends on the relative move in S from one fixing to the next*. The point of inflection is not near any particular value of S. The conclusion has to be that any deterministic, volatility surface model fitted to vanilla prices is not going to be able to model the risk associated with changing volatility. This is true even if you allow the local volatility surface to move up and down *and* to rotate.

For this reason we are going to focus on using the uncertain volatility model described in the above-mentioned references. In this model the actual volatility is chosen to vary with the variables in such a way as to give the option value its worst (or best) possible value. The actual volatility is assumed to lie in the range σ^- to σ^+. The worst option value is when actual

volatility is highest for negative Gamma and lowest for positive Gamma:

$$\sigma(\Gamma) = \begin{cases} \sigma^+ & \text{if } \Gamma < 0 \\ \sigma^- & \text{if } \Gamma > 0. \end{cases}$$

Now let us look at the pricing of the cliquet option.

Path dependency, constant volatility

We will be working in the classical lognormal framework for the underlying:

$$dS = \mu S\, dt + \sigma S\, dX$$

Assuming for the moment that volatility is constant, or at most a deterministic function of stock price S and time t, we can approach the pricing from the two most common directions, Monte Carlo simulation and partial differential equations. A brief glance at the term sheet shows that there are none of the nasties such as early exercise, convertibility or other decision processes that make Monte Carlo difficult to implement.

Monte Carlo

Monte Carlo pricing requires a simulation of the risk-neutral random walk for S, the calculation of the payoff for many, tens of thousands, say, of paths, and the present valuing of the resulting average payoff. This can be speeded up by many of the now common techniques. Calculation of the greeks is slightly more time consuming but still straightforward.

PDE

To derive a partial differential equation which one then solves via, for example, finite-difference methods, requires one to work out the amount of path dependency in the option and to count the number of dimensions. This is not difficult, see Wilmott (2000).

In all non-trivial problems we always have the two given dimensions, S and t. In order to be able to keep track, before expiry, of the progress of the possible option payoff we also need the following two new 'state variables':

$$S' \quad \text{and} \quad Q.$$

where:

$$S' = \text{the value of } S \text{ at the previous fixing} = S_i$$

and:

$$Q = \text{the sum to date of the bit inside the max function}$$

$$= \sum_{j=1}^{i} \max\left[0, \min\left(\text{Cap}, \frac{S_j - S_{j-1}}{S_{j-1}}\right)\right]$$

Here I am using the index i to denote the fixing just prior to the current time, t. This is all made clear in the figure.

Since S' and Q are only updated discretely, at each fixing date, the pricing problem for $V(S, t, S', Q)$ becomes:

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

where r is the risk-free interest rate. In other words, the vanilla Black–Scholes equation. The twist is that V is a function of four variables, and must further satisfy the jump condition across the fixing date:

$$V(S, t_i^-, S\prime, Q) = V\left\{S, t_i^+, S, Q + \max\left[0, \min\left(E_1, \frac{S - S'}{S'}\right)\right]\right\}$$

and the final condition:

$$V(S, T, S', Q) = \max(Q, E_2).$$

Here E_1 is the local cap and E_2 the global floor. (More general payoff structures can readily be imagined.)

Being a four-dimensional problem, it is a toss up as to whether a Monte Carlo or a finite-difference solution is going to be the faster. However, the structure of the payoff, and the assumption of lognormality, mean that a similarity reduction is possible, taking the problem down to only three dimensions and thus comfortably within the domain of usefulness of finite-difference methods. The similarity variable is:

$$\xi = \frac{S}{S'}.$$

The option value is now a function of ξ, t and Q. The governing equation for $V(\xi, t, Q)$ (loose notation, but the most clear) is:

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\sigma^2 \xi^2 \frac{\partial^2 V}{\partial \xi^2} + r\xi\frac{\partial V}{\partial \xi} - rV = 0.$$

The jump condition becomes:

$$V(\xi, t_i^-, Q) = V(1, t_i^+, Q + \max(0, \min(E_1, \xi - 1)))$$

and the final condition is:

$$V(\xi, T, Q) = \max(Q, E_2).$$

All of the results that I present are based on the finite-difference solution of the partial differential equation. The reason for this is that I want to focus on the volatility dependence, in particular I need to be able to implement the uncertain volatility model described above and this is not so simple to do in the Monte Carlo framework (the reason being that volatility

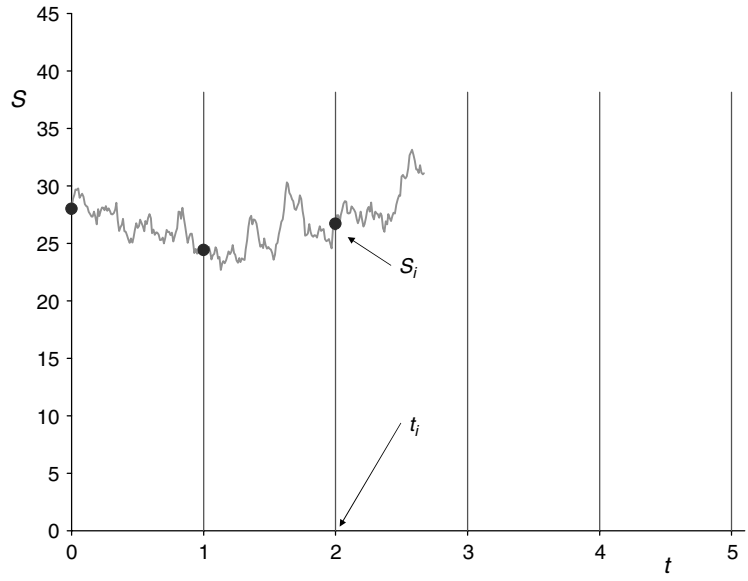


Figure 1: Stock price path and data points used in calculating the cliquet payoff

depends on Gamma in this model and Gamma is not calculated in the standard Monte Carlo implementation).

Results

The following results are based on the cliquet option described in the term sheet. In particular, it is a five-year contract with annual fixings, a global floor of 16% and local caps of 8%. The interest rate is 3% and there are no dividends on the underlying.

To understand the following you must remember that the cliquet value is a function of three independent variables, ξ, Q and t. I will be showing plots of value against various variables at certain times before expiry. These will assume a constant volatility. Then we will look at the effect of varying volatility on the prices.

Constant volatility

In the following five plots volatility is everywhere 25%. Figure 2 shows the cliquet value against Q and ξ at 4.5 years before expiry. The contract has thus been in existence for six months. At this stage there have been no fixings yet and the state variable Q only takes the value 0. The non-convex contract value can be clearly seen.

At 3.5 years before expiry, and therefore 1.5 years into the contract's life, the value is shown in Figure 3. The state variable Q now ranges from zero to 8%.

One year later (see Figure 4) the contract is exactly half way through its life. The state variable Q lies in the range zero to 16%. For small values of Q the option value is very close to being the present value of the 16% floor. This represents the small probability of getting a payoff in excess of the floor at expiry.

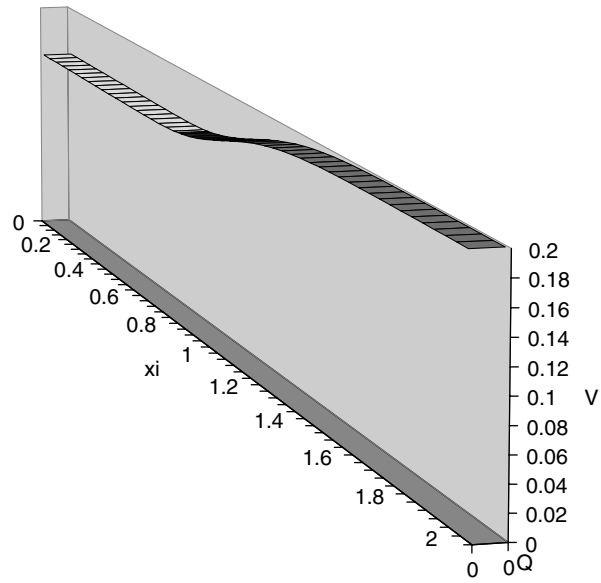


Figure 2: 4.5 years before expiration

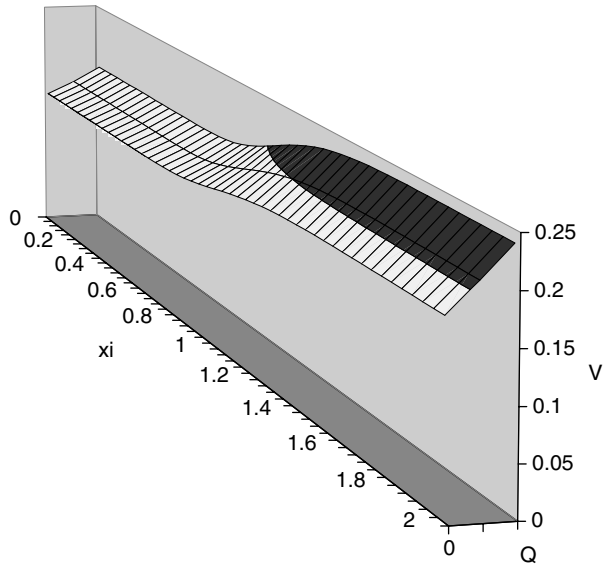


Figure 3: 3.5 years before expiration

After being in existence for 3.5 years, and having only 1.5 years left to run, the cliquet value is as shown in Figure 5. Now Q ranges from zero to 24%. When Q is zero there is no chance of the global floor being exceeded and so the contract value there is exactly the present value of 16%.

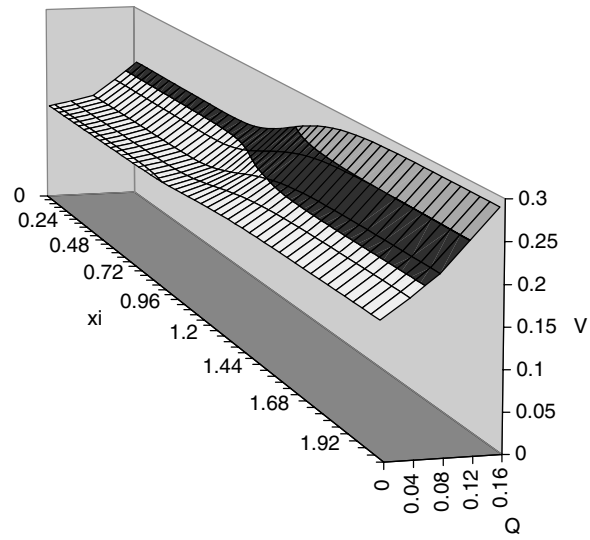


Figure 4: 2.5 years before expiration

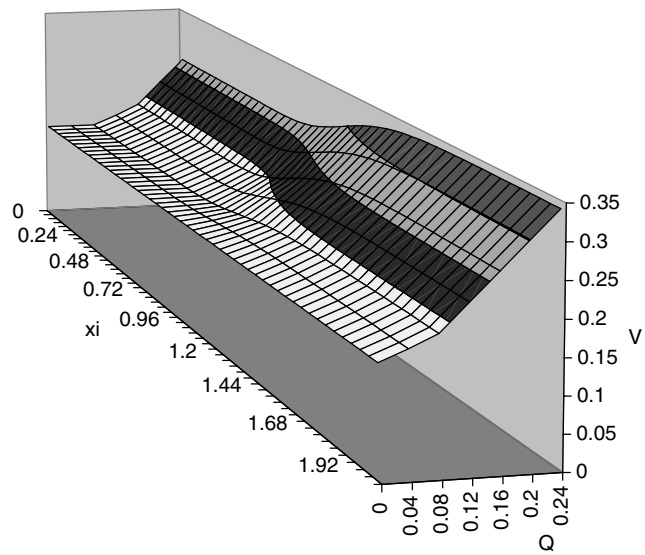


Figure 5: 1.5 years before expiration

Six months before expiry the option value is as shown in Figure 6. Q ranges from zero to 32% and for any values below 8% the contract is again only worth the present value of 16%.

Uncertain volatility

The above shows the evolution of the option value for constant volatility. There is diffusion in the ξ direction and a 'jump condition' to be applied at every fixing. The amount of the diffusion is constant. (Or rather, is constant on a logarithmic scale.)

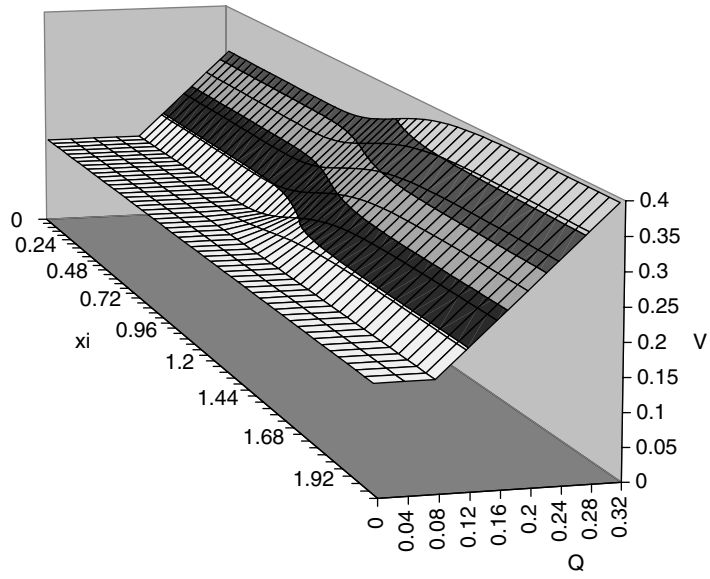


Figure 6: 0.5 years before expiration

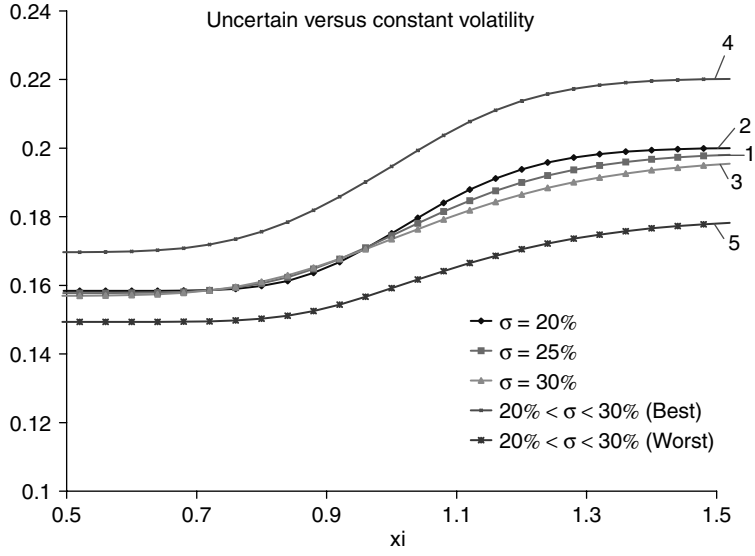


Figure 7: Cliquet values for different volatility models

To price the contract when volatility is uncertain we must use a volatility that depends on (the sign of) Gamma. Some results are shown below.

Figure 7 plots the contract value against ξ at five years before expiry with $Q = 0$. Five calculations have been performed.

1. The first line to examine is the middle line in Figure 7. This corresponds to a constant volatility of 25%. This is the base case with which we compare other prices.
2. The second line to examine is close to the middle line. This is the cliquet value with a constant volatility of just 20%.
3. The third case has a constant volatility of 30%.
4. The fourth line represents the cliquet value when the volatility is allowed to range between 20 and 30%, taking a value locally that maximizes the cliquet value overall.
5. The fifth and final curve is the one for which volatility has again been allowed to range from 20 to 30%, but now such that it gives the option its lowest possible value.

The first observation to make is how close the constant volatility curves are, i.e. curves 1–3. As stated above, a good rule of thumb is that high volatility and positive Gamma give a high option value. Because Gamma changes sign in this contract a result of this is that there is a ξ value at which the contract value does not appear to be sensitive to the volatility. In this case the value is around 0.95, close to the point of inflection.

