
WILMOTT COLLECTION
THE BEST OF
WILMOTT 2

The Best of Wilmott

Volume 2

Edited by

Paul Wilmott

John Wiley & Sons, Ltd

The Best of Wilmott

Volume 2

The Best of Wilmott

Volume 2

Edited by

Paul Wilmott

John Wiley & Sons, Ltd

Published in 2006 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.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, 42 McDougall Street, 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.

Library of Congress Cataloging-in-Publication Data

The best of Wilmott 2 / edited by Paul Wilmott.
p. cm.
Includes bibliographical references and index.
ISBN-13 978-0-470-01738-8 (cloth : alk. paper)
ISBN-10 0-470-01738-4 (cloth : alk. paper)
1. Derivative securities. 2. Finance—Mathematical models. 3. Risk management.
4. Options (Finance) I. Title: Best of Wilmott two. II. Wilmott, Paul.
HG6024.A3B517 2005
332.64′5—dc22 2005020005

British Library Cataloguing in Publication Data

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

ISBN-13 978-0-470-01738-8 (cloth : alk. paper)
ISBN-10 0-470-01738-4 (cloth : alk. paper)

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

Preface

The team at Wilmott is very proud to present this compilation of magazine articles and presentations from our second year. We have selected some of the very best in cutting-edge research, and the most illuminating of our regular columns. Our columnist, the Collector, contributes his infamous 'Know Your Weapon' series in which he espouses the principle that it is more important to have a robust model that you understand than a fancy one you don't. Dr Z gets down to basic concepts of money management, and Aaron gives us a history lesson.

The technical papers include state-of-the-art pricing tools and models. You'll notice there's a bias towards volatility modelling in the book. Of course, it's one of my favourite topics, but volatility is also the big unknown as far as pricing and hedging are concerned. We present research in this area from some of the best newcomers in this field. You'll see ideas that make a mockery of 'received wisdom', ideas that are truly paradigm shattering – for we aren't content with a mere 'shift'. Several of these articles are from that hive of original thought that is ITO33. Elie Ayache has also written his own introduction to this compilation. And, in true French philosopher tradition, he's been at the absinthe again!

Finally a big 'thank you' to all supporters, the subscribers and the sponsors!

Paul Wilmott
2005

Foreword

Elie Ayache

And so it fell to me to write an introduction for *Best of Wilmott 2*. To quote from the introduction of *Best of Wilmott 1*, by Paul Wilmott: 'In September 2002 a small, keen group... joined forces with a book publisher to create a new magazine, *Wilmott*...'

'*In September 2002*', '*Create*', '*New*': These words speak of birth and novelty; they set a 'source point'. Somehow Paul's attempt at introducing *Best of Wilmott 1* is easier than mine today. His introduction is self-giving and originary, whereas mine is a sequel. Mine is unoriginal and derivative. Also, the title of the first book speaks for itself: 'This is the *first* edition of the *best* of *Wilmott*.' What better way to present a subject than the conjunction of these two superlatives?

'*To write*', '*Introduction*': Mark these words as I will revisit them later and remark on them. To give you a hint: *This is a book about derivatives and derivatives are essentially all about writing* — they are said to be written on the underlying. How then do you introduce the derivatives or write about them? By first introducing their underlying? And how do you introduce that? By floating it? (The French word for 'floating' is '*introduire en bourse*'.) What better way of introducing the derivatives than joining their market at once? Shouldn't we all stop writing and start trading? And how can you introduce a market, or *introduce* somebody to trading?

Why me?

The name of Paul Wilmott imposes itself as best introducer of *Best of Wilmott*. I have been considering a variant of the title with the name of Wilmott crossed out. In private correspondence Paul Wilmott indeed refers to the book simply as *Best of 2*. Call it selflessness, or self-evidence. Simply, the man could not get over speaking both in his name and for his name. Imagine him asking me: 'Could you please write me an introduction for *Best of me, volume 2*?'

'*Best of 2*': The formula almost strikes me like a derivative payoff. And this suits my purpose just fine. As it severs the link with the original name of the initial introducer, this elliptical formula seems, as a consequence, to dispense with personality and proper name altogether. Writing is impersonal. Just as anybody can write a derivative payoff, anybody can write an introduction for *Best of 2*. The market is impersonal. Writing derivatives is just a way of handing back to the market, i.e. to impersonality, the skewed and exotic and idiosyncratic scenarios that the market may have

Contact address: ITO33 SA, 36 rue Lacépède, 75005 Paris, France
E-mail: NumberSix@ito33.com

inspired you personally. Writing is derivative. It always comes after speech. A compilation book always comes after the articles compiled in the book, and the introduction of the compilation book always comes after the compilation book, never before—have you noticed? (Not mentioning that volume 2 always comes after volume 1.)

Let us pursue the thread of the derivative for a while, that is to say, of impersonality and unoriginality, and let us forget about the best and the privilege of writing about the *Best*. Essentially, what I inherit today is the endless task of rewriting. Since a compilation book is a repackaging and a rewriting of articles initially published in the magazine, writing an introduction for the compilation book is writing about the rewriting of articles initially written about the derivatives which are all about writing. How can I even *start* to do that? Did the writing of derivatives start one day or has it always been going on? Did the writing of articles about the derivatives start one day? Did the market start one day? Or has the writing always been going on? From my personal and localized point of view, something has definitely always preceded my writing. This is volume 2, remember?

Having thus dissolved the superlative and the privilege of introducing it in the impersonal chain of writing, I may as well move, without further notice or introduction, to what interests me personally. I am not Paul Wilmott after all, the editor-in-chief and impartial arbitrator of *Wilmott*, so the reader will have to excuse a little extremism on my part. And what interests me, what interests me in general and in the particular instance (which is, as expected, an instance of writing and rewriting and writing about writing), what has always interested me to the exclusion of anything else, is *replication*. (Imitation?)

When you neutralize the primary meaning of the *best of* (the value judgement) and retain only the derivative meaning (of a compilation and a rewriting), all you end up with is a replication argument. Buy this book, so the argument goes, invest in it an initial fee, and you will have replicated a process of writing, editing and publishing that has lasted for a whole year. From which it appears that the process of selection of the 'best' articles—whose other side is the rejection of others—is just the necessary consequence of idealization. It has nothing to do with good or bad, with best or worst, only with relevant and significant. It is a modelling assumption like any other, with its expected share of choice and sacrifice. Mustn't you specify a robust dynamic model before you try to replicate a given payoff?

All of which brings me to volatility. And to writing an introduction for the second issue of *Best of Wilmott*, where there is contained, as you will see, a lot of volatility papers. How do you introduce volatility? Isn't it, by essence, the subject that has always already started and has always already been introduced? A lot has been written on volatility (otherwise, I wouldn't be today in the position of writing an introduction for a compilation of papers written about volatility). However, what interests me in volatility today—as you must have guessed by now—is to write about it *derivatively*. Not only because I am in the business of writing about the writing of volatility papers, but because volatility, as an original and underived concept, is now disappearing everywhere. 'Writing about volatility derivatively'—for those who didn't catch my drift—just means 'writing about volatility *by way of the derivative*'. What else? Where is volatility to be observed in the world, apart from the traded prices of derivative instruments?

I might as well say it straight, at the risk of shocking the reader and shaking him (but this, according to Paul Wilmott in the introduction of *Best of 1*, is exactly what I am supposed to do): *There is no meaning to volatility outside the derivative and nobody today knows how to price the derivative!* 'A lot has been written on volatility' therefore can only mean 'A lot of derivatives have been written', and it is only through this writing, which is constantly submitted to the impersonal rewriting of the market, that volatility can mean anything at all and ever get introduced. Volatility as the (unobservable) measure of risk, volatility as historical volatility, does not in the least interest us. And certainly no book—let alone the introduction to a book—can teach us its meaning. Volatility can only be meaningful within the *language* of volatility, which is the language of derivative prices. It can only be meaningful within the fabric of the derivative market, that is to say, the market as both a texture and a text.

So volatility can only mean something in the derivative sense of *volatility-for-a-derivative*. And this, my dear reader, is all about replication. The only way to introduce the subject of volatility today is to kiss goodbye to the myth of the origin and the myth of the original, introductory talk about volatility—to kiss goodbye to the models where volatility is posited as an independent and originary parameter. It is to join at once a market with no origin or starting point, where the only activity is the activity of derivative writing and model rewriting and the only sense one can make of a derivative price is the cost of replicating its payoff with other derivative instruments (which may include the underlying), under a dynamics previously calibrated with the market prices of the latter. There is no volatility or derivative pricing models *per se*, only recalibration and replication episodes.

This of course brings me close to fulfilling the task of writing about writing on a subject (volatility, the derivatives) which is all about writing, in other words, the task of rehearsing, in an introduction, nothing more than a replication argument. It also leaves me with a question that no writing or replication can help answer: 'How the hell is anybody able to price a CDO?'

FOOTNOTE

1. And I don't mean implied volatility, as this concept is dying with the Black–Scholes paradigm and the derivative pricing models now imply several parameters.

1
Time's Up

Dan Tudball

Dan Tudball winds back the clock and takes a look at the major issues of 2004 and what they might bode for 2005.

Timetables were supreme in 2004. The ticking of the clock was omnipresent and frighteningly audible and no doubt many in the industry wished they could jump into a temporal vortex and transport themselves back a couple of decades. Back a few decades before Enron, Worldcom, Adecco, to a time when a gentleman's word and some academic credentials might have been enough. But 2004 was the year of the timetable, and no such time tunnel was opening up promising a return to the comfort of the unquestioning past.

Accountancy was to the fore this year. Internal risk management the repository of both hope for the future of the industry and the focus of questions of 'who watches the watchmen?' To be custodian of both the firm's profitability and public perception? A question lost in the rush to comply as Sarbanes Oxley 404 became a reality, and public accountancy firms found themselves reinvigorated after their time in the wilderness.

The year has brought, under the demands of regulation, new questions and new channels of communication for the quantitative finance community. Old pastures meanwhile have looked less than fertile, with equities mostly inactive after the rebound of 2003. Exciting new departures, such as volatility trading, have faced a shakeup after a false dawn back in late 2003. The credit derivatives market continues to excite, and grow at a staggering pace. Tightening margins and technological development have forced the sell side to innovate ever more complex structured trades. Meanwhile, thanks to the very global nature of the market at this point foreign exchange has been blooming, despite many a premature obituary. Finally, on the periphery movements have been made to introduce brand new markets which may represent a massive opportunity in 2005.

Quis custodiet custodes?

As an issue corporate governance and regulatory accounting have been ever present throughout 2004. The heavyweight Sarbanes Oxley Act section 404 (SOX 404), Management's Reports on Internal Control Over Financial Reporting and Certification of Disclosure in Exchange Act Periodic Reports, has been dominant. More specific compliance rules such as the International Accounting Standards Board's amendments to standards on financial instruments disclosure and presentation (IAS32) and recognition and measurement (IAS39) have also been front and center in discussion and activity.

The flurry of activity, of course, has all been down to deadlines. SOX 404 had to be implemented in the States in 2004 beginning with entities whose financial year ended last November. For foreign private issuers and non-US institutions it's in effect at the end of calendar year 2005. IAS32 and IAS39 are effective as of January 2005. Both amendments bring non-US-based financial institutions under a near identical regime to that which the Federal Accounting Standards Board presides over in the United States, and represent a further move towards convergence in accounting oversight globally.