Now ask yourself the following question. "Do I believe that volatility is a constant, and this constant is somewhere between 20% and 30%? Or do I believe that volatility is highly uncertain, but is most likely to stay within the range 20% to 30%?"

If you believe the former, then the calculation we have just done, in curves 1–3, is relevant. If, on the other hand, you think that the latter is more likely (and who wouldn't?) then you must discard the calculations in curves 1–3 and consider the whole spectrum of possible option values by looking at the best and worst cases, curves 4 and 5.

Such calculations show that the real sensitivity to volatility is much, much larger than a naive vega calculation would suggest.

Table 1 shows how the cliquet value (five years before expiry at $Q = 0$ and $\xi = 1$) varies with the allowed range for volatility. The table is to be read as follows. When volatility takes one value only, read along the diagonal, the dark tinted cells, to see the contract values. For example, when the volatility is 22% the contract value is 0.1739. And when the volatility is 27% the contract value is 0.1726.

TABLE 1: RANGE OF VOLATILITY AND RESULTING RANGE OF CLIQUET PRICES

		$V_{\max}$										
		0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.3
$V_{\min}$	0.2	0.1743	0.1720	0.1700	0.1680	0.1662	0.1645	0.1629	0.1615	0.1601	0.1588	0.1576
	0.21	0.1763	0.1741	0.1719	0.1699	0.1680	0.1663	0.1646	0.1631	0.1617	0.1603	0.1591
	0.22	0.1784	0.1761	0.1739	0.1718	0.1698	0.1680	0.1663	0.1647	0.1632	0.1618	0.1605
	0.23	0.1804	0.1780	0.1757	0.1736	0.1716	0.1698	0.1680	0.1663	0.1648	0.1633	0.1620
	0.24	0.1824	0.1799	0.1776	0.1754	0.1734	0.1715	0.1696	0.1679	0.1663	0.1648	0.1634
	0.25	0.1843	0.1818	0.1794	0.1772	0.1751	0.1731	0.1713	0.1695	0.1679	0.1663	0.1648
	0.26	0.1863	0.1837	0.1812	0.1789	0.1768	0.1748	0.1729	0.1711	0.1694	0.1678	0.1662
	0.27	0.1881	0.1855	0.1830	0.1807	0.1785	0.1764	0.1744	0.1726	0.1708	0.1692	0.1676
	0.28	0.1900	0.1873	0.1847	0.1824	0.1801	0.1780	0.1760	0.1741	0.1723	0.1706	0.1690
	0.29	0.1918	0.1890	0.1865	0.1840	0.1817	0.1796	0.1775	0.1756	0.1738	0.1720	0.1704
	0.3	0.1935	0.1908	0.1881	0.1857	0.1833	0.1811	0.1790	0.1770	0.1752	0.1734	0.1717

Now consider a range of possible volatilities. Suppose you believe volatility will not stray from the range 22% to 27%. The worst case is in the light tinted cells, in this case 0.1647. The best case is to be found in the untinted cells, 0.1830. So, when volatility ranges from 22 to 27% the correct range for the contract value is 0.1647 to 0.1830.

When volatility is a constant, but a constant between 22% and 27%, the contract value range is $0.1739-0.1726 = 0.0013$ or 0.75% relative (to mid-price) range. When volatility is allowed to varying over the 22–27% range we find that the contract value itself has a value range of $0.1830-0.1647 = 0.0183$ or 10.5% relative (to mid-price) range. The true sensitivity to volatility is 14 times greater than that estimated by vega.

Code sample: Cliquet with uncertain volatility, in similarity variables

Below is some Visual Basic code that can be used for pricing these cliquet options in the uncertain volatility framework.

The range for volatility is `VolMin` to `VolMax`, the dividend yield is `Div`, risk-free interest rate `IntRate`, the local cap is `Strike2` and the global floor `Strike1`. Expiry is `Expiry`. The numerical parameter is `NumAssetSteps`, the number of steps in the S and Q directions.

This program clearly leaves much to be desired, for example in the discretization, the treatment of the jump condition, etc. But it does have the benefit of transparency.

A. The timestep is set so that the explicit finite difference method is stable. If the timestep is any smaller than this the method will not converge.
B. Here the payoff is set up, the dependent variable as a function of the independent variables.
C. The timestepping engine. Delta and Gamma are discretized versions of the first- and second-order derivatives with respect to S. This part of the code also treats the uncertain volatility. See how the volatility depends on the sign of Gamma.
D. The boundary conditions, for $\xi = 0$ and large ξ.
E. Updating the next step back in the grid.
F. Here the code tests for a fixing date.
G. Across fixing dates the updating rule is applied. This is really the only point in the code that knows we are pricing a cliquet option.

```
Option Explicit
Function cliquet (VolMin, VolMax, Div, IntRate, Strike1, Strike2, NumFixes, _
   Fixing, Expiry, NumAssetSteps)
ReDim xi (-NumAssetSteps To NumAssetSteps)
Dim Vmax, AssetStep, TStep, QStep, Delta, Gamma, Theta, Tim, Vol, qafter, frac, _
   V1, V2 As Double
Dim i, j, k, M, iafter, kafter, N, NumSoFar, NumQSteps As Integer
ReDim jtest (1 To NumFixes) As Integer

Vmax = Application.Max(VolMin, VolMax)
AssetStep = 1 / NumAssetSteps
TStep = 0.95 * AssetStep ^ 2 / Vmax ^ 2 / 2 ^ 2 ' This ensures stability of the _
   explicit method
```

```
M = Int(Expiry / TStep) + 1
TStep = Expiry / M
QStep = AssetStep
NumQSteps = Int (Strike2 / QStep) * NumFixes

ReDim Q (0 To NumQSteps)
ReDim Vold (-NumAssetSteps To NumAssetSteps, 0 To NumQSteps) ' First dimension _
   centred on xi = 1
ReDim VNew (-NumAssetSteps To NumAssetSteps, 0 To NumQSteps)

NumSoFar = 1
For j = 1 To NumFixes - 1
jtest(j) = Int (j * Fixing / TStep) ' Used in testing whether fixing date has _
   been passed
Next j

For k = 0 To NumQSteps
Q(k) = k * QStep
For i = -NumAssetSteps To NumAssetSteps
xi(i) = 1 + AssetStep * i ' i = 0 corresponds to xi = 1
Vold(i, k) = Application.Max(Strike1, Q(k) + _
                  Application.Max(0, Application.Min(Strike2, xi(i) - 1))) ' _
                     Payoff
Next i
Next k

For j = 1 To M

For k = 0 To NumQSteps
For i = -NumAssetSteps + 1 To NumAssetSteps - 1
Delta = (Vold(i + 1, k) - Vold(i - 1, k)) / 2 / AssetStep ' Central difference
Gamma = (Vold(i + 1, k) - 2 * Vold(i, k) + Vold(i - 1, k)) / AssetStep _
   / AssetStep
Vol = VolMax
If Gamma > 0 Then Vol = VolMin ' Volatility depends on Gamma in the uncertain _
   volatility model
Theta = IntRate * Vold(i, k) - 0.5 * Vol * Vol * xi(i) * xi(i) * Gamma _
                                    -  (IntRate - Div) * xi(i) * Delta ' The _
                                       Black-Scholes equation
VNew(i, k) = Vold(i, k) - TStep * Theta
Next i

VNew(-NumAssetSteps, k) = Vold(-NumAssetSteps, k) * (1 - IntRate * TStep) ' _
   Boundary condition at xi = 0
VNew(NumAssetSteps, k) = VNew(NumAssetSteps - 1, k) ' Boundary condition at xi = _
   infinity. Delta = 0

For i = -NumAssetSteps To NumAssetSteps
Vold(i, k) = VNew(i, k)
Next i

Next k

If jtest (NumSoFar) = j Then ' Test for a fixing date

    For i =  -NumAssetSteps To NumAssetSteps
```

A

B

C

D

E

F

```
    For k =  0 To NumQSteps
    qafter =  Q(k) + Application.Max(0, Application.Min(Strike2, xi(i) - 1)) ' _
       The updating rule
    kafter =  Int(qafter / QStep)
    frac =  (qafter - QStep * kafter) / QStep

    V1 = 0
    V2 = 0
    If kafter < NumQSteps Then
    V1 = VNew(0, kafter)
    V2 = VNew(0, kafter + 1)
    End If

    Vold(i, k) = (1 - frac) * V1 + frac * V2 ' The jump condition. Linear _
       interpolation
    Next k
    Next i
    NumSoFar = NumSoFar + 1

End If

Next j

cliquet = Vold ' Output the whole array

End Function
```

G

REFERENCES

- Avellaneda, M. Levy, A. & Parás, A. (1995) Pricing and hedging derivative securities in markets with uncertain volatilities. *Applied Mathematical Finance* 2, 73–88.
- Avellaneda, M. & Parás, A. (1996) Managing the volatility risk of derivative securities: the Lagrangian volatility model. *Applied Mathematical Finance* 3, 21–53.
- Lyons, T. J. (1995) Uncertain volatility and the risk-free synthesis of derivatives. *Applied Mathematical Finance* 2, 117–133.
- Wilmott, P (2000) *Paul Wilmott on Quantitative Finance*. Wiley, Chichester.

25
Long Memory and Regime Shifts in Asset Volatility

Jonathan Kinlay

Wilmott magazine, January 2003

The conditional distribution of asset volatility has been the subject of extensive empirical research in the last decade. The overwhelming preponderance of evidence points to the existence of pronounced long-term dependence in volatility, characterized by slow decay rates in autocorrelations and significant correlations at long lags (e.g. Crato and de Lima, 1993, and Ding, Granger and Engle, 1993). Andersen *et al*, 1999, find similar patterns for autocorrelations in the realized volatility processes for the Dow 30 stocks – autocorrelations remain systematically above the conventional Bartlett 95% confidence band as far out as 120 days. Comparable results are seen when autocorrelations are examined for daily log range volatility, as Figure 1 illustrates. Here we see significant autocorrelations in some stocks as far back as two years.

Long memory detection and estimation

Among the first to consider the possibility of persistent statistical dependence in financial time series was Mandelbrot (1971), who focused on asset returns. Subsequent empirical studies, for example by Greene and Fielitz (1977), Fama and French (1988), Porteba and Summers (1988) and Jegadeesh (1990), appeared to lend support for his findings of anomalous behavior in long-horizon stock returns. Tests for long-range dependence were initially developed by Mandelbrot using a refined version of a test statistic, the Rescaled Range, initially developed by English hydrologist Harold Hurst (1951). The classical rescaled range statistic is

Contact address: Caissa Capital Partners, New York, USA.
E-mail: jkinlay@caissacapital.com Telephone: 212 786 1514 Fax: 212 786 1516 www.caissacapital.com

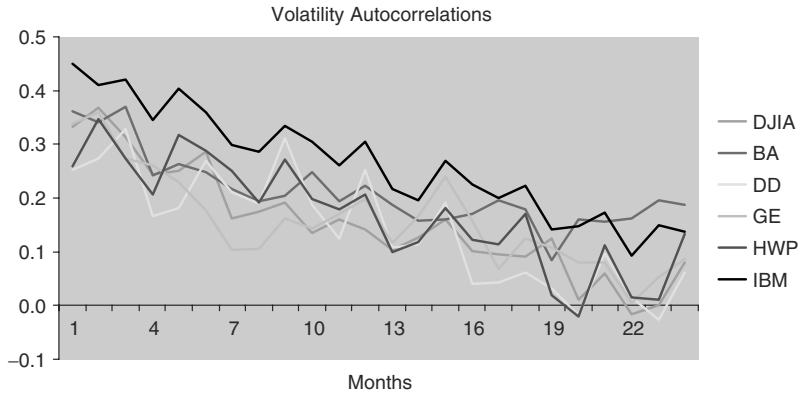


Figure 1

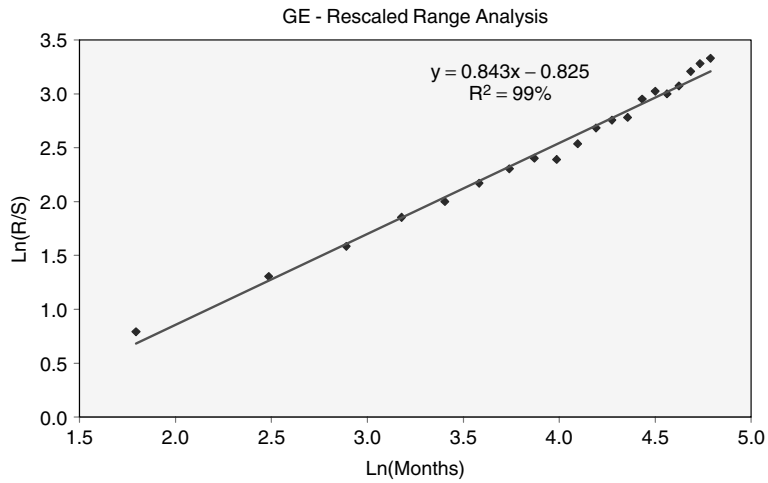


Figure 2: Estimation of Hurst exponent for GE volatility process

defined as:

$$R/S(n) = \frac{1}{s_n}\left[\begin{array}{c}\text{Max}\sum_{j=1}^{k}(X_j - \tilde{X}_n) - \text{Min}\sum_{j=1}^{k}(X_j - \tilde{X}_n) \\ 1 \le k \le n\end{array}\right]$$

where s_n is the sample standard deviation:

$$s_n = \left[\frac{1}{n}\sum_j (X_j - \tilde{X}_n)^2\right]^{1/2}$$

The first term is the maximum of the partial sums of the first k deviations of X_j from the sample mean. Since the sum of all n deviations of the X_j's from their mean is zero, this term

is always non-negative. Correspondingly, the second term is always nonpositive and hence the difference between the two terms, known as the range for obvious reasons, is always nonnegative.

Mandelbrot and Wallis (1969) use the R/S statistic to detect long-range dependence in the following way. For a random process there is scaling relationship between the rescaled range and the number of observations n of the form:

$$R/S(n) \sim n^H$$

where H is known as the Hurst exponent. For a white noise process $H = 0.5$, whereas for a persistent, long memory process $H > 0$. The difference $d = (H - 0.5)$ represents the degree of fractional integration in the process.

Mandelbrot and Wallis suggest estimating the Hurst coefficient by plotting the logarithm of $R/S(n)$ against $\log(n)$. For large n, the slope of such a plot should provide an estimate of H. The researchers demonstrate the robustness of the test by showing by Monte Carlo simulation that the R/S statistic can detect long-range dependence in highly non-Gaussian processes with large skewness and kurtosis. Mandelbrot (1972) also argues that, unlike spectral analysis which detects periodic cycles, R/S analysis is capable of detecting nonperiodic cycles with periods equal to or greater than the sample period.

The technique is illustrated below for the volatility process of General Electric Corporation, a DOW Industrial Index component. The estimated Hurst exponent given by the slop of the regression, approximately 0.8, indicates the presence of a substantial degree of long-run persistence in the volatility process. Analysis of the volatility processes of other DOW components yield comparable Hurst exponent estimates in the region of 0.76–0.96.

A major shortcoming of the rescaled range is its sensitivity to short-range dependence. Any departure from the predicted behavior of the R/S statistic under the null hypothesis need not be the result of long-range dependence, but may merely be a symptom of short-term memory. Lo (1991) shows that this results from the limiting distribution of the rescaled range:

$$\frac{1}{\sqrt{n}} R/S(n) \Rightarrow V$$

where V is the range of a Brownian bridge on the unit interval.