Preparation for SOX 404 requires two things. It requires management to provide an attestation as to the sufficiency of the financial controls and it requires the external audit firm to do two things, one to review management's attestation and then to do their own evaluation to report back. 'The Act requires company's management to conduct an assessment as to the company's internal control over financial reporting.'

It is the aspect dealing with the sufficiency of the financial controls that has naturally prompted questions within some quarters of the quantitative finance community. Most commonly, concerns have been voiced over whether or not accountants can really adequately assess the risks and methodologies employed in complex trading. 'If you were to ask "Would your average accountant be able to do this?" Probably not', says Chris Lucas, at PricewaterhouseCoopers in London. 'But part of the preparation for this is the involvement of specialists both in internal audit functions and in the public accounting arena who aren't actually qualified accountants. They may be qualified risk managers, ex traders etc. I'm not suggesting that it's easy, but those people are available to complement the core criteria skills which are financial reporting and control activities.'

The question is really a straw man. Firstly the contemporary place that quantitative skills have in financial institutions dates the original contention. Quantitative finance is pervasive, and this is something to celebrate. Despite the poor performance and shocks of the past years, quant hiring has at worst remained steady and in other instances boomed in response to regulatory demands. Internal risk management exists as a result of quants as much as price discovery, model validation and program trading. To find someone within the institution to explain the methodology will not be difficult. To find a third party able to understand that methodology and corroborate its findings is also not difficult. More quants contracted by public accounting firms, more competition to hire the best candidates. Who benefits here?

The second aspect really goes to the heart of why standards are a necessity. 'You can either treat compliance as an evil and look for minimum compliance or you can use it as an opportunity to enhance the reputation of the organization,' Lucas explains. 'Clearly history has shown what happens if people are not able to comply, the newspapers are littered with stories of investment banks which have struggled. The flip side is never getting yourself into that position, but the other thing is recognizing that recognition and brand are important and are an important part of running the business. On a macro "what is in it for me?" level, it helps people be reassured about the type of organization they are dealing with. It helps regulators form that positive view; it certainly assists when you are trying to get regulatory approval for acquisitions or strategic transactions or whatever. So I think there are small specific advantages, but there is a broader view in terms of what the broader stakeholders see of the organization.'

Oh, so COSO

The Committee of Sponsoring Organizations of the Treadway Commission

COSO was originally formed in 1985 to sponsor the National Commission on Fraudulent Financial Reporting, an independent private sector initiative which studied the causal factors that can lead to fraudulent financial reporting and developed recommendations for public companies and their independent auditors, for the SEC and other regulators, and for educational institutions.

The National Commission was jointly sponsored by five major professional associations in the United States, the American Accounting Association, the American Institute of Certified Public Accountants, the Financial Executives Institute, the Institute of Internal Auditors, and the National Association of Accountants (now the Institute of Management Accountants). The Commission was wholly independent of each of the sponsoring organizations, and contained representatives from industry, public accounting, investment firms, and the New York Stock Exchange.

The Chairman of the National Commission was James C. Treadway, Jr., Executive Vice President and General Counsel, Paine Webber Incorporated and a former Commissioner of the US Securities and Exchange Commission. (Hence, the popular name Treadway Commission.) Currently, the COSO Chairman is John Flaherty, Chairman, Retired Vice President and General Auditor for PepsiCo Inc.

Internal control is a process, effected by an entity's board of directors, management and other personnel, designed to provide reasonable assurance regarding the achievement of objectives in the following categories:

- Effectiveness and efficiency of operations
- Reliability of financial reporting
- Compliance with applicable laws and regulations

Key concepts

- Internal control is a process. It is a means to an end, not an end in itself.
- Internal control is effected by people. It's not merely policy manuals and forms, but people at every level of an organization.
- Internal control can be expected to provide only reasonable assurance, not absolute assurance, to an entity's management and board.
- Internal control is geared to the achievement of objectives in one or more separate but overlapping categories.

Source: `www.COSO.org`

Cynics might say that public accountancy firms were thrown a bone after the catastrophes of Enron, Worldcom, Adecco, you know the litany. Well, true enough, public accountancy firms have turned in a very healthy profit in the last year—but that's a natural part of a cycle that predicates those firms' existence. No third party, no audit, not much business done.

An interesting part of the compliance process has been the contrast between the two major aspects under scrutiny. Investment banks have internal checks in place, are able to explain the process by which financial instruments work and are able to put a figure to it. If they don't have these things then it is very surprising that they are in business at all. The framework has been in place since the 1980s in the guise of COSO (Oh, so COSO) and has informed the proper running of investment banks since even before Nick Leeson went wild on Commodity Quay.

No fair!

The IASB raised that old bugbear 'fair value', but this time got their way

Accounting standards are hardly the most exciting proposal in the world, but if you brought in the issue of fair value reporting then many a face on the banking side would be aflush with fury. This has certainly been the case for at least 15 years. Now the IASB has enshrined the principle as a preferred option in standards related to financial instrument disclosure and presentation. But why the fuss?

The banking industry has long preferred to take a modified historical cost basis approach to measuring banking book performance. One major reason why is that in marking financial instruments to the current market value of the underlying increases the volatility of earnings. The counterargument has been that fair value assessments improve the ability to forecast violations of requirements.

Despite the view that fair value assessment would, for example, force lenders to ignore higher risk borrowers and cause a flight to quality—thus undermining banks' roles as long-term lenders—it is the change in public perception of these issues that has largely informed the result.

The world knows what happens when derivatives are not marked to market in the books, and it has happened too often for the argument to be acceptable anymore. If anything this set of new standards will create further impetus for structural change within traditional investment banking and greater impetus for the growth of the, still lightly regulated, hedge fund industry across the board. Watch this space.