Suppose now that the underlying process $\{X_j\}$ is short range-dependent, in the form of a stationary $AR(1)$, i.e.:

$$r_t = \rho r_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2), |\rho| \in (0, 1)$$

The limiting distribution of $R/S(n)/\sqrt{n}$ is $V[(1 + \rho)/(1 - \rho)]^{1/2}$. As Lo points out, for some common stocks the estimated autoregressive coefficient is as large at 0.5, implying that the mean of $R/S(n)/\sqrt{n}$ may be biased upward by as much as 73%. In empirical tests, Davies and Harte (1987) show that even though the Hurst coefficient of a stationary Gaussian $AR(1)$ is precisely 0.5, the 5% Mandelbrot regression test rejects this null hypothesis 47% of the time for an autoregressive parameter of 0.3

To distinguish between long-range and short-term dependence, Lo proposes a modification of the R/S statistic to ensure that its statistical behavior is invariant over a general class of short memory processes, but deviates for long memory processes. His version of the R/S test

statistic differs only in the denominator. Rather than using the sample standard deviation, Lo's formula applies the standard deviation of the partial sum, which includes not only the sums of squares of deviations for X_j, but also the weighted autocovariances (up to lag q):

$$\hat{\sigma}_n^2(q) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \tilde{X}_n)^2 + 2\sum_{j=1}^{q}\omega_j(q)\hat{\gamma}_j, \omega_j(q) = 1 - \frac{j}{q+1}, q < n$$

where the γ_j are the usual autocovariance estimators.

While in principle this adjustment to the R/S statistic ensures its robustness in the presence of short-term dependency, the problem remains of selecting an appropriate lag order q. Lo and MacKinlay (1989) have shown that when q becomes relatively large to the sample size n, the finite-sample distribution of the estimator can be radically different from its asymptotic limit. On the other hand, q cannot be taken too small as the omitted autocovariances beyond lag q may be substantial. Andrews (1991) provides some guidance on the choice of q, but since criteria are based on asymptotic behavior little is known about the optimal choice of lag in finite samples.

Another method used to measure long-range dependence is the detrended fluctuation analysis (DFA) approach of Peng *et al* (1994) and further developed by Viswanathan *et al* (1997). Its advantage over the rescaled range methodology is that it avoids the spurious detection of apparent long-run correlation due to non-stationarities. In the DFA approach the integrate time series $y(t')$ is obtained:

$$y(t') = \sum_{T=1}^{t'} x(t).$$

The series $y(t')$ is divided into non-overlapping intervals, each containing m data points, and a least squares line is fitted to the data. Next, the root mean square fluctuation of the detrended time series is calculated for all intervals:

$$F(m) = \sqrt{\frac{1}{T}\sum_{t'=1}^{T}[y(t') - y_m(t')]^2}$$

A log–log plot of $F(m)$ vs the interval size m indicates the existence of a power-scaling law. If there is no correlation, or only short-term correlation, then $F(m) \infty m^{1/2}$, but if there is long-term correlation then $F(m)$ will scale at rates greater than $\frac{1}{2}$.

A third approach is a semi-parametric procedure to obtain an estimate of the fractional differencing parameter d. This technique, due to Geweke and Porter-Hudak (1983), is based on the slope of the spectral density around the angular frequency $w = 0$. The spectral regression is defined by:

$$\ln\{I(\omega_\lambda\} = a + b\ln\left\{4\sin^2\frac{\omega_\lambda}{2}\right\} + n_\lambda, \lambda = 1, \ldots, v$$

where $I(W_\lambda)$ is the periodogram of the time series at frequencies $w_\lambda = 2\pi\lambda/T$ with $l = 1, \ldots, (T-1)/2$. T is the number of observations and v is the number of Fourier frequencies

included in the spectral regression. The least squares estimate of the slope of the regression line provides an estimate of d. The error variance is $\pi^2/6$ and allows for the calculation of the t-statistics for the fractional differencing parameter d. An issue with this procedure is the choice of v, which is typically set to $T1/2$, with Sowell (1992) arguing that u should be based on the shortest cycle associated with long-run correlation.

The final method we consider is due to Sowell (1992) and is a procedure for estimating stationary ARFIMA models of the form:

$$\Phi(L)(1-L)^d(y_t-\mu)=\Theta(L)\varepsilon_t$$

where Φ and Θ are lag polynomials, d is the fractional differencing parameter, μ is the mean of the process $y_t \sim N(\mu, \Sigma)$ and ε_t is an error process with zero mean and constant variance σ_e^2. We can use any set of exogenous regressors to explain the mean: $z=\mathbf{y}-\mu$, $\mu=f(\mathbf{X},\beta)$.

The spectral density function is written in terms of the model parameter d, from which Sowell derives the autocovariance function at lag k in the form:

$$\gamma(k)=\frac{1}{2\pi}\int_0^{2\pi} f(W)e^{twk}\,dw$$

The parameters of the model are then estimated by exact maximum likelihood, with log likelihood:

$$\log L(d,\phi,\theta,\beta,\sigma_e^2)=-\frac{T}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}$$

Structural breaks

Granger and Hyung (1999) take a different approach to the analysis of long term serial autocorrelation effects. Their starting point is the standard $I(d)$ representation of an fractionally integrated process y_t of the form:

$$(1-L)^d y_t=\varepsilon_t$$

where d is the fractional integration parameter and, from its Maclaurin expansion:

$$(1-L)^d=\sum_{j=0}^{\infty}\pi_j L^j,\ \pi_j=\frac{j-1-d}{j}\pi_{j-1},\ \pi_0=1$$

The researchers examine the evidence for structural change in the series of absolute returns for the SP500 Index by applying the sequential break point estimation methodology of Bai (1997) and Bai and Perron (1998) and Iterative Cumulative Sums of Squares (ICSS) technique of Aggarwal, Inclan and Leal, 1999. Bai's procedure works as follows. When the break point is found at period k, the whole sample is divided into two subsamples with the first subsample consisting of k observations and the second containing the remaining $(T-k)$ observations. A break point is then estimated for the subsample where a hypothesis test of parameter consistency is rejected. The corresponding subsample is then divided into further subsamples at the

estimated break point and a parameter constancy test performed for the hierarchical subsamples. The procedure is repeated until the parameter constancy test is not rejected for all subsamples. The number of break points is equal to the number of subsamples minus 1. Bai shows how the sequential procedure coupled with hypothesis testing can yield a consistent estimate for the true number of breaks.

Aggarwal, Inclan and Leal's (1999) approach uses the Iterative Cumulative Sums of Squares (ICSS) as follows. We let $\{\varepsilon_t\}$ denote a series of independent observations from a normal distribution with zero mean and unconditional variance σ_t^2. The variance within each interval is denoted by τ_j^2, $j = 0, 1, \ldots, N_t$, where N_t is the total number of variance changes in T observations and $1 < k_1 < k_2 < \cdots < k_{NT} < T$ are the set of change points.

$$\text{So} \quad \sigma_t = \tau_j k_j < t < k_{j+1}$$

To estimate the number of changes in variance and the point in time of the shift a cumulative sum of squares is used.

Let $C_\lambda = \sum_{t=1}^{k} \varepsilon_t^2$, $k = 1, \ldots, T$ be the cumulative sum of the squared observations from the start of the series until the k^{th} point in time. Then define $D_k = (C_k/C_T) - k/T$.

If there are no changes in variance over the sample period, the D_k oscillate around zero. Critical values based on the distribution of D_k under the null hypothesis of no change in variance provide upper and lower bounds to detect a significant change in variance with a known level of probability. Specifically, if $\max_k \sqrt{(T/2)}|D_k|$ exceeds 1:36, the 95$^{\text{th}}$ percentile of the asymptotic distribution, then we take k^*, the value of k at which the maximum value is attained as an estimate of the change point.

Figure 3 illustrates the procedure for a simulated GBM process with initial volatility of 20%, which changes to 30% after 190 periods, and then reverts to 20% once again in period 350. The test statistic $\sqrt{(T/2)}|D_k|$ reaches local maxima at $t = 189(2.313)$ and $t = 349(1.155)$, clearly and accurately identifying the two break points in the series.

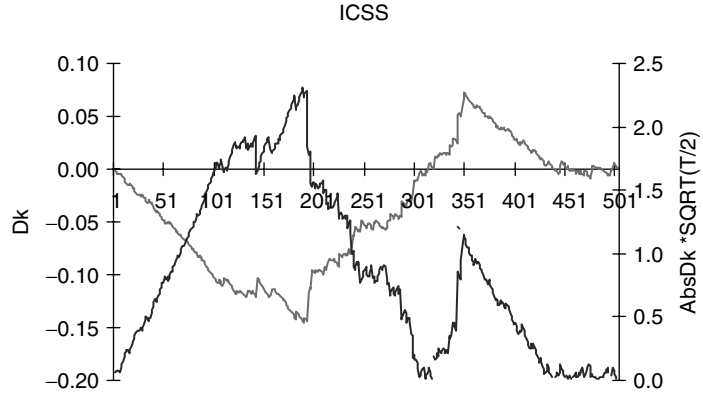


Figure 3: Testing for structural breaks in simulated GBM process using iterative cumulative sums of squares

A similar analysis (Figure 4) is carried out for the series of weekly returns in the SP500 index from April 1985 to April 2002. Several structural shifts in the volatility process are

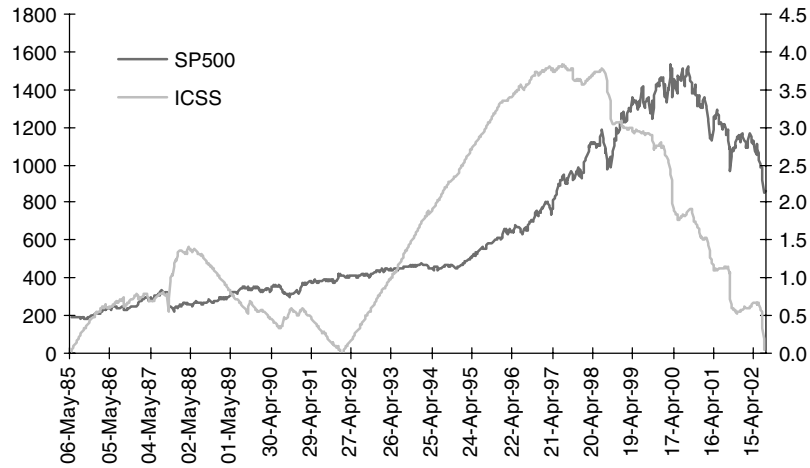


Figure 4: Testing for structural breaks in SP500 index returns using iterative cumulative sums of squares

apparent, including the week of 19 Oct 1987, 20 July 1990 (GulfWar), the market tops around Aug 1997, Aug 1998 and Oct 2000.

In their comprehensive analysis of several emerging and developed markets, Aggarwal *et al* identify numerous structural shifts relating to market crashes, currency crises, hyperinflation and government intervention, including, to take one example, as many as seven significant volatility shifts in Argentina over the period from 1985–1995.

It is common for structural breaks to result in ill-conditioning in the volatility processes distribution, often in the form of excess kurtosis. This kind of problem can sometimes be resolved by modeling the different regime segments individually. Less commonly, regime shifts can produce spurious long memory effects. For example, Granger and Hyung (1999) estimate the degree of fractional integration d in daily SP500 returns for 10 subperiods from 1928–1991 using the standard Geweke and Porter Hudak approach. All of the subperiods have strong evidence of long memory in the absolute stock return. They find clear evidence of a positive relationship between the time-varying property of d and the number of breaks, and conclude that the SP500 Index absolute returns series is more likely to show the "long memory" property because of the presence of a number of structural breaks in the series rather than being an $I(d)$ process.

Stocks in Asian-Pacific markets typically exhibit volatility regime shifts at around the time of the regional financial crisis in the latter half of 1997. The case of the ASX200 Index component stock AMC is typical (see Figure 5). Rescaled range analysis of the entire volatility process history leads to estimates of fractional integration of the order of 0.2. But there is no evidence of volatility persistence is the series post-1997. The conclusion is that, in this case, apparent long memory effects are probably the result of a fundamental shift in the volatility process.

Conclusion

Long memory effects that are consistently found to be present in the volatility processes in financial assets of all classes may be the result of structural breaks in the processes themselves, rather than signifying long-term volatility persistence.

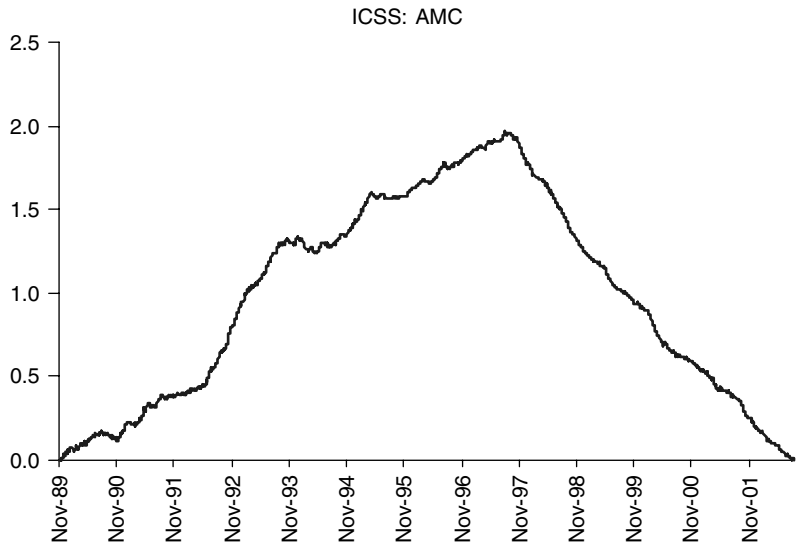


Figure 5: Structural breaks in the Asian crisis period for ASX component stock AMC

Reliable techniques for detecting regime shifts are now available and these can be used to segment the data in a way that reduces the risk of model misspecification.

However, one would be mistaken to conclude that all long memory effects must be the result of regime shifts of one kind or another. Many US stocks, for example, show compelling evidence for volatility persistence both pre- and post-regime shifts. Finally, long memory effects can also result from the interaction of a small number of short-term correlated factors.

REFERENCES

■ Aggarwal, R., Inclan, C. and Leal, R. (1999) Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis* (forthcoming).

■ Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (1999) The distribution of exchange rate volatility. Wharton Financial Institutions Center Working Paper 99-08 and NBER Working Paper 6961.

■ Bai, J. (1997) Estimating multiple breaks one at a time. *Econometric Theory* 13, 315–352.

■ Bai, J. and Perron, P. (1998) Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.

■ Breidt, J. F., Crato, N. and de Lima, P. (1998) The detection and estimation of long memory in stochastic volatility. *Journal of Economics* 83, 325–348.

■ Davies, R. and Harte, D. (1987) Tests for Hurst effect. *Biometrika* 74, 95–101.

■ Ding, Z., Granger, C. W. J. and Engle, R. E. (1993) Long memory properties of stock market returns and a new model. *Journal of Empirical Finance* 1, 83–106.

■ Fama, E. and French, K. (1988) Permanent and temporary components of stock prices. *Journal of Political Economy* 96, 246–273.

■ Geweke, J. and Porter-Hudak, S. (1983) The estimation and application of long memory time series models. *Journal of Time Series Analysis* 4, 15–39.
■ Granger, C. W. J. and Hyung, N. (1999) Occasional structural breaks and long memory. University of California, Discussion Paper.
■ Greene, M. and Fielitz, B. (1997) Long-term dependence in common stock returns. *Journal of Financial Economics* 4, 339–349.
■ Jegadeesh, N. (1990) Evidence of predictable behavior of security returns. *Journal of Finance* 45, 881–898.
■ Lo, A. (1991) Long-term memory in stock market prices. *Econometrica* 59, 1279–1313.
■ Lo, A. and MacKinlay (1988) Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.
■ Mandelbrot, B. (1971) When can price be arbitraged efficiently? A limit to the validity of the random walk and Martingale models. *The Review of Economics & Statistics* 53, 225–236.
■ Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. and Goldberger, A. L. (1994) *Phys. Rev. E* 49, 1684.
■ Porteba, J. and Summers, L. (1988) Mean reversion in stock returns evidence and implications. *Journal of Financial Economics* 22, 27–60.
■ Sowell, F. (1989) *Maximum likelihood estimation of fractionally integrated time series models*. Mimeo, Carnegie-Mellon University.
■ Sowell, F. (1992) Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics* 53, 165–188.
■ Teyssiere, G. (1996) Double long-memory financial time series. Paper No 348, Department of Economics, QMW, London.
■ Viswanathan, G. M., Buldyrev, S. V., Havlin, S. and Stanley, H. E. (1997) *Biophys. J.* 72, 866.

26

Heston's Stochastic Volatility Model: Implementation, Calibration and Some Extensions

Sergei Mikhailov and Ulrich Nögel

Wilmott magazine, July 2003

The paper discusses theoretical properties, shows the performance and presents some extensions of Heston's (1993) stochastic volatility model. The model proposed by Heston extends the Black and Scholes (1973) model and includes it as a special case. Heston's setting takes into account non-lognormal distribution of the assets returns, leverage effect, important mean-reverting property of volatility and it remains analytically tractable. The Black–Scholes volatility surfaces generated by Heston's model look like empirical implied volatility surfaces. The complication is related to the risk-neutral valuation concept. It is not possible to build a riskless portfolio if we formulate the statement that the volatility of the asset varies stochastically. This is principally because the volatility is not a tradable security.

Heston's stochastic volatility model

In this section we specify Heston's stochastic volatility model and provide some details of how to compute options prices. We use the following notation:

Contact address: Fraunhofer ITWM, Financial Engineering, Gottlieb-Daimler-Strasse 49, 67663 Kaiserslautern, Germany.
E-mails: Mikhailov@itwm.fhg.de and noegel@itwm.fhg.de

$S(t)$	Equity spot price, financial index.
$V(t)$	Variance.
C	European call option price.
K	Strike price.
$W_{1,2}$	Standard Brownian movements.
r	Interest rate.
q	Dividend yield.
κ	Mean reversion rate.
θ	Long run variance.
V_0	Initial variance.
σ	Volatility of variance.
ρ	Correlation parameter.
t_0	Current date.
T	Maturity date.

Heston's stochastic volatility model (1993) is specified as follows:

$$\frac{dS(t)}{S(t)} = \mu\, dt + \sqrt{V(t)}\, dW_1 \tag{1}$$

$$dV(t) = \kappa(\theta - V(t))\, dt + \sigma\sqrt{V(t)}\, dW_2 \tag{2}$$

To take into account leverage effect, Wiener stochastic processes W_1, W_2 should be correlated $dW_1 \cdot dW_2 = \rho\, dt$. The stochastic model (2) for the variance is related to the square-root process of Feller (1951) and Cox, Ingersoll and Ross (1985). For the square-root process (2) the variance is always positive and if $2\kappa\theta > \sigma^2$ then it cannot reach zero. Note that the deterministic part of process (2) is asymptotically stable if $\kappa > 0$. Clearly, that equilibrium point is $V_t = \theta$.

Applying the Ito lemma and standard arbitrage arguments we arrive at Garman's partial differential equation:

$$\frac{\partial C}{\partial t} + \frac{S^2 V}{2}\frac{\partial^2 C}{\partial S^2} + (r-q)S\frac{\partial C}{\partial S} - (r-q)C + [\kappa(\theta - V) - \lambda V]\frac{\partial C}{\partial V} + \frac{\sigma^2 V}{2}\frac{\partial^2 C}{\partial V^2} + \rho\sigma SV\frac{\partial^2 C}{\partial S \partial V} = 0 \tag{3}$$

where λ is the market price of volatility risk.

Heston builds the solution of the partial differential equation (3) not in the direct way but using the method of characteristic functions. He is looking for the solution in the form of the corresponding Black and Scholes model:

$$C(S_0, K, V_0, t, T) = SP_1 - Ke^{-(r-q)(T-t)}P_2 \tag{4}$$

where P_1 is the delta of the European call option and P_2 is the conditional risk neutral probability that the asset price will be greater than K at the maturity. Both probabilities P_1, P_2 also satisfy

PDE (3). Provided that characteristic functions φ_1, φ_2 are known the terms P_1, P_2 are defined via the inverse Fourier transformation:

$$P_j = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\left[\frac{e^{-iu\ln K}\varphi_j(S_0, V_0, t, T, u)}{iu}\right]du, \quad j = 1, 2 \tag{5}$$

Heston assumes the characteristic functions φ_1, φ_2 having the form:

$$\varphi_j(S_0, V_0, \tau; \phi) = \exp\{C_j(\tau;\phi) + D_j(\tau;\phi)V_0 + i\phi S_0\}, \quad \tau = T - t \tag{6}$$

After substitution of φ_1, φ_2 in the Garman equation (3) we get the following ordinary differential equations for unknown functions $C_j(\tau;\phi)$ and $D_j(\tau;\phi)$:

$$\frac{dC_j(\tau;\phi)}{d\tau} - \kappa\theta D_j(\tau;\phi) - (r-q)\phi i = 0 \tag{7}$$

$$\frac{dD_j(\tau;\phi)}{d\tau} - \frac{\sigma^2 D_j^2(\tau;\phi)}{2} + (b_j - \rho\sigma\phi i)D_j(\tau;\phi) - u_j\phi i + \frac{\phi^2}{2} = 0 \tag{8}$$

with zero initial conditions:

$$C_j(0,\phi) = D_j(0,\phi) = 0 \tag{9}$$

The solution of the system (7) (9) is given by:

$$\begin{aligned} C(\tau,\phi) &= (r-q)\phi i\tau + \frac{\kappa\theta}{\sigma^2}\left\{(b_j - \rho\sigma\phi i + d)\tau - 2\ln\left[\frac{1-ge^{d\tau}}{1-g}\right]\right\} \\ D(\tau;\phi) &= \frac{b_j - \rho\sigma\phi i + d}{\sigma^2}\left[\frac{1-e^{d\tau}}{1-ge^{d\tau}}\right] \end{aligned} \tag{10}$$

where:

$$\begin{aligned} &g = \frac{b_j - \rho\sigma\phi i + d}{b_j - \rho\sigma\phi i - d}, \quad d = \sqrt{(\rho\sigma\phi i - b_j)^2 - \sigma^2(2u_j\phi i - \phi^2)}. \\ &u_1 = 0.5, \quad u_2 = -0.5, \quad a = \kappa\theta, \quad b_1 = \kappa + \lambda - \rho\sigma, \\ &b_2 = \kappa + \lambda. \end{aligned} \tag{11}$$

Realization of Heston's stochastic volatility model

How to use the model

Implementing such a model consists of different parts that can be divided under a lot of people:

- The first thing is to implement the closed-form solutions for a standard call for the Heston model and the Heston model with jump diffusion, trying to optimize the numerics for speed, such that the calibration can be done as fast as possible.
- The closed-form solution should be verified with a Monte-Carlo (MC) simulation and by directly solving the resulting PDEs using the Finite Difference Method (FDM).

- With the closed-form solutions a suitable set-up should be established to calibrate the models to traded standard calls.
- With the now calibrated model we finally should be able to calculate the price and the greeks of volatility sensitive products such as cliquets using again Monte-Carlo simulation and the Finite Difference Method.

Everything should be done in C++ and be usable as a DLL in Microsoft Excel.

Implementing the Fourier integral

Inverse Fourier transformation (5) is the main point in numerical implementation of the option valuation algorithm provided that characteristic function is known.

The complex numbers can be conveniently implemented by using the complex <> class from the C++ Standard Library. Because the integral should be computed with a high precision for a wide range of parameters (parameters of the stochastic vol process, different strikes and maturities) we decided to use an adaptive quadrature for the first try. Then the algorithm can adjust to changes in the integrand on its own, saving us from the need to do so. We use an adaptive Simpson and an adaptive Gauss–Lobatto quadrature which both give good results, where the Gauss–Lobatto one uses less computation time for the same precision. But after some experience with the model, we ended up with a special optimized fixed stepwidth Gauss quadrature for faster computation.

The pitfalls of the complex logarithm

Due to the fact that the complex logarithm is multiple valued (see Figure 1):

$$\log z = \log|z| + i(\arg(z) + 2\pi n) \tag{12}$$

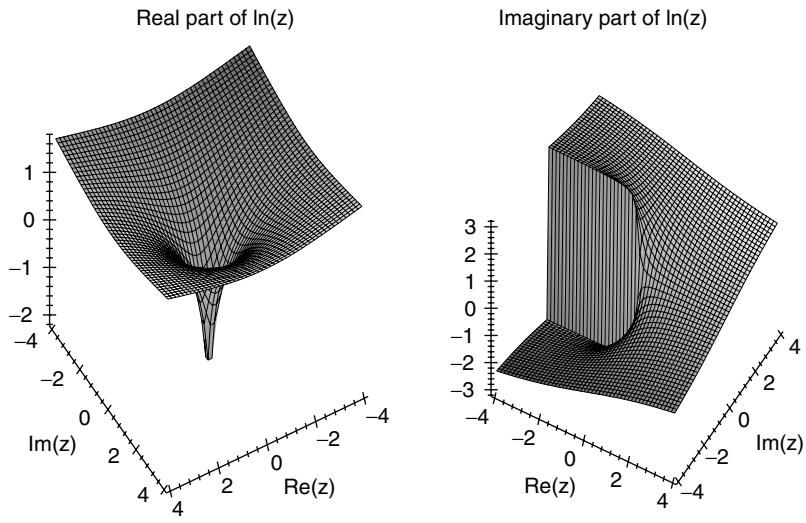


Figure 1: Shows the real part and principle branch of imaginary part of complex logarithm

with n being an integer, one usually restricts the logarithm to its principle branch by restricting $\arg(z) \in [-\pi, \pi]$ and setting $n = 0$. This choice is used by the standard C++ log-function and it is necessarily discontinuous at the cut along the negative real axis. At first we had problems with numerical implementation of complex logarithm. Fortunately, we found help at www.wilmott.com in the thread on stochastic volatility models. After implementing the code with a complex logarithm function that maintains continuous over the cut (thanks Roger for the instruction), the results of our three different numerical approaches (Monte-Carlo simulations, Finite Difference method and closed-form solution) agreed nicely and this gave us the confidence to continue our work.

Calibration of Heston's model to market data

With the now stable implementation of the closed-form solution we are able to calibrate the models to some traded plain vanilla calls.

Calibration scheme

We decide to do a least squared error fit in the following way.

Let $\tau_1, \tau_2, \ldots, \tau_M$ be some times to maturities with $fwd_1, fwd_2, \ldots, fwd_M$ being the corresponding forwards and $dfs_1, dfs_2, \ldots, dfs_M$ the corresponding discount factors. Let $X_1, X_2, \ldots, X_N$ be a set of strikes and σ_{ij}^{imp} the corresponding market implied volatility. The aim of the calibration is to minimize the least squared error:

$$SqErr(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{M} w_{ij} [C_{MP}(X_i, \tau_j) - C_{SV}(S(t), X_i, fwd_j, dfs_j, \tau_j, \theta)]^{\alpha} + Penalty(\Theta, \Theta_0) \tag{13}$$

where $C_{MP}(X_i, \tau_j)$ denotes the market price for a call with strike X_i and maturity τ_j. C_{SV} is the price calculated with the stochastic volatility model which depends on the vector of model parameters $\Theta = (\kappa, \theta, \sigma, \rho, V_0, \lambda)$ for the Heston model. Further, typically $\alpha = 2n, n = 1, 2, \ldots$. The penalty function may be e.g. the distance to the initial parameter vector $Penalty(\Theta, \Theta_0) = \|\Theta - \Theta_0\|^2$ and may be used to give the calibration some additional stability.

As it turns out, the suitable choice of the weight factors w_{ij} is crucial for good calibration results.

Local vs. global optimization

Minimizing the objective function (13) is clearly a nonlinear programming (NLP) problem with the nonlinear constrain $2\kappa\theta - \sigma^2 > 0$. This condition ensures that the volatility process cannot reach zero. Unfortunately the objective function is far from being convex and it turned out that usually there exist many local extrema. As a consequence we decided to try both local and global optimizers:

- *Local (deterministic) algorithms.* Within these types of algorithms one has to choose an initial guess (hopefully a good one) for the parameter vector $\Theta_0 \in R^d$. The algorithm then determines the optimal direction and the stepsize and is moving downhill on the

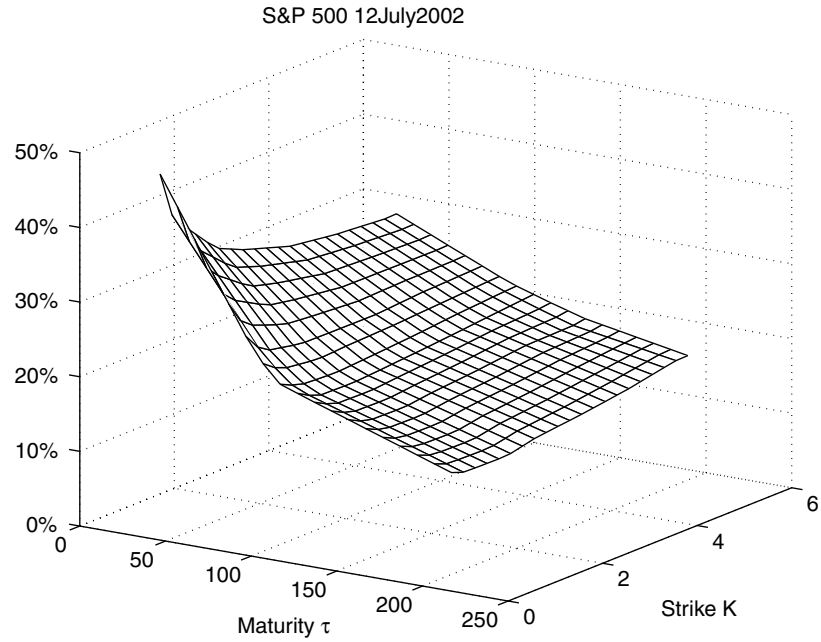


Figure 2: Volatility surface for the S&P 500 index

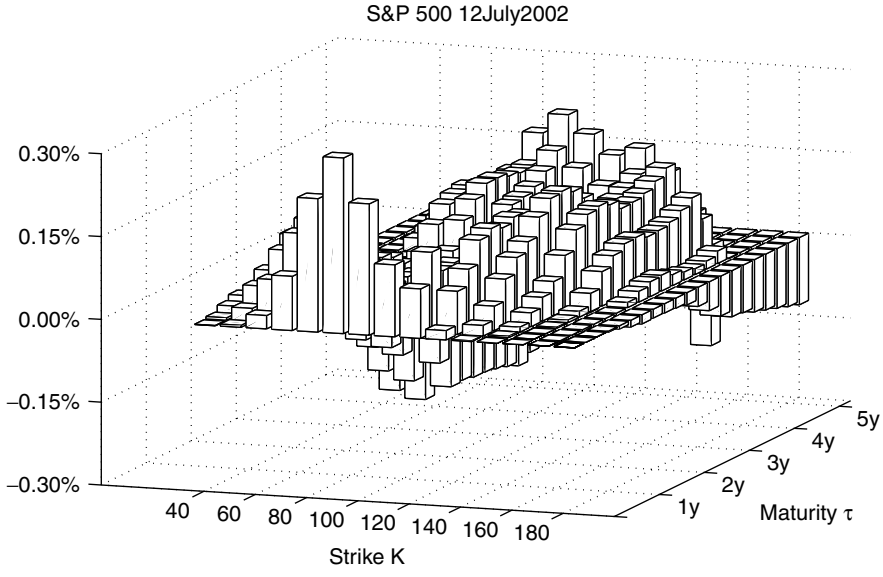


Figure 3: Errors after calibration of the Heston model to the S&P 500 index

parameter manifold to the minimum of the objective function. There are a lot of algorithms available both for unconstrained and constrained problems and they are usually based on simplex or some kind of gradient method. Most of these algorithms work reasonably fast, but one always has the risk to end up in a local minimum. As a consequence a good initial guess is crucial.

- *Stochastic algorithms*. In contrast to the local optimizers the initial guess is (hopefully) irrelevant in the concept of stochastic optimization. The simulated annealing algorithm chooses the direction and stepsize randomly, it "searches everywhere". It moves always downhill but may accept an uphill move with a certain probability p_T which depends on the annealing parameter T. This parameter is called "the temperature" for historical reasons. During the optimization process the temperature is gradually reduced. There exist some convergence theorems, which state that the algorithm always ends up in the global minimum if the annealing process is sufficiently slow. There are different variants (e.g. FA, VFSRA, ASA) available which differ from the original simulated annealing (SA) in the annealing scheme, but in general these stochastic algorithms are computationally more burdensome than the local optimizers.