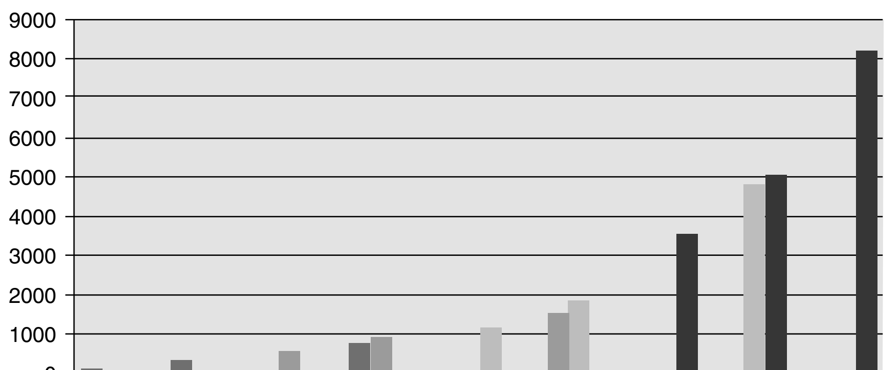
'A lot of the discussion is around the detailed control activity as opposed to financial reporting, and that falls into the area of hard data,' Lucas explains. 'But there is another element of the COSO framework which looks at the entity level controls, which look at things like tone from the top, overall control environment etc. I think those are the more difficult to measure pieces to assess. Some of the preparation has been around that. An example would be the extent to which there are codes of conduct, but then we drill down to how does it get distributed, how do we know employees read it, how does it get translated into the language of the employees. That's where you're looking for harder evidence in relatively soft areas.'

A fact of life

How many more times can people say that credit derivatives have been and are inescapable? This year should see the end of it, largely because the market is here and highly unlikely to go away. In April of 2004 the worst kept secret of the year finally was secret no more. In 2003 the main discussion within credit derivatives circles had been around the need for a single strong index to take the market onto the next evolutionary step. The situation at that point was a duopoly—or the less kind would say it was a monopoly with an understudy. Trac-X was the frontrunner, created by JP Morgan and Morgan Stanley, then handed over to Dow Jones Indexes in early 2004. Iboxx was London registered and shareholders were 'just about everyone else'.

After much talk of power broking the inevitable happened with, we are sure, a little ceding of power here and a little circumspection there. Iboxx and Trac-X merged to form Dow Jones iTraxx.

The new figure on everyone's lips was \$8.2 trillion. This is where the global market in credit derivatives is expected to be by the end of 2006 according to a report by the British Bankers Association in the last quarter of 2004; that figure, however, only stands if you include asset swaps. Without including them the figure is somewhat more conservative; at the end of 2004 the figure is estimated to stand at around \$5 billion and is estimated to exceed \$8 billion by the end of 2006. In terms of growth, however, this is still remarkable, with the market having grown by 50% over the last year alone (see Figure 1).



	1997	1998	1999	2000	2001	2002	2003	2004	1996
'97/'98 survey	108	350		740					
'99/2000 survey			586	893		1,581			
2001/'02 survey					1,189	1,952		4,799	
2003/'04 survey							3,548	5,021	8,206

Figure 1: Global credit derivatives market (excluding asset swaps in \$bns)

In terms of product categories (see Figure 2) credit default swaps still outstrip other classifications; however, what is notable is a reduction in synthetics—which is largely accounted for by the increasing significance of subcategories therein—and the emergence of indices. What is likely to be a massive influence in 2005 is the opening up of emerging markets (see Indicators, below).

At the center of all this is Mark It Partners. Originally an offshoot of TD Securities, Mark It owns RED, the repository for 99% of the world's reference data, and absolutely essential to the

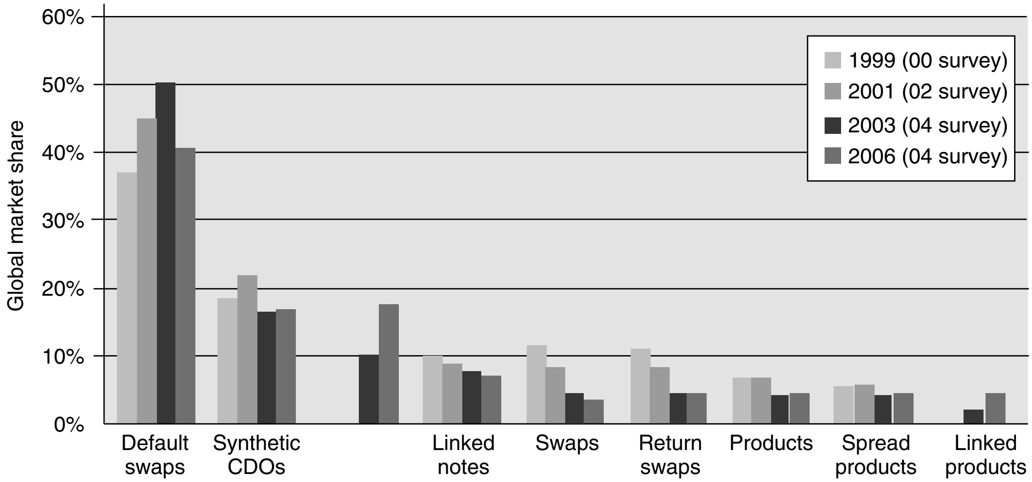


Figure 2: Credit derivatives products

running and growth of the credit derivatives market. The year has been an active one, explains Lance Uggla, CEO of Mark It, with acquisitions such as the high profile Totem Partners.

'It's been a year of consolidating and standardization of the reference data and also making more transparent the underlying data that supports the market,' says Uggla. 'By having standard reference entities it makes it very easy to transfer data accurately, and to consolidate that data accurately and make it available to research, trading, sales, origination, hedge funds, asset managers, insurance companies, rating agencies. The data has been made readily available to a lot of users; we've over 250 unique customers now.'