Results

We tested different local optimizers and surprisingly the built-in Excel solver, which comes with Excel for free, turned out to be very robust and reliable. It is based on the Generalized Reduced Gradient (GRG) method (www.solver.com for details) and is our favored optimizer when we have some "good" initial guess for our parameter vector, e.g. if one has to recalibrate the model every day and the volatility surface has not changed much. We were able to calibrate the Heston model to the S&P 500 index with an maximum error of less than 0.15% for ATM calls (Figures 2 and 3). The Excel solver may however sometimes end up in a local minimum instead of reaching the global minimum. In such cases or when there is no good initial guess available, we use the adaptive simulated annealing (ASA) algorithm (www.ingberg.com), which allows a faster annealing scheme than the standard SA. It further turned out that adding jump-diffusion to the Heston model often does not improve the quality of the calibration any more. This may be due to the fact that the market now frequently shows an inverted yield curve and the model is simply overtaxed with this situation.

Stochastic volatility model with time-dependent parameters

Why are more complex stochastic models required? The answer is simple – because the prices from stochastic engines are not supported by market prices. As a result financial engineers have to recalibrate model parameters every day to new market data. It is not consistent with an accurate description of the dynamics. The next (but not the last) step in the stochastic volatility models history is to models with time-dependent parameters. Since the Riccati differential equation (8) is non-linear, the generalization of Heston model for variable parameters is not straightforward.

Analytical solutions to the Riccati equation

We rewrite the Riccati equation (8) in the standard form:

$$\frac{dx(t)}{dt} + a(t)x^2(t) + b(t)x(t) + c(t) = 0. \tag{14}$$

Recall, that the general solution of a Riccati equation (14) cannot be expressed by means of quadratures except in some particular cases.

The simplest case is $a(t) \equiv 0$. In this case we have a linear differential equation with variable parameters that has an analytical solution.

After change of variable $y(t) = -1/x(t)$ we arrive again at Riccati equation:

$$\frac{dy(t)}{dt} + c(t)y^2(t) + b(t)y(t) + a(t) = 0. \tag{15}$$

Therefore if $c(t) \equiv 0$ in the original Riccati equation then after transformation we obtain again the linear equation with analytical solution.

The general solution of the Riccati equation can be written by means of two quadratures if one particular solution of a Riccati equation is known.

For the Heston stochastic volatility model the ordinary extension for the time-dependent coefficients is long run variance θ. Since this parameter does not appear in the Riccati equation (8) the analytical solution for arbitrary $\theta(t)$ can be constructed. For the other Heston models coefficients κ, ρ, σ the generalization to the time-dependent model is not so straightforward. Some analytical solutions are possible. For example if $\kappa(t) = at + b$, or $\kappa(t) = ae^{-\alpha t}$. In this case the Riccati equation (8) has closed form solutions expressed by means of hypegeometric function. The drawback – numerical implementation of this analytical solution might be more time consuming than direct numerical integration of equations (7) and (8).

Asymptotic solution to Riccati equation

As finding the general solution of the Riccati equation with time-variable coefficient is not possible, the natural approach is to apply asymptotic methods. Let, for simplicity, all Heston model parameters except the correlation coefficient be constant. The approximate solution to the Riccati equation can be found in the form of the asymptotic expansion:

$$\begin{aligned} &\rho(t) = \rho_0 + \varepsilon\rho_1(t) + \varepsilon^2\rho_2(t) + \cdots, \\ &D(t) = D_0(t) + \varepsilon D_1(t) + \varepsilon^2 D_2(t) + \cdots, \quad \varepsilon \ll 1. \end{aligned} \tag{16}$$

In the first approximation we arrive at a linear equation with time-variable coefficients. To obtain the solution of this ODE is straightforward:

$$\begin{aligned} D_1(t) = -\sigma u i \int_0^t \rho_1(\tau) D_0(\tau) \exp\left\{ \int_0^\tau D_0(\xi)\, d\xi - (-\rho_0\sigma u i + b_j)\tau \right\} d\tau \\ \times \exp\left(-\int_0^t D_0(\tau)\, d\tau + (-\rho_0\sigma u i + b_j)t \right) \end{aligned} \tag{17}$$

The alternative to the above-discussed approach is asymptotic analysis of the systems with slow varying parameters.

Analytical solution to Riccati with piece-wise constant parameters

The second extension of the standard Heston stochastic volatility model to time-dependent coefficients is the setting with piecewise-constant parameters. We can define the solution of the Riccati equation (8) with piecewise-constant coefficients by means of adjusting of initial conditions.

At first we need a solution of the equations (7) and (8) with arbitrary initial conditions:

$$C_j(0,\phi) = C_j^0,\ D_j(0,\phi) = D_j^0. \tag{18}$$

The solution was build by means of computer-algebra system Maple.

$$C_j(\tau,\phi) = (r-q)\phi\tau + \frac{\kappa\theta}{\sigma^2} \times \left((b_j - \rho\sigma\phi i + d)\tau - 2\ln\left(\frac{1-ge^{\tau d}}{1-g}\right)\right) \tag{19}$$

$$D_j(\tau,\phi) = \frac{b_j - \rho\sigma\phi i + d - (b_j - \rho\sigma\phi i - d)ge^{\tau d}}{(1-ge^{\tau d})\sigma^2} \tag{20}$$

where:

$$g = \frac{b_j - \rho\sigma\phi i + d - D_j^0\sigma^2}{b_j - \rho\sigma\phi i - d - D_j^0\sigma^2},\quad d = \sqrt{(\rho\sigma\phi i - b_j)^2 - \sigma^2(2u_j\phi i - \phi^2)}. \tag{21}$$

The solution is close to the Heston one (10), (11). The time interval to maturity $[t, T]$ is divided into n subintervals $[t, t_1], \ldots, [t_i, t_j], \ldots, [t_{n-1}, T]$ where $t_k, k = 1, \ldots, n-1$ is the time of model parameters jumps. Model parameters are constant during $[t_i, t_j]$ but different for each subinterval. Further on it is convenient to use the inverse time $\tau = T - t$. The initial condition for the first subinterval from the end $[0, \tau_1]$ where $\tau_k = T - t_{n-k}, k = 1, \ldots n-1$ is zero. Therefore we can use Heston's solution (10), (11). For the second subinterval $[\tau_1, \tau_2]$ we employ the general solution (19)–(21) with arbitrary initial conditions (18). Provided that functions $C_j(\tau,\phi)$, $D_j(\tau,\phi)$ are continuous in the time of parameters jump τ_1 the initial conditions for the second subinterval can be found from the following condition:

$$C_j(0,\phi) = C_j^0 = C_j^H(\tau_1,\phi),\ D_j(0,\phi) = D_j^0 = D_j^H(\tau_1,\phi) \tag{22}$$

where $C_j^H(\tau_1,\phi)$, $D_j^H(\tau_1,\phi)$ are Heston's solutions with zero initial conditions, Solving the above equations relative to C_j^0, D_j^0 we obtain the initial values for the second time interval. The same procedure is repeated at each time moment $\tau_k, k = 2, \ldots, n-1$ of the parameters jumps.

Thus the calculation of the option price for the model with piecewise-constant parameters consists of two phases:

1. Determine the initial conditions for each time interval in accordance with formulas (22).

2. Calculate the functions $C_j(\tau, \phi)$, $D_j(\tau, \phi)$ using the solution (19)–(21) with initial conditions (21).

For the numerical realization this solution is close to the Heston one. Additionally we have to calculate initial conditions for the second time interval.

Numerical verification of the model with time-dependent parameters

Here we compare options prices calculated according to techniques described in the previous section and options prices from a Monte Carlo engine. The algorithm was implemented in C/C++ code. We assume that mean reversion parameter κ is time-dependent and all other model parameters are constant. Opening price S_0 of the underlying asset is 1, the maturity of the option considered is 5 years, interest rate is 0, start value for volatility V_0 is 0.1, the long run variance θ is 0.1, volatility of variance σ is 0.2, correlation coefficient ρ is -0.3, market price of volatility risk λ is 0. The results of the numerical simulations for various strikes K are presented in Table 1.

TABLE 1: COMPARISON OF ANALYTIC SOLUTION WITH MONTE CARLO SIMULATIONS

$k = \{4, 2, 1\}$, $T = 5$					
	Monte-Carlo		Analytical solution		
	$N = 150000,\ n\_b = 150$				
K	Value	StdDev	Value	Abs Err	Rel Err
0.5	0.545298	0.001	0.543017	0.002281	0.004201
0.75	0.387548	0.001048	0.385109	0.002439	0.006333
1	0.275695	0.001021	0.273303	0.002392	0.008752
1.25	0.197629	0.000949	0.195434	0.002195	0.011231
1.5	0.143341	0.00086	0.14121	0.002131	0.015091

Conclusions

The attractive features of the Heston stochastic volatility model are:

- Its volatility updating structure permits analytical solutions to be generated for standard plain vanilla European options and thus the model allows a fast calibration to given market data.
- The form of the Heston stochastic process used to model price dynamics allows for non-lognormal probability distributions.
- Heston stochastic model takes into account the leverage effect.

- The Heston and the HJD model are able to nicely reproduce a wide range of the volatility surfaces implied from option prices in the market.

On the other hand, there remain some disadvantages and open questions:

- The integrals needed for the computation of the option prices do not always converge nicely.
- To perform well across a large time interval of maturities further extensions of the model are necessary (such as time-dependent parameters).
- Heston's model implicitly takes systematic volatility risk into account by means of a linear specification for the volatility risk premium.
- The standard Heston model usually fails to create a short term skew as strong as the one given by the market; the HJD model is often unable to fit an inverse yield curve.

REFERENCES

■ Black, F. and Sholes M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3), 637–659.
■ Cox, J., Ingersoll, J. and Ross, S. (1985) A theory of the term structure of interest rates. *Econometrica* 53: 389–408.
■ Feller, W. (1951) Two singular diffusion problems. *Annals of Mathematics* 54, 173–182.
■ Heston, S. (1993) A closed-form solution for options with stochastic volatility. *Review of Financial Studies* 6, 327–343.
■ Heston, S. and Nandi, S. (1997) A closed form GARCH option pricing model. Federal Reserve Bank of Atlanta Working Paper 97-9, 1–34.

27
Forward-start Options in Stochastic Volatility Models

Vladimir Lucic

Wilmott magazine, September 2003

This article presents a treatment of forward-start options in stochastic volatility models via change of numeraire. By choosing the asset price stopped at the strike setting as the numeraire, the original problem is transformed into valuing a vanilla call option on a hypothetical asset that becomes "alive" only between the strike setting and the expiry of the option. In the particular case of the Heston model, this enables straightforward extension of the closed-form formula of Heston (1993) to forward-start options.

Change of numeraire for forward-start options

Forward-start option is one of the simplest exotics, with the terminal payoff

$$[S_T - KS_{T_0}]^+ \tag{1}$$

where T is the expiry, $T_0 < T$ is the strike set date, and K is the (percentage) strike. Another common version of this contract, which primarily serves as a building block for cliquet options, has the payoff:

$$\left[\frac{S_T}{S_{T_0}} - K\right]^+ \tag{2}$$

Contact address: TD Securities, Triton Court, 14–18 Finsbury Square, London, EC2A 1DB, UK.
E-mail: vladimir.lucic@tdsecurities.com

Valuing these options in the Black–Scholes framework is standard (see Wilmott (1998)). To examine the problem in a more general setting, suppose that dynamics of the asset price and variance process are given on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ under the equivalent martingale measure $\mathbb{Q}$ by:

$$dS_t = r_t S_t\, dt + \sigma_t(v_t, S_t) S_t\, dW_t^{(1)} \tag{3a}$$

$$dv_t = \alpha_t(v_t)\, dt + \beta_t(v_t)(\rho\, dW_t^{(1)} + \sqrt{1-\rho^2}\, dW_t^{(2)}), \quad S_0, v_0 \in \mathbb{R}^+ \tag{3b}$$

where $\alpha_t(\cdot)$, $\beta_t(\cdot) \geq 0$, and $\sigma_t(\cdot) \geq 0$ are deterministic and sufficiently regular to ensure existence and strong uniqueness for (3), as well as the martingale property of the discounted asset price $S_t \exp(-\int_0^t r_s\, ds)$. Within this framework we can study a number of named stochastic volatility models (Heston, Hull–White, Scott, Stein–Stein), together with the local volatility model of Dupire.

Put $P(s,t) := \exp(-\int_s^t r_u I_{\{s \leq u\}}\, du)$, and let $S_t^{T_0}$ be the price process stopped at T_0, $S_t^{T_0} := S_{t \wedge T_0}$. We can write the payoff in (1) as:

$$\left[S_T - K S_T^{T_0}\right]^+$$

so the value of this option at time $t \geq 0$ is:

$$V_t^{(1)} = P(t,T) \mathrm{E}_{\mathbb{Q}}\left[\left[S_T - K S_T^{T_0}\right]^+ \middle| \mathcal{F}_t\right]$$

Fix $t \in [0, T_0]$, and for notational simplicity drop the functional dependence of the parameters. Since:

$$S_u = S_0 \exp\left(\int_0^u \left(r_s - \frac{1}{2}\sigma_s^2\right) ds + \int_0^u \sigma_s\, dW_s^{(1)}\right), \quad u \in [0,T] \tag{4}$$

changing the numeraire to $N_u := S_u^{T_0}/P(T_0,u)$ yields:

$$\begin{aligned} V_t^{(1)} &= N_t \mathrm{E}_{\hat{\mathbb{Q}}}\left[\left[\frac{S_T}{S_T^{T_0}/P(T_0,T)} - KP(T_0,T)\right]^+ \middle| \mathcal{F}_t\right] \\ &= S_t P(T_0,T) \mathrm{E}_{\hat{\mathbb{Q}}}\left[\left[\exp\left(\int_{T_0}^T \left(r_s - \frac{1}{2}\sigma_s^2\right) ds + \int_{T_0}^T \sigma_s\, dW_s^{(1)}\right) - K\right]^+ \middle| \mathcal{F}_t\right] \end{aligned} \tag{5}$$

where $\hat{\mathbb{Q}}$ is the measure corresponding to N_u as the numeraire:

$$\frac{d\hat{\mathbb{Q}}}{d\mathbb{Q}} = \frac{N_T P(0,T)}{N_0} = \exp\left(-\frac{1}{2}\int_0^T I_{\{s \leq T_0\}} \sigma_s^2\, ds + \int_0^T I_{\{s \leq T_0\}} \sigma_s\, dW_s^{(1)}\right)$$

From (3a), (3b), and (5) we have:

$$dS_u = (r_u + \sigma_u^2 I_{\{u \le T_0\}}) S_u \, du + \sigma_u S_u \, d\hat{W}_u^{(1)} \tag{6a}$$

$$dv_u = (\alpha_u + \rho\beta_u\sigma_u I_{\{u \le T_0\}}) \, du + \beta_u(\rho \, d\hat{W}_u^{(1)} + \sqrt{1-\rho^2} \, d\hat{W}_u^{(2)}) \tag{6b}$$

$$V_t^{(1)} = S_t P(T_0, T) \mathrm{E}_{\hat{\mathbb{Q}}} \left[\left[\exp\left(\int_{T_0}^T \left(r_s - \frac{1}{2}\sigma_s^2 \right) ds + \int_{T_0}^T \sigma_s \, d\hat{W}_s^{(1)} \right) - K \right]^+ \middle| \mathcal{F}_t \right] \tag{6c}$$

where, by the Girsanov theorem, $\hat{W}_u^{(1)} := W_u^{(1)} - \int_0^u I_{\{s \le T_0\}} \sigma_s \, ds$ and $\hat{W}_u^{(2)} := W_u^{(2)}$ are independent Wiener processes under $\hat{\mathbb{Q}}$. The expression for $V_t^{(1)}$ in (6c) involves only one Itô exponential, so $V_t^{(1)}/S_t$ can be viewed as the value of a vanilla call option on an asset whose risk-neutral dynamics are:

$$d\hat{S}_u = \hat{r}_u \hat{S}_u \, du + \hat{\sigma}_u \hat{S}_u \, d\hat{W}_u^{(1)}, \quad \hat{S}_0 = 1$$

where $\hat{r}_u := r_u I_{\{T_0 \le u\}}$, $\hat{\sigma}_u := \sigma_u(v_u, S_u) I_{\{T_0 \le u\}}$. As a result of the measure change, the asset $\hat{S}_t$ is "frozen" in time until the strike set date, which is a probabilistic analogue of the similarity reduction commonly used for pricing these options in the Black–Scholes setting (see Wilmott (1998)).

The analogous formula for the payout in (2) is obtained more easily, since in that case the change of numeraire is not needed. Denoting by $V_t^{(2)}$ the value of this option, from (4) we get:

$$V_t^{(2)} = P(t, T) \mathrm{E}_{\mathbb{Q}} \left[\left[\exp\left(\int_{T_0}^T \left(r_s - \frac{1}{2}\sigma_s^2 \right) ds + \int_{T_0}^T \sigma_s \, dW_s^{(1)} \right) - K \right]^+ \middle| \mathcal{F}_t \right] \tag{7}$$

Similarly as before, $V_t^{(2)}/P(t, T_0)$ can be interpreted as the value of the vanilla call option written on an asset that follows:

$$d\hat{S}_u = \hat{r}_u \hat{S}_u \, du + \hat{\sigma}_u \hat{S}_u \, dW_u^{(1)}, \quad \hat{S}_0 = 1$$

where $\hat{r}_u := r_u I_{\{T_0 \le u\}}$, $\hat{\sigma}_u := \sigma_u(v_u, S_u) I_{\{T_0 \le u\}}$.

In this manner we have related pricing forward-start call options to pricing vanilla call options. The obvious drawback is that now the asset has more complicated dynamics, with the local volatility depending exogenously on S_u. This additional complexity, however, is not present in the important class of two-factor models in which $\sigma_u = \sigma_u(v_u)$. This fact will be further explored in the next section when we examine the Heston stochastic volatility model in more detail.

From (6) and (7) it also follows that the only difference between $V_t^{(1)}/S_t$ and $V_t^{(2)}$ (apart from discounting) comes from the Girsanov change of drift in (6a) and (6b). Since β_t and σ_t are nonnegative, the drift in (6b) dominates the drift in (3b) iff $\rho \ge 0$, resulting, by the comparison theorem for diffusions,[1] in a.s. uniformly higher level of the corresponding volatility paths. Therefore, if $\sigma_u = \sigma_u(v_u)$, then $V_t^{(1)} \ge V_t^{(2)} S_t / P(t, T_0)$ iff $\rho \ge 0$.

Forward-start options in the Heston model

The Heston model is under risk-neutral measure given by the following pair of SDEs:

$$dS_t = r_t S_t\,dt + \sqrt{v_t} S_t\,dW_t^{(1)} \tag{8a}$$

$$dv_t = \lambda(\overline{v} - v_t)\,dt + \eta\sqrt{v_t}(\rho\,dW_t^{(1)} + \sqrt{1-\rho^2}\,dW_t^{(2)}) \tag{8b}$$

This is evidently a special case of (3), so the calculations of the previous section apply, provided we can show that $M_t := S_t P(0,t)$ is a true martingale. Establishing this property turns out to be a quite involved technical task, which is therefore relegated to the Appendix.

Examining (6) and (7) we conclude that $V^{(m)}$, $m = 1, 2$ can be obtained via the premiums of vanilla call options written on assets that under $\mathbb{Q}$ follow:

$$d\hat{S}_t^{(m)} = I_{\{T_0 \le t\}} r_t \hat{S}_t^{(m)}\,dt + I_{\{T_0 \le t\}}\sqrt{v_t^{(m)}}\,\hat{S}_t^{(m)}\,dW_t^{(1)} \tag{9a}$$

$$dv_t^{(m)} = \left(\lambda\overline{v} - (\lambda - \rho\eta\ (2-m) I_{\{t \le T_0\}}) v_t^{(m)}\right)\,dt + \eta\sqrt{v_t^{(m)}}(\rho\,dW_t^{(1)} + \sqrt{1-\rho^2}\,dW_t^{(2)}) \tag{9b}$$

The above dynamics differ from (8) only by having piecewise constant (in time) coefficients. As a consequence, the original Heston procedure can still be applied successively over the intervals of parameters constancy, yielding an extension of the Heston closed-form formula. We carry out this procedure in the rest of this section, closely following the derivation of the Heston formula from Gatheral's (2002) notes.

Put $\hat{V}_t^{(m)} := V_t^{(m)}/S_t^{2-m}$ and note that, after ignoring discounting, we get:

$$\hat{V}_t^{(m)} = \mathrm{E}_{\mathbb{Q}}\left[\hat{S}_T^{(m)} I_{\{\hat{S}_T^{(m)} > K\}}\middle|\,\mathcal{F}_t\right] - K\mathrm{E}_{\mathbb{Q}}\left[I_{\{\hat{S}_T^{(m)} > K\}}\middle|\,\mathcal{F}_t\right] \quad m = 1, 2$$

This can be written as:

$$\hat{V}_\tau^{(m)}(x, v) = K\left(e^x P_1^{(m)}(x, v, \tau) - P_0^{(m)}(x, v, \tau)\right) \tag{10}$$

where $x := \ln(\hat{S}_t^{(m)} P(t, T_0)/P(t, T)/K)$, $\tau := T - t$, and $P_0^{(m)}$, $P_1^{(m)}$ stand for probabilities of exercise under the risk-neutral measure and the measure corresponding to $\hat{S}_t^{(m)}$ as numeraire. Since the summands on the right-hand side are values at time t of tradable assets they satisfy the valuation PDE, which in turn implies:

$$-\frac{\partial P_j^{(m)}}{\partial \tau} + \theta_\tau\left(\frac{1}{2}v\frac{\partial^2 P_j^{(m)}}{\partial x^2} + \rho\eta v\frac{\partial^2 P_j^{(m)}}{\partial x\,\partial v} - v u_j\frac{\partial P_j^{(m)}}{\partial x}\right) + \frac{1}{2}\eta^2 v\frac{\partial^2 P_j^{(m)}}{\partial v^2}$$

$$+ (\lambda\overline{v} - b_j^{(m)} v)\frac{\partial P_j^{(m)}}{\partial v} = 0,$$

where:

$$u_j = 1/2 - j, \quad b_j^{(m)} = \lambda + \rho\eta[m - 2 + \theta_\tau(2 - m - j)], \quad \theta_\tau = I_{[0,T-T_0]}(\tau).$$

Switching to Fourier transforms:

$$\tilde{P}_j^{(m)}(k, v, \tau) = \int_{-\infty}^{\infty} \exp(-ikx) P_j^{(m)}(x, v, \tau)\, dx$$

yields:

$$-\frac{\partial \tilde{P}_j^{(m)}}{\partial \tau} + \theta_\tau \left(-\frac{1}{2}k^2 v \tilde{P}_j^{(m)} + ik\rho\eta v \frac{\partial \tilde{P}_j^{(m)}}{\partial v} - ikvu_j \tilde{P}_j^{(m)} \right) + \frac{1}{2}\eta^2 v \frac{\partial^2 \tilde{P}_j^{(m)}}{\partial v^2}$$
$$+ (\lambda\overline{v} - b_j^{(m)} v)\frac{\partial \tilde{P}_j^{(m)}}{\partial v} = 0,$$

so with the Heston's ansatz:

$$\tilde{P}_j^{(m)}(k, v, \tau) = \frac{1}{ik} \exp\{C(m, k, \tau)\overline{v} + D(m, k, \tau)v\}, \tag{11}$$

we get at the following pair of ODEs for C and D:

$$\frac{\partial C}{\partial \tau} = \lambda D, \quad C(0) = 0, \tag{12a}$$

$$\frac{\partial D}{\partial \tau} = \alpha_\tau - \beta_\tau^{(m)} D + \eta^2 D^2/2, \quad D(0) = 0, \tag{12b}$$

where:

$$\alpha_\tau = \theta_\tau(-k^2/2 - ik/2 + ijk),$$
$$\beta_\tau^{(m)} = b_j^{(m)} - i\rho\eta k\theta_\tau = \lambda + \rho\eta[m - 2 + \theta_\tau(2 - m - j - ik)].$$

At this point we have reduced the original problem to integration of (12) over $[0, \tau]$, which is completed considering two separate cases.

First suppose $\tau \in [0, T - T_0]$, which is the case of a vanilla call option. The parameters in (12) are constant over the region of integration, so according to Heston (1993) we have:

$$D(m, k, \tau) = r_-^{(m)} \frac{1 - \exp(-d^{(m)}\tau)}{1 - g^{(m)} \exp(-d^{(m)}\tau)}, \tag{13}$$

$$C(m, k, \tau) = \lambda \left\{ r_-^{(m)}\tau - \frac{2}{\eta^2} \ln\left(\frac{1 - g^{(m)} \exp(-d^{(m)}\tau)}{1 - g^{(m)}} \right) \right\}, \tag{14}$$

where:

$$d^{(m)} = \sqrt{\left(\beta_0^{(m)}\right)^2 - 2\alpha_0\eta^2},$$

$$r_{\pm}^{(m)} = \frac{\beta_0^{(m)} \pm d^{(m)}}{\eta^2},$$

$$g^{(m)} = \frac{r_-^{(m)}}{r_+^{(m)}}.$$

Next suppose $\tau > T - T_0$, which is the case of a genuine forward-start option. Using $C(m, k, T - T_0)$ and $D(m, k, T - T_0)$ as the initial conditions, and integrating over $[T - T_0, \tau]$ yields:

$$D(m, k, \tau) = \frac{2\beta_T^{(m)}}{\eta^2(1 + c\exp(\beta_T^{(m)}(\tau - T + T_0)))},$$

$$C(m, k, \tau) = C(m, k, T - T_0) + \frac{2\beta_T^{(m)}\lambda(\tau - T + T_0)}{\eta^2} - \frac{2\lambda}{\eta^2}\ln\left(\frac{1 + c\exp(\beta_T^{(m)}(\tau - T + T_0))}{1 + c}\right),$$

$$c = \frac{2\beta_T^{(m)}}{\eta^2 D(m, k, T - T_0)} - 1.$$

According to (11) this completes the calculation of the Fourier transform of the option price.

As noted previously, the fact that (8) and (9) have the same form (modulo time-dependence of the parameters) was crucial in extending the original Heston procedure. This property is shared by the Stein–Stein model:

$$d\ln(S_t) = (r_t - 1/2v_t^2)\,dt + v_t\,dW_t^{(1)}$$

$$dv_t = k(\theta - v_t)\,dt + \sigma(\rho\,dW_t^{(1)} + \sqrt{1 - \rho^2}\,dW_t^{(2)})$$

so the same approach can potentially be applied to this model as well. (The general closed-form formula for vanilla options in Stein–Stein model is due to Schöbel and Zhu (1999)).

Future work

In view of the increasing importance of models incorporating jumps in asset and/or volatility dynamics (Duffie *et al.* (2000) Overhaus *et al.* (2002)), generalising the present result to that class of models would be a natural extension of this work. Another potential area of application is the perturbation analysis of the forward skew along the lines of Fouque *et al.* (2000).

Appendix: Asset as numeraire in the Heston model

Ensuring that the exponential local martingale $M_t = S_t P(0, t)$ is a true martingale is instrumental in the application of the Girsanov theorem at the beginning of this chapter. Unlike the

classical Black–Scholes case, this condition need not hold in stochastic volatility models, potentially creating problems with numeraire change (see Sin (1998), Lewis (2000) for examples of such models and related consequences for option pricing). The first comprehensive study of conditions under which this property holds in the Heston model was carried out in Wong and Heyde (2002), who present several sufficient conditions. These authors also study the conditions under which equivalent martingale measure exists in the Heston model, a problem that we do not address in this work.

In the rest of this section we establish that M_t is a martingale if $\lambda \geq 0$, which is the case in virtually all practical applications. This result relies on techniques different from those used in Wong and Heyde (2002), and extends the corresponding result of that paper. The main tool for ensuring that M_t is a martingale is the Novikov-type condition of Exercise VIII.1.40 of Revuz and Yor (1999), requiring existence of two constants $a > 0$, $c > 0$ such that:

$$\mathrm{E}_{\mathbb{Q}}[e^{av_t}] \leq c, \quad \forall t \in [0, T] \tag{15}$$

Starting with the formula from Pitman and Yor (1982) (see also Wong and Heyde (2002)), for an extension) we get for $\Re(p) \geq 0$

$$\mathrm{E}_{\mathbb{Q}}[\exp(-pv_t)] = A(p, t)\exp(-B(p, t)v_0), \tag{16}$$

$$A(p, t) = \left(\frac{2\lambda \exp(\lambda t)}{\eta^2 p(\exp(\lambda t) - 1) + 2\lambda \exp(\lambda t)}\right)^{2\lambda\overline{v}/\eta^2}$$

$$B(p, t) = \frac{2p\lambda}{\eta^2 p(\exp(\lambda t) - 1) + 2\lambda \exp(\lambda t)}$$

Fix $t \in [0, T]$, and let $F(p, t)$ denote the expression on the right-hand side in (16). Since $F(p, t)$ is analytic on $\Re(p) > \frac{2\lambda \exp(\lambda T)}{\eta^2(1-\exp(\lambda T))} =: \alpha_0$ (the singularity occurring for $\lambda = 0$ is removable), from Lemma A.3 of Filipović *et al* (2003) it follows that (16) holds for $p = \alpha_0/2$. Thus, since the function $g(t) := F(\alpha_0/2, t)$ is continuous on $[0, T]$, with $c := \max_{t\in[0,T]} g(t) \in (0, \infty)$, $\alpha := -\alpha_0/2 > 0$ we have

$$\mathrm{E}_{\mathbb{Q}}[e^{\alpha v_t}] \leq c, \quad \forall t \in [0, T]$$

Therefore, the version of the Novikov condition (15) holds, and M_t is a martingale.

FOOTNOTES & REFERENCES

1. e.g. Theorem IX.3.7 of Revuz and Yor (1999).

■ Duffie, D., Pan, J., and Singleton, K. (2000) Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68, 1343–1376.

■ Filipović, D., Duffie, D., and Schachermayer, W. (2003) Affine processes and applications in finance. *Ann. Appl. Probab.* 13(3), 984–1053.

■ Fouque, J-P., Papanicolaou, G., and Sircar, K. R. *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press, Cambridge, 2000.

■ Gatheral, J. *Stochastic Volatility and Local Volatility, Lecture Notes*. Courant Institute of Mathematical Sciences, New York, Fall 2002.
■ Heston, S. L. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Finan. Stud.* 6, 327–343, 1993.
■ Lewis, A. *Option Valuation Under Stochastic Volatility*. Finance Press, Newport Beach, 2000.
■ Overhaus, M., Ferraris, A., Knudsen, T., Mao, F., Milward, R., Nguyen-Ngoc, L., and Schindlmayr, G. *Equity Derivatives. Theory and Applications*. Wiley, Chichester, 2002.
■ Pitman, J. and Yor, M. A decomposition of Bessel bridges. *Zeit Wajrsch. Geb.* 59, 425–457, 1982.
■ Revuz, D. and Yor, M. *Continuous Martingales and Brownian Motion*. Springer-Verlag, Berlin, 1999.
■ Schöbel, R. and Zhu, J. Stochastic volatility with Ornstein–Uhlenbeck process: An extension. *Eur. Finance Rev.* 3, 23–46, 1999.
■ Sin, C. Complications with stochastic volatility models. *Adv. Appl. Prob.* 30, 256–268, 1998.
■ Wilmott, P. *Derivatives. The Theory and Practice of Financial Engineering*. Wiley, Chichester, 1998.
■ Wong, B. and Heyde, C. C. Change of measure in stochastic volatility models. *Preprint, University of New South Wales*, 2002.