'Mark It through Totem Valuations does the price testing in the correlation space with banks,' says Uggla. 'Within the broker you can see quotes that give evidence of the correlation in the index space. Now is that same correlation that you are seeing in the index space representative of correlation in something more bespoke? In the index space it's a known set of names and correlation can be executed in a quite liquid fashion in a broker around the index names, then within banks. Banks do all these customized CDOs or bespoke CDOs and is that same correlation existing in bespoke products as is existing in the index tranches? Those are the quantitative discussions, which are occurring right now — we are active in those discussions. I think as more and more information becomes available then developing and fine-tuning models in the bespoke areas become a little easier.'

Keep on keeping on

'Retirement at 65 is ridiculous. When I was 65 I still had pimples.' The foreign exchange markets are very much the George Burns of the financial world, the number of times a sell-by date has been applied and then hastily removed would, to mix metaphors completely beyond reason, put a bikini wax specialist to shame. But such is the case. Since 2001, however, the volume of foreign exchange transactions has grown by 57%.

Within this sample OTC derivatives, consisting of 'non-traditional' foreign exchange derivatives (cross currency swaps etc.) and all interest rate derivatives, have seen average daily turnover increase by 112% over the same period.

The Bank of International Settlements reports, in its Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity:

> The growth in turnover was driven by all types of counterparties. Trading between banks and financial customers rose markedly, and its share in total turnover went up from 28 per cent to 33 per cent. Based on market commentary, the higher activity between reporting banks and financial customers may to a large extent have reflected a sizeable increase in activity by hedge funds and commodity trading advisers, as well as robust growth of trading by asset managers. This is in contrast with the period between 1998 and 2001, when activity in this market segment had been driven mainly by asset managers, while the role of hedge funds had reportedly declined. Trading between reporting dealers also rose between 2001 and 2004, although its share continued to fall, from 59 per cent in 2001 to 53 per cent in 2004. Restraining factors might include the continuing consolidation in the banking industry, as well as efficiency gains derived from the use of electronic brokers in the interbank spot market. For its part, the share of trading between banks and non-financial customers edged up slightly to 14 per cent.

Trading was again largely driven by the combination of a weak US dollar but also augmented by the inaction on the stock markets over the year. Dollar–euro pairings accounted for 28% of daily turnover, dollar–yen was 17% and dollar–sterling took 14%. Bets on the dollar's continued decline against the euro were heavy throughout the year. With no sign of the US winding back

TABLE 1: GLOBAL FOREIGN EXCHANGE MARKET TURNOVER[a]
(daily averages in April, in billions of US dollars)

Instrument	1989	1992	1995	1998	2001	2004
Spot transactions	317	394	494	568	387	621
Outright forwards	27	58	97	128	131	208
Foreign exchange swaps	190	324	546	734	656	944
Estimated gaps in reporting	56	44	53	60	261	07
Total 'traditional' turnover	590	820	1,190	1,490	1,200	1,880
Memorandum item:						
Turnover at April 2004 exchange rates[b]	650	840	1,120	1,590	1,380	1,880

[a]Adjusted for local and cross-border double-counting.
[b]Non-US dollar legs of foreign currency transactions were converted into original currency amounts at average exchange rates for April of each survey year and then reconverted into US dollar amounts at average April 2004 exchange rates.

its unstated weak dollar policy, the need for businesses with large revenue streams from the US to hedge against swings will remain a constant.

Volatile behavior

The year before last was seat-of-the-pants stuff for most people in the equities markets. After the nightmare of 2002, 2003 looked to be beginning in the same fashion. Down. But then, after the first quarter, the choppy, yet inexorable, rise began—the markets were transfixed. Volatility was king.

But then 2004 happened. One long yawn—barely a shift over the long term and volatility was now as hard to come by as vodka at an AA meeting. Despite this, though, industry estimates put the total assets under management by long volatility funds at between $1.5 and 2 billion–an up to 33% rise on the same time last year. So what's happening?

Rami Habib runs the FIMAT volatility funds index under the auspices of SocGen in London, and has his finger on the pulse of a thoroughly exciting (whether for good or bad) market.

'The main reason why we chose to look at the vol-arb space was because at the time—a year and a half ago—I found that if I was talking to investors, there was very little understanding of what vol-arb managers were doing,' explains Habib. 'They could understand long-short equity bias, and although they would say that they were well diversified across many different hedge funds they would still have a long equity bias—so where they thought they had a well-diversified portfolio it wasn't really all that well diversified.'

Habib started to hear investors suggesting they wanted to look at the vol-arb space. 'Most people saw vol-arb funds as being long volatility funds so they would have a long options profile, they wouldn't do a great deal for 90% of the year, they would be the hedge of the portfolio. They would have one vol-arb manager and they wouldn't care what he did most of the time unless there was a situation.'

Naturally, in the early days most of the managers Habib and his team were talking to were long volatility purely in equity markets. 'As time has gone by, with the squeeze in volatility, managers have suffered. Some of the long-running long-vol managers who have been going since the mid nineties are still around and will probably still be around in years to come even though they have had a drawdown of 30 to 40%.' But these managers have proved they can make money in the past and they will keep on going. 'We saw quite a few long volatility managers launching just prior to October 2003,' Habib says. 'Starting a fund with a 30% drawdown is not going to help anybody, so we've seen quite a few closures. None of these funds were ever really able to hit their stride. Quite a few managers are looking at volatility as a relative value strategy. They don't necessarily have to have a long bias. Also there are many who are looking further than equity and are looking at the fixed income side and the commodities side and currencies. So the opportunities are there for managers who are becoming more like a global macro with a volatility focus.'