28
Stochastic Volatility and Mean-variance Analysis

Hyungsok Ahn* and Paul Wilmott†

Wilmott magazine, November 2003

Stochastic volatility models usually lead to a linear option pricing equation containing a market price of risk term. This term is the source of endless problems and argument.

The main reason for the argument is that this quantity is not directly observable. At best it can be deduced from the prices of derivatives, so called 'fitting.' But this is far from adequate, the fitting will only work if those who set the prices of derivatives are using the same model and they are consistent in that the fitted market price of risk does not change when the model is refitted a few days later.

In practice, refitted parameters are always significantly different from the original fit. This is why practitioners use static hedging, to minimize model error. However, static hedging may be considered to be an afterthought, since it is, in the classical framework, no more than a patch for mending a far-from-perfect model.

Whether we have a deterministic volatility surface or a stochastic volatility model with prescribed or fitted market price of risk, we will always be faced with how to interpret refitting. Was the market wrong before but is now right, or was the market correct initially and now there are arbitrage opportunities? We won't be faced with awkward questions like this if we don't expect our model, whatever it may be, to give unique values. In this paper we'll see how to estimate probabilities for prices being correct. We do this by only delta hedging and not dynamically vega hedging. Instead we look at means and variances for option values.

What's wrong

In the mark-to-market accounting framework, the price of a security should be marked at the prevailing market price. Thus, we do not need a theoretical model to price vanilla products in this framework. A model plays a significant role for determining the price of custom products

Contact addresses: *E-mail: Hyungsok.Ahn@CommerzbankIB.com
† E-mail: paul@wilmott.com

as well as for risk management. A typical approach in practice is to select a suitable model and to calibrate its parameter to match the model price of a vanilla product with the market quote:

$$\text{model::price}(\text{product}(\alpha), \text{quote}(\text{product}(\cdot), t), \text{parameter}, t) = \text{quote}(\text{product}(\alpha), t)$$

whenever a quote is available. Thus, by the implicit function theorem, the model parameter is a function of market prices and time:

$$\text{parameter} = \theta(\text{quote}(\text{product}(\cdot), t), t)$$

This function is *supposed to be invariant* under the change of time and quote, but there is no physical constraint that it *has to be invariant*. A model uses its parameter to describe the random behavior of the market prices. Therefore, if the model parameter changes, the prices before and after the change are not consistent any more. This will generate P&L that is unexplained by the model.

In the stochastic volatility framework, it looks as if the model allows its parameter to change. However, the volatility in this context is not a parameter any more. It is just an index which is assumed observable or estimable. The parameter is one that describes the dynamics of the volatility. The obvious merit of a stochastic volatility model is that it has more parameters to fit the market quotes better (for example, the smile). Nevertheless, its parameter is not immune from changing randomly in time. This is simply because the market itself has a higher order of complexity than that of a stochastic volatility model.

Suppose that a stochastic volatility model has its parameters invariant under the change of time and quote. Does this mean that this model leads us to a risk-free land? Here's an extreme example. The market is pricing all the vanilla options with a flat volatility at 50%. One calibrates the Black–Scholes model perfectly, always. What happens if the realized volatility of the underlying price won't agree with 50%? Thus, a perfect and stable calibration does not necessarily immunize the portfolio. Market prices are subject to supply and demand. Since the buy-side and sell-side may have different rules (such as a short selling constraint) and asymmetric information, there is a possibility of price elevation, also known as a bubble. The typical approach, where model parameters are fitted to match the market prices, will not help you to manage the risk better in such a case.

There are other problems. A significant one in practice is that the meaning of vega hedging is ambiguous. One interpretation is that it is a hedge against the change of the portfolio value with respect to the change of implied volatilities (i.e. market prices of vanilla products). In this case, one has to re-calibrate the model by bumping the market prices to obtain the sensitivity. Another interpretation is that it is a hedge against the change of the portfolio value with respect to the change of volatility index, that is assumed observable but never is. In this case, the sensitivity is obtained from the model without necessarily bumping the market prices. The first one complies with the motivation of the mark-to-market framework. The second is more faithful to theory. Neither one is perfect. When these two are different, we are in serious trouble, as a wrong choice will give a mishedge.

The model for the asset and its volatility

We are going to work with a general stochastic volatility model:

$$dS(t) = \mu(S(t), Z(t), t)S(t)\ dt + \sigma(S(t), Z(t), t)S(t)\ dX_1(t)$$

and:

$$dZ(t) = p(S(t), Z(t), t)\ dt + q(S(t), Z(t), t)\ dX_2(t)$$

where X_1 and X_2 are standard Brownian motions under physical measure with an instantaneous correlation $d[X_1, X_2](t) = \rho(S(t), Z(t), t)\, dt$. If the coefficient function $\sigma(s, z, t) = z$, the above specification agrees with the classical setting. We'll only consider a non-dividend-paying asset; the modifications needed to allow for dividends are the usual. In what follows we will drop the time index (t) and function arguments $(S(t), Z(t), t)$ as long as the expressions are clear.

We are going to examine the statistical properties of a portfolio that tries to replicate as closely as possible the original option position. We will not hedge the portfolio dynamically with other options so our portfolio will not be risk free. Instead we will examine the mean and variance of the value of our portfolio as it varies through time.

With Π representing the discounted cash-flow of maintaining $-\Delta$ in the asset dynamically:

$$\Pi(t, T) = \sum_{i=1}^{n} e^{-r(\tau_i - t)} P(S(\tau_i), \tau_i) - \int_t^T e^{-r(\tau - t)} \Delta(dS(t) - rS(t)\, dt)$$

where P denotes a payoff of a contingent claim at $\tau_i \in [t, T]$, which can be a stopping time, and where r denotes a funding cost rate. One can make r time dependent, but we'll keep things simple here. Except those times when a claim is settled, the change in this cash-flow in time is continuous:

$$\Pi(t, T) = (1 - r\, dt)\Pi(t + dt, T) - \Delta(dS(t) - rS(t)\, dt) \tag{1}$$

We are going to vary Δ dynamically so as to replicate as closely as possible the option payoff. At expiration we will hold stock, and have a cash account containing the results of our trading. We are going to analyze the mean and the variance of our total position and interpret this in terms of option prices and probabilities.

Analysis of the mean

Naturally we are to determine the trading strategy Δ in a Markovian way. In fact, the stochastic control problem is reduced to a Markov control problem under a mild regularity condition, and therefore we will simply start from this for now. Define the mean (or the expected future cash flow) m at any time by:

$$m(S(t), Z(t), t) = E_t[\Pi(t, T)]$$

where the expectation E_t is a shorthand notation for the conditional expectation given the state of the world at time t. Using the equation (1), we obtain:

$$m = E_t[(1 - r\, dt)(m + dm) - \Delta(dS(t) - rS(t)\, dt)]$$

Thus, using Itô's formula we obtain the following partial differential equation (PDE):

$$\frac{\partial m}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 m}{\partial S^2} + \rho\sigma S q \frac{\partial^2 m}{\partial S \partial Z} + \frac{1}{2} q^2 \frac{\partial^2 m}{\partial Z^2} + \mu S \frac{\partial m}{\partial S} + p \frac{\partial m}{\partial Z} - rm = (\mu - r) S \Delta$$

Once again, we emphasize that all the drift coefficients are from the physical dynamic of the spot process not from risk-adjusted dynamic. For simplicity, we will write:

$$\mathcal{L} = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2}{\partial S^2} + \rho\sigma S q \frac{\partial^2}{\partial S \partial Z} + \frac{1}{2} q^2 \frac{\partial^2}{\partial Z^2} + \mu S \frac{\partial}{\partial S} + p \frac{\partial}{\partial Z}$$

and the equation for the mean becomes:

$$\frac{\partial m}{\partial t} + \mathcal{L} m - rm = (\mu - r) S \Delta \tag{2}$$

We still have to decide on Δ. We will choose it to minimize the variance locally, so we can't choose it until we've analyzed the variance in the next section. Note also that the final condition for (2) will be the payoff for our original option that we are trying to replicate.

This equation for m was easy to derive, the equation for the variance is a bit harder.

Analysis of the variance

The variance $v(S(t), Z(t), t)$ is defined by:

$$v(S(t), Z(t), t) = E_t[(\Pi(t, T) - m(S(t), Z(t), t))^2]$$

We may write:

$$\Pi(t, T) - m(S(t), Z(t), t) = (1 - r\, dt) A_1 + A_2 + O(dt)$$

where:

$$A_1 = \Pi(t + dt, T) - (m + dm)$$
$$A_2 = dm - \Delta\, dS$$

Also note that A_1 and A_2 are uncorrelated. Therefore:

$$v = E_t[(1 - r\, dt)^2 (v + dv) + (dm - \Delta\, dS)^2] + o(dt)$$

which further reduces to:

$$o(dt) = E_t[dv] - 2rv\, dt + E_t\left[\left(-\sigma S \Delta\, dX_1 + \frac{\partial m}{\partial Z} q\, dX_2 + \frac{\partial m}{\partial S} \sigma S\, dX_1\right)^2\right]$$

The end result, for an arbitrary Δ, is:

$$0 = \frac{\partial v}{\partial t} + \mathcal{L}v - 2rv + \sigma^2 S^2 \left(\frac{\partial m}{\partial S}\right)^2 + 2\rho\sigma Sq \frac{\partial m}{\partial S}\frac{\partial m}{\partial Z} + q^2 \left(\frac{\partial m}{\partial Z}\right)^2 + \sigma^2 S^2 \Delta^2 - 2\Delta \left(\sigma^2 S^2 \frac{\partial m}{\partial S} + \rho\sigma Sq \frac{\partial m}{\partial Z}\right) \tag{3}$$

Choosing Δ to minimize the variance

Only the last two terms in (3) contain Δ. We therefore choose Δ to minimize this quantity, to ensure that the variance in our portfolio is as small as possible. This gives:

$$\Delta = \frac{\partial m}{\partial S} + \frac{\rho q}{\sigma S}\frac{\partial m}{\partial Z} \tag{4}$$

The mean and variance equations

Define a risk-adjusted differential operator:

$$\mathcal{L}^* = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2}{\partial S^2} + \rho\sigma Sq \frac{\partial^2}{\partial S \partial Z} + \frac{1}{2}q^2 \frac{\partial^2}{\partial Z^2} + rS\frac{\partial}{\partial S} + p\frac{\partial}{\partial Z}$$

Substituting (4) into (2) and (3) we get:

$$\frac{\partial m}{\partial t} + \mathcal{L}^* m - rm = \frac{\mu - r}{\sigma}\rho q \frac{\partial m}{\partial Z} \tag{5}$$

and:

$$\frac{\partial v}{\partial t} + \mathcal{L}v - 2rv + q^2(1-\rho^2)\left(\frac{\partial m}{\partial Z}\right)^2 = 0 \tag{6}$$

The final conditions for these are obviously the payoff, for $m(S, Z, T)$, and zero for $v(S, Z, T)$. If the portfolio contains options with different maturities, the equations must satisfy the corresponding jump conditions as well.

Since the final condition for v is zero and the only 'forcing term' in (6) is $\left(\frac{\partial m}{\partial Z}\right)^2$, equation (6) shows that the only way we can have a perfect hedge is for either q to be zero, i.e. deterministic volatility, or to have $\rho = \pm 1$. In the latter case the asset and volatility (changes) are perfectly correlated. The solution of (5) is then different from the Black–Scholes solution.

Equation (5) is very much like the pricing equation for stochastic volatility in a risk-neutral setting. It's rather like having a market price of volatility risk of $(\mu - r)\rho/\sigma$. But, of course, the reasoning and model are completely different in our case.

The system of equations is nonlinear (actually two linear equations, coupled by a nonlinear forcing term). We are going to exploit this fact shortly.

How to interpret and use the mean and variance

Take an option position in a world with stochastic volatility, and delta hedge as proposed above. Because we cannot eliminate all the risk we cannot be certain how accurate our hedging will be. Think of the final value of the portfolio together with accumulated hedging as being the 'outcome.' The distribution of the outcome will generally not be Normal. The shape will depend very much on the option position we are hedging. But we have calculated both the mean and the variance of the hedged portfolio.

If the distribution of profit/loss were Normal then we could interpret the mean and the variance as in Figure 1.

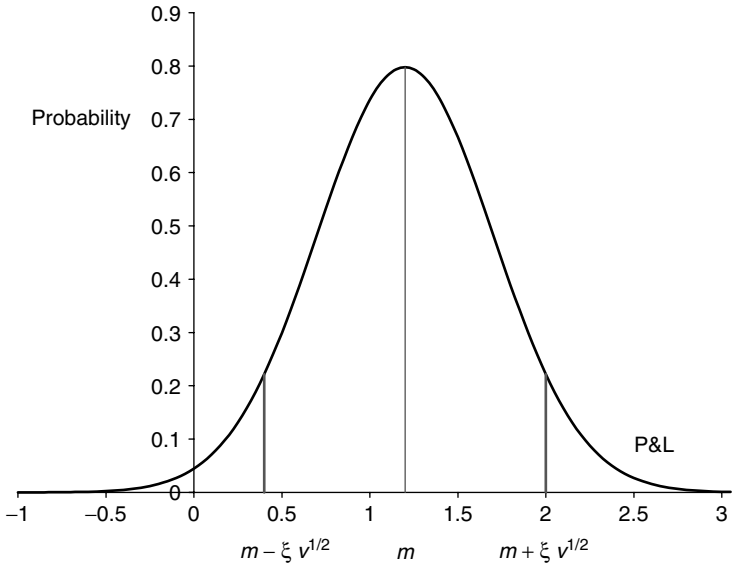


Figure 1: Distribution of profit/loss

Since this is likely to be one of very many trades, the Central Limit Theorem tells us that only the mean and the variance matter as far as our long-term profitability is concerned.

It is therefore natural to price the contract so as to ensure that it has a specified probability of being profitable. If we made the assumption that the distribution was not too far from Normal then the mean and the variance are sufficient to describe the probabilities of any outcome. If we wanted to be 95% certain that we would make money then we would have to sell the option for:

$$m + 1.644853v^{1/2}$$

or buy it for

$$m - 1.644853v^{1/2}$$

The 1.644853 comes from the position of the 95th percentile, assuming a Normal distribution.

More generally we would price at:

$$m \pm \xi \ v^{1/2}$$

where the ξ is a personal choice.

Clearly the larger ξ the greater the potential for profit from a single trade, see Figure 2.

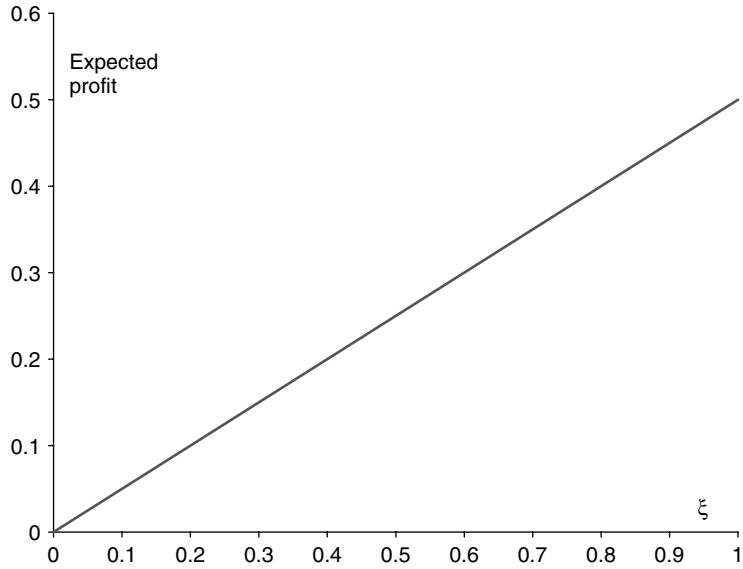


Figure 2: Expected profit from a single trade versus ξ

However, the larger ξ, the fewer trades, see Figure 3.

The net result is that the total profit potential, being a product of the number of trades and the profit from each trade, is of the form shown in Figure 4. Don't be too greedy or too generous.

We'll use this idea in the example below, but we will insist that we are within one standard deviation of the mean so that $\xi = 1$. This is simply so that we have fewer parameters to carry around.

Static hedging and portfolio optimization

If we use as our option (portfolio) 'price' the following:

$$\text{mean} - (\text{variance})^{1/2} = m - v^{1/2}$$

then we have a non-linear model.

Whenever we have a non linear model we have the potential for improving the price by static hedging (see Avellaneda and Parás, 1995, and Wilmott, 2000). This static hedging is, unlike the static hedging of linear problems, completely internally consistent. We will see how this works in the example.

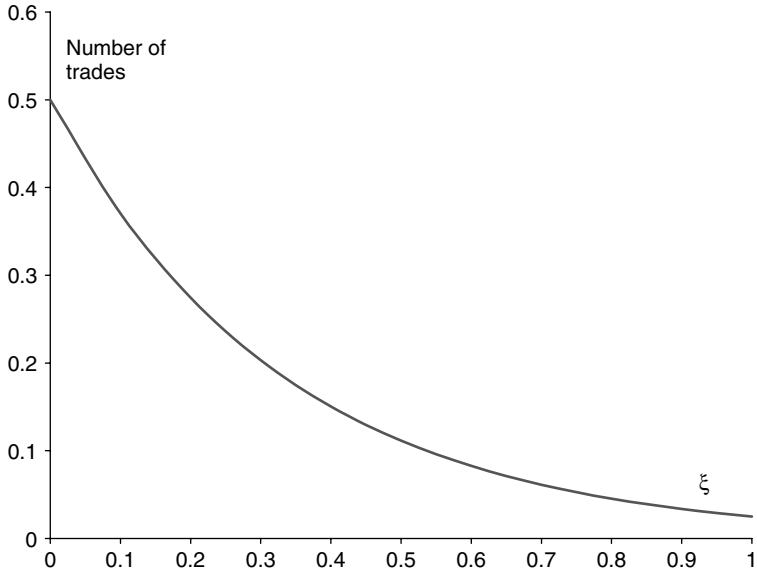


Figure 3: Number of trades versus ξ

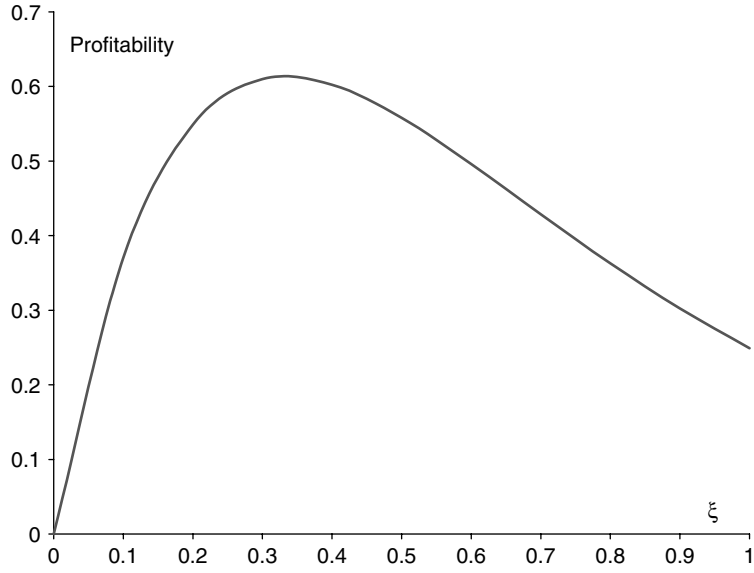


Figure 4: Total profit potential versus ξ

Example: valuing and hedging an up-and-out call

In this section, we consider the pricing and hedging of a short up-and-out call. Furthermore. we will consider a special case when the stochastic volatility is parameterized in a classical way:

$\sigma(S, Z, t) = Z$. Throughout this section, our choice of mean-variance combination is:

$$m - v^{1/2} \tag{7}$$

First consider a single up-and-out call with barrier located at S_u. In this case, we solve the equations (5) and (6) subject to:

(a) $m(S_u, \sigma, t) = v(S_u, \sigma, t) = 0$ for each $(\sigma, t) \in (0, \infty) \times [0, T)$ where T is maturity;
(b) $m(S, \sigma, T) = -\max(S - E, 0)$ for each $(S, \sigma) \in (0, X) \times (0, \infty)$ where E is the strike;
(c) $v(S, \sigma, T) = 0$ for each (S, σ).

The discontinuity of the payoff at the knock-out barrier makes this position particularly difficult to hedge. In fact this can be easily seen from our equations. Figure 5 and Figure 6 are the pictures of calculated mean and variance respectively with strike at 100, barrier at 110, and expiry in 30 days. We have chosen the model:

$$p(\sigma) = 0.8(\sigma^{-1} - 0.2), \quad q(\sigma) = 0.5 \quad \text{with} \quad \rho = 0.$$

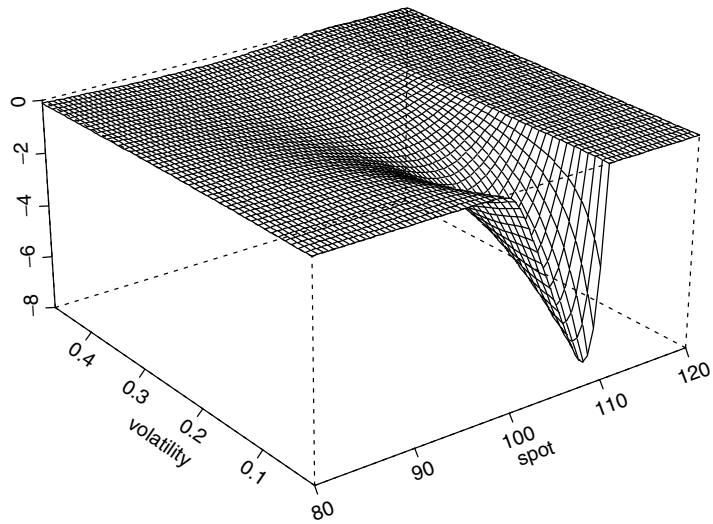


Figure 5: Mean for a single up-and-out call

Near the barrier, $\left(\frac{\partial m}{\partial \sigma}\right)^2$ is huge (see Figure 5) and this feeds the variance, being the source term in (6). If the spot S is 100, and the current spot volatility σ is 20% per annum, the mean is -1.1101 and the variance is 0.3290. Thus if there is no other instrument available in the market, one would price this option at \$1.6836 to match with Equation (7).

These results are shown in the table:

	Mean (m)	Var. (v)		Value
Unhedged	-1.1101	0.329		1.6836

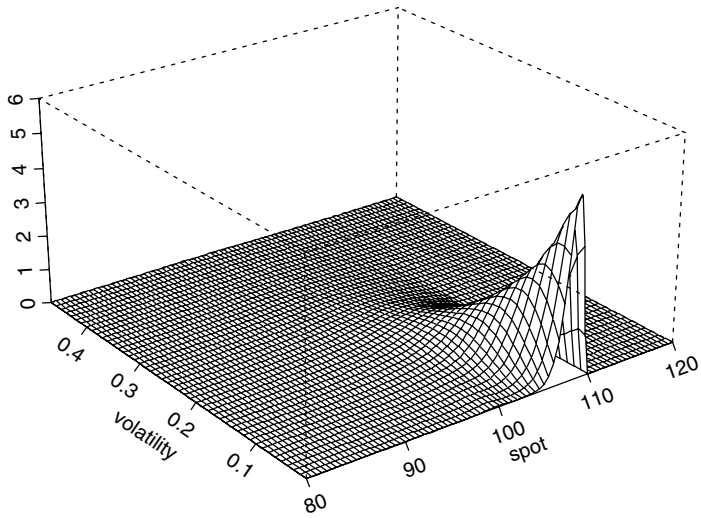


Figure 6: Variance for a single up-and-out call

Static hedging

Suppose that there are six 30-day vanilla call options available in the market with the following specifications:

Option	1	2	3	4	5	6
Strike	96.62	100.00	104.17	108.70	112.36	116.96
Bid Price	4.6186	2.6774	1.1895	0.4302	0.1770	0.0557
Ask Price	4.6650	2.7043	1.2014	0.4345	0.1788	0.0562

(**Aside:** These hypothetical market prices were generated by computing the mean of each call option, with:

$$d\sigma = \left(\frac{1}{\sigma} - 0.2\right) dt + 0.5\, dX_2 \tag{8}$$

where X is a Brownian motion with respect to the risk-neutral measure. Then 0.5% bid-ask spread was added.)

Now we employ the optimal static vega hedge. Suppose we trade $(q_1, \ldots, q_6)$ of the above instruments and let E_i be the strikes among the payoffs. Furthermore, let $(m^{(0)}, v^{(0)})$ be the mean variance pair after knock out and $(m^{(1)}, v^{(1)})$ be that before knock out. Then $(m^{(i)}, v^{(i)})$, $i = 0$, 1, satisfy the equations (5) and (6) subject to:

(a) $m^{(1)}(110, \sigma, t) = m^{(0)}(110, \sigma, t)$ and $v^{(1)}(110, \sigma, t) = v^{(0)}(110, \sigma, t)$ for each (σ, t) in $(0, \infty) \times [0, T)$;

(b) $m^{(0)}(S, \sigma, t) = \sum_{i=1}^{6} q_i \max(S - E_i, 0)$ for each $(S, \sigma) \in (0, \infty) \times (0, \infty)$;

(c) $m^{(1)}(S, \sigma, T) = \sum_{i=1}^{6} q_i \max(S - E_i, 0) - \max(S - 100, 0)$ for each (S, σ) in $(0, 110) \times (0, \infty)$;

(d) $v^{(1)}(S, \sigma, T) = v^{(0)}(S, \sigma, T) = 0$ for each (S, σ) in $(0, \infty) \times (0, \infty)$.

Thus $m^{(1)}(S, \sigma, 0)$ stands for the mean of the cashflows excluding the up-front premium. We find a $(q_1, \ldots, q_6)$ that maximizes:

$$m^{(1)}(S, \sigma, 0) - \sqrt{v^{(1)}(S, \sigma, 0)} - \sum_{i=1}^{6} p(q_i)$$

where $p(q_i)$ is the market price of trading q_i shares of strike E_i. In the case of $S = 100$ and $\sigma = 0.2$, our optimal choice for vega hedge is given by:

Option	1	2	3	4	5	6
Strike	96.62	100.00	104.17	108.70	112.36	116.96
Quantity	0.0000	−1.1688	1.0207	3.1674	−3.6186	0.8035

The cost of this hedge position is \$1.1863. Figure 7 and Figure 8 are the pictures of $m^{(1)}$ and $v^{(1)}$ after the optimal static vega hedge. After the optimal static vega hedge, the mean is 0.0398 and the variance is reduced to 0.0522. Thus the price for the up-and-out call that matches with our mean-variance combination (7) is $\$1.3752(1.1863 - 0.0398 + \sqrt{0.0522})$. In the risk-neutral set-up (8), the price for this up-and-out call is \$1.1256. The difference mainly comes from the standard deviation term (variance$^{1/2}$) in (7) which is $\sqrt{0.0522} = 0.2286$.

These results are shown in the table:

	Mean (m)	Var. (v)	Hedge	Value
Unhedged	−1.1101	0.329		1.6836
Hedged	0.0398	0.0522	1.1863	1.3752

By statically hedging we have reduced the price at which we can safely sell the option, from 1.6836 to 1.3752, while still making money 84% of the time. Alternatively, we can still sell the option for 1.6836 and make even more profit.

At the same time the variance has been dramatically reduced so that we are less exposed to volatility risk than if we had not statically hedged the position.

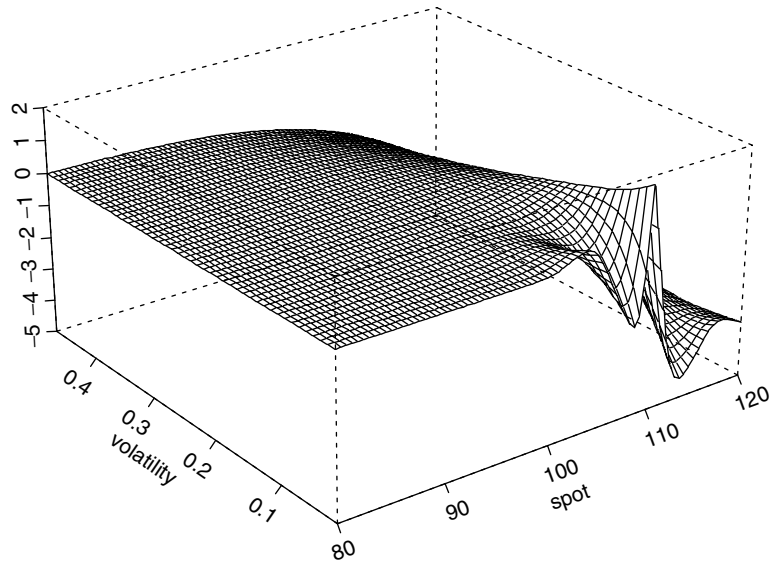


Figure 7: Mean of portfolio after optimal static vega hedging

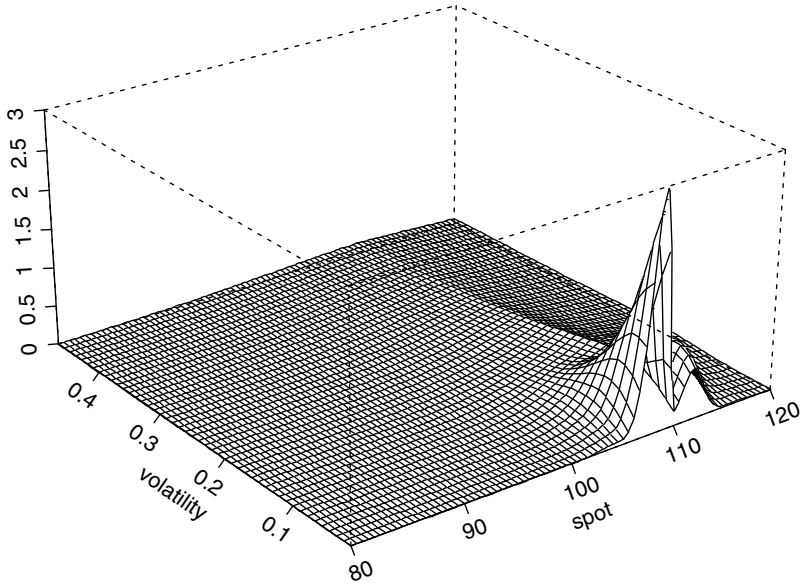


Figure 8: Variance of portfolio after optimal static vega hedging

Other definitions of 'value'

In the above example we have statically hedged so as to find the best value according to our definition of value. This is by no means the only static hedging strategy. One can readily imagine different players having different criteria.

Obvious strategies that spring to mind are as follows:

- Minimize variance, that is minimize the function v. This has the effect of reducing model risk as much as possible using all available instruments (the underlying and all traded options). This may be a strategy adopted by the sell side.
- Maximize the returns risk ratio. This is perhaps more of a buy-side strategy, for maximizing Sharpe ratio, for example.

Summary

Constructing a risk-neutral model to fit the market prices of exchange traded options consistently over a reasonable time period is a difficult task. Putting aside the fundamental question of whether the axiomatic risk-neutral model for stochastic volatility is legitimate or not, we must face potential financial losses due to re-calibration. In this paper we have taken another approach. We first evaluate the mean and variance of the discounted future cashflow and then find market instruments that reduce the volatility risk optimally.

We've set this problem up in a mean-variance framework but it could easily be extended to a more general utility theory approach.

REFERENCES

■ Ahn, H., Bouabci, M. and Penaud, A. (2000) Tailor-made for tails, *RISK Magazine*, January.
■ Avellaneda, M. and Parás, A. (1996) Managing the volatility risk of derivative securities: the Lagrangian volatility model: *Applied Mathematical Finance* 3, 21–53.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance*, Wiley, Chichester.

Index