'We're probably having the most interest in volatility funds opening up that we've had for a long time,' Habib reports. 'The thing that has saved this market is that the whole hedge fund market has been difficult throughout the year. It's not that long-vol funds have done particularly well but that is in line with other strategies being hurt as well.'

As for next year, Habib says, with volatility so low so long, people are itching to take a bet on it starting up soon. 'Now we are seeing standalone funds which are alpha generated, new territory for some of the funds. We've not seen significant asset outflows. We're seeing a change in the field with money going from the long-vol funds into relative value operations. It's a strategy where the universe is still small, there is difficulty in finding two guys doing the same thing, and there are still few managers who are sufficiently knowledgeable to make this kind of strategy work.'

Indicators

Options markets up, up, up

With equities somnambulant it was no surprise that 2004 was a banner year for the options exchanges. The Chicago Board Options Exchange (CBOE) reported that October volume totaled 33 357 205 contracts traded, an increase of 17% over the October 2003 volume of 28 635 741 contracts. Through the end of October, CBOE's year-to-date volume of over 295 million contracts traded is up 27% over 2003 on track to establish a new all-time annual volume record, surpassing the previous high of 326 million contracts in 2000.

The Chicago Board of Trade (CBOT) announced that total exchange volume continued its strong growth, reaching 47 830 745 contracts in October, up 12.8% from last year. Year-to-date (YTD) volume through October was up 29.5% to 494 383 000 from January through October last year. Average daily volume in October increased 23.6% to 2 277 655 contracts from October 2003 levels.

CBOT President and CEO Bernard Dan said, 'In October the CBOT reached a new all-time annual trading volume record, surpassing the prior record set by the exchange in 2003.

'The impressive gains in our volume underscore the confidence our customers have in the CBOT's risk management products, in its superior electronic trading platform, innovative clearing system, and in its markets, known for their liquidity, transparency and integrity.'

Approximately 88.4 million contracts were traded on the international derivatives market Eurex in October. This equates to an average daily volume of approximately 4.2 million contracts. At roughly 893 million contracts, total turnover for the current year exceeds previous-year levels by around 20 million contracts or 2%. Furthermore, the world's largest derivatives market recorded its highest open interest to date with open interest of 76 million contracts. The number of open positions has climbed 22% since October 2003.

Outsourcing

It's not been talked about too openly but following on from the well-known customer service outsourcing, vendors in India are providing more and more analytics and research to departments in Europe and the US. It's a natural result of greater regulatory demands to quantify research and the corresponding squeeze this puts on the bottom line. Also, with the squeeze on fees occurring for hedge funds, outsourcing model testing and the like to firms in the subcontinent.

Some are going further, with funds setting up dedicated teams in India whilst running their operations in the main financial market centers. Shariar Shahida of New York's Constellation fund told the *Financial Times* that 'We haven't moved the advanced quant work to India yet . . . but there's no reason we couldn't do that in the future.'

One interesting aspect to look out for over the next year will be the natural result of this labor arbitrage. With reports estimating annual growth of 45% for outsourced jobs in India and a total of 1 million Indians employed by an outsource vendor, it is only a matter of time before secondary markets begin to move in to take advantage of the price increase which will inevitably occur.

Real estate

A quiet revolution occurred in the UK in March 2004. After many years of lobbying the UK government brought taxation on property derivatives in line with that of other derivatives. This has been the major stumbling block in making index-based property derivatives a viable offering thus far.

Property is the last asset class remaining without a liquid derivatives market either in the UK or the US. In the UK, it has been argued that the ideal index exists for a nascent liquid derivatives market. The Investment Property Databank (IPD) Index was established in 1986, it is now based on over 12 000 commercial properties with a current value in excess of £100 billion. This represents some 75% of the total institutional investment property market.

The IPD also publishes a UK Monthly Index, which is increasing in importance. Both the Annual and Monthly Indices provide data on Capital Growth, Income Return as well as Total Return. Most major property market participants both contribute to and use these indices. This is all according to a report by Deutsche Bank.

2
First Cause

Dan Tudball

Louis Bachelier's *Théorie de la spéculation* defied categorization, but its ideas gave birth to the field of mathematical finance. Dan Tudball looks at the life and work of the man who started it all ...

Quantitative finance enjoys a rare distinction amongst the sciences in being able to identify the single event that brought it into existence. When Louis Bachelier successfully defended his thesis *Théorie de la spéculation* on March 29th 1900 he effectively inaugurated year zero on the quantitative finance calendar. March 29th really ought to be marked by champagne toasts and new resolutions across investment banks, hedge funds and campuses the world over.

The subject of Louis Bachelier has developed a mini area of study in itself, with a few particularly committed researchers dedicated to discovering more about the man dubbed the father of mathematical finance. Bachelier's work contains so much that is familiar today, but predates the work of so many whose contributions were acknowledged during their lifetimes. His work influenced Wiener, Kolmogorov, Ito, Black, Scholes and Merton to name but a few. Periodically 'rediscovered' over the last hundred years Bachelier's contribution has a talismanic quality about it.

The road to defending the thesis was not an easy one. Bachelier's life is strewn with misfortune and tragic misunderstanding, obstacles which, had they not been present, might have allowed the acceptance of finance as a legitimate area of study and the subsequent evolution of the financial markets to have occurred earlier.

The context

Shortly after his graduation from secondary school in Caen, northern France, first his father then mother died in quick succession. Bachelier was forced to assume control of his father's wine business. It was early 1889, Bachelier was not yet 19 years of age. Having achieved the degree of *baccalauréat es sciences*, his education was unavoidably interrupted—unfortunate for a young mathematical mind beginning to get to grips with the great theoretical debates of the day. Those interested in math applied themselves either to mathematical physics or geometry. Probability simply did not exist as an area of study or research.

While his contemporaries such as Emile Borel continued on their path through academia, Bachelier was instead tending to the needs of the family business and assuming responsibility for his sister. The Bachelier family was very much part of the community in Le Havre, where his father both worked as a wine merchant and was the Vice Consul of Venezuela. His mother was the daughter of a local banker who also dabbled in poetry. The loss of his mother and father dragged Bachelier away from his formal education, and thrust him into the practical considerations of the market. At the helm of *Bachelier Fils* he had his first interactions with the Paris Bourse, in particular the heavily traded *Rente*s contracts. This practical education was to prove beneficial in terms of his ability to be original but the lack of the formal training that should have occurred at this point was to prove a burden he would carry till the end of his life.

After his baccalauréat Bachelier should have gone on to a Lycée for two years. The grounding in science required for his later choice of career could only be acquired here. Bachelier's passion for science was innate, and something that he did not neglect even given his practical responsibilities. But being self-taught meant that there were inevitably gaps in his knowledge. Once at the Sorbonne he struggled; although he eventually did succeed at each level it was only by a very narrow margin. To have passed at all, however, is not to be undervalued—the standards were painfully high—but had he had the benefit of the Lycée education doubtless he would have performed far better at an earlier stage. This unorthodox aspect of his curriculum vitae was perceived as a handicap, and it was largely due to this that he was never offered a university chair.

After three years as a businessman, somewhat against his will, the problems were further compounded when Bachelier was drafted into the French army, to serve for one year. By the time he was demobbed he was 22. Entirely self-taught he then entered the Sorbonne to sit for his Bachelor of Science which he attained in 1895 after much struggle. This was followed in 1897 by a certificate in mathematical physics.

Since the death of Pierre Laplace in 1827 probability theory had been in the doldrums, it was not deemed worthy of any serious effort by mathematicians. As a recognized discipline it dates from after 1925. Laplace had introduced various ideas and techniques in his book *Théorie analytique des probabilités*. Prior to Laplace probability theory had predominantly been stimulated by and directed toward the mathematical analysis of gambling.

Although the theory of errors, actuarial mathematics and statistical mechanics arose during the nineteenth century, the difficulty in arriving at a definition of probability that was precise enough for use in mathematics yet comprehensive enough for application to a range of phenomena meant that it was largely left to people looking for a quick franc. Due to this, mathematicians largely abandoned the study of probability for nearly a century.

Bachelier's thesis could not have applied itself to questions that were more out of vogue, distinguishing two types of probabilities with reference to operations on the exchange. The thesis could not be considered a probability thesis and instead was slotted into the mathematical physics pigeonhole. But it wasn't about physics; it was about the stock exchange—rather a trivial pursuit in the minds of the intellectual elite. The paper is remarkable in that although the reasoning did not display the sort of rigor one expects, the intuitive aspect is largely correct. There was no mathematical foundation for probability in the late nineteenth century, yet here we find the origins of mathematical finance, stochastic calculus, the theory of Brownian motion, Markov processes, diffusion processes and so on. But what the thesis board at the Sorbonne saw was a paper lacking in technical rigor. A paper dealing with finance! Talking about probabilities! There was only one person who was willing to give papers that refused simple categorization any time: Henri Poincaré, the greatest mathematician at the turn of the century.

The thesis

Let us now consider the remarkable list of precedents Bachelier set with this single paper.

- Initiated the theory of Brownian motion. Predating Einstein's Nobel winning paper by five years.
- The paper represented the first attempt to mathematically model price movements and evaluate contingent claims in financial markets.
- His formulation that the speculator's expectation is zero was seminal, implicitly creating the axiom that the market evaluates assets using a martingale measure.
- Bachelier proposed the further hypothesis that price evolves as a continuous Markov process, homogeneous in time and space. Markov did not begin work on this until 1906.
- He showed that the density of the one-dimensional distributions of this process satisfies relations now known as the Chapman–Kolmogorov equation. He noted that the Gaussian density with linearly increasing variance solved this equation. He also arrived at this result by considering the price process as a limit of random walks.
- Bachelier observed the family of distribution functions of the process satisfies the heat equation. Probability diffuses. This model is applied to calculate various option prices.
- With path dependent options in mind Bachelier calculated the probability that Brownian motion does not exceed a fixed level. He found the distribution of the supremum of Brownian motion.

Poincaré was impressed. Despite the unorthodox subject matter and almost cavalier approach to rigorous proof he wrote a highly positive report. 'The hypothesis . . . that the probability of a deviation from the current market price is independent of the absolute value of this price. The hypothesis holds provided that the deviations are not too large. The author states this clearly, without perhaps emphasizing it as much as he ought to. It is enough that he has stated it explicitly so that his reasoning is correct.'

Bachelier was fortunate that Poincaré made such a careful study, and already drawn to the paper by the application of the heat equation and development of ideas of trajectories. In the future Bachelier would suffer for his assumptions. Poincaré and the committee awarded the distinction '*Honorable*' which apparently was the highest distinction that could be conferred on a paper that was not purely mathematical and lacked some of the rigor required for the higher awards.

Despite receiving a positive assessment of his primary thesis from the pre-eminent mathematical mind of the age, Bachelier fell into relative obscurity soon after achieving the doctorate. Although *Théorie de la spéculation* was published in the most respected journal of the time, other factors militated towards Bachelier receding into the shadows. His second thesis, on the movements of a sphere in fluid, was nowhere near as innovative as his first. Furthermore, his resumé did not fit with the demands of the upper echelons of academia. Bachelier must have been employed in something with regard to the Bourse in order to survive; there are records of his having received scholarships to continue his studies. Poincaré continued to provide a benevolent force in helping to keep Bachelier's head above water, but the way into the establishment was proving remarkably unyielding.

As mentioned Bachelier's academic efforts were primarily funded by scholarships. Many of these were granted by Emile Borel—the founder of the modern theory of functions—less than a year younger than Bachelier, but already well ensconced in the establishment. He was the youngest person to have ever received a chair at the Sorbonne at 25, he had a prominent position on the Council of the Faculty of Sciences. Borel would report favorably upon Bachelier's applications for funding, but despite a deep interest in probability he took no interest in Bachelier. Amongst the reasons for this were Bachelier's subject matter, Bachelier didn't fit the necessary criteria to be 'one of us'. Borel enjoyed a rarefied view of proceedings, he did not see the point of hyperasymptotic diffusion, Bachelier's obsession after 1900. So on the one hand he could afford to be magnanimous and keep Bachelier's efforts alive, but on the other he could completely ignore the results of those efforts and block Bachelier's progress simply through ignorance. And that was very much the way of things for Bachelier until 1909.

The rentes

How the French Revolution created a massively liquid market in bonds

Louis Bachelier predominantly concerned himself with the *Rentes*, perpetual government bonds traded on the Paris Bourse in the nineteenth and early twentieth centuries. These instruments came about after landowners, who had fled France during the Revolution, returned to discover that their holdings had been sold as national property. As recompense the French state took a loan of a billion francs in 1815.

Interest was paid on this by the state (but the capital was never paid), thus creating a perpetual bond — the success of this initial issue led to further new offerings along the same lines. At the time Bachelier wrote his thesis the nominal capital of this debt was around 26 billion francs against an annual national budget of 4 billion francs.

Rentes provided the dispossessed noblemen with a quarterly income, and the certificates were passed on through families and actively traded. The market was very active, with price fluctuations happening in continuous time. Prices did not generally deviate much from par value, absolute price changes were roughly the same as relative price changes, an average standard deviation over the year of 2.5% was normal. *Rentes* ceased to exist in 1914, when the franc collapsed with the outbreak of war.

In that year Bachelier lectured at the Sorbonne as what was then known as a 'free professor', he only began to receive payment for his work in 1913. He presented on probability calculus with applications in the financial markets. In 1912 he published his lecture notes as the book *Calcul des probabilités*, the first work to surpass Laplace. This was followed in 1914 by *Le jeu, la chance et le hasard* which reiterated his argument that continuous distributions best describe random phenomena. His systematic use of the concept of continuity in probabilistic modeling and not simplification through the use of discrete distributions was, he felt, his major contribution to science. The book was an enormous success, selling over 6000 copies. That year was the first to look truly positive for Bachelier's career since his thesis.

In that year the Council of the Paris University actually supported a move to make Bachelier's appointment permanent and paid. But the Great War erupted and destroyed this plan. Once again Bachelier was drafted, as a private, and served in the army until the end of 1918. World War I was a destroyer of illusions, it left elites on shaky ground. Bachelier survived the war, inevitably other mathematicians, with tenure, did not. An unfortunate irony of the war was that it left Bachelier with more opportunity. He was able to lecture, first at Besançon, then Dijon and Rennes.

The misunderstanding

It was 1926 when Bachelier passed through perhaps the most trying period of his life. A position had become available at Dijon, where he had taught between 1922 and 1925. But the application was turned down and he was blackballed by the university due to an unfavorable report from Paul Lévy.

In his thesis Bachelier progressed from a 'drunkard's' random walk with n discrete steps each of size d in time t, to a continuous distribution of where the drunkard might be at time t. He realized that there had to be a relationship between n and d, with d proportional to $(t/n)^{1/2}$ in order for the limit process to work as n increased.

In a paper of 1913 (*Les probabilités cinematiques et dynamiques)* Bachelier had shown that if a random walk on the y-axis is represented as a graph in time the path was such that the tangent of the path angle d divided by t/n became increasingly large as n increased. The paths in the time graph got more and more vertical with increasing n but the resulting distribution of where the drunkard might be became increasingly regular. Lévy had been asked by Maurice Gevrey, then professor of Mechanics at Dijon, to comment on this single page from Bachelier's

Market appreciation

A few well-known names provide their view on Bachelier's influence

'Regarding Bachelier, his life is the beginning of a wonderful fascination of academics, mostly physicists and later the probabilists, with the stock market. There is a new book coming out by Emanuel Derman on his experiences moving from physics to Wall Street. Although I haven't read it, I have heard excerpts, and I suspect that we will learn that many of the motivations of Bachelier are alive and well today. Also, I think due credit has to be given to the academic peers, Poincaré of course in Bachelier's case, for allowing people to work on such "bastard" topics.' **Alan Lewis**

'To me one important aspect of the Bachelier story is that, as far as I understand it, he himself never had the satisfaction of hearing his work praised to the skies or getting associated material rewards — and nothing we can do or say now can change that. It's sad and sobering, and makes one reflect on the importance of thinking for yourself and of respecting others' independent thoughts. On a professional level, like everyone else, I think the main point is the power of a very simple idea — a random Bachelier-motion walk, completely defined by its drift and volatility — to explain vast realms of financial behavior. Where I differ is that I believe that one simple 'friendly amendment' to the Bachelier worldview — namely that fundamentals occasionally shift from one Bachelier-type process to another without clear signal — explains nearly all of the puzzles that ordinary Bachelier processes can't explain. Again I regard this as a friendly amendment — it increases my respect for Bachelier's core approach.' **Kent Osband**

1913 paper. Bachelier, as we have noted often took shortcuts in his work, and both Gevrey and Lévy merely scanned the paper without recourse to looking at the original thesis. Lévy concluded that Bachelier had made a mistake by making the tangent of the path constant. The problem in Bachelier's style was that he often skipped details that were obvious to him but perhaps not to others. He only made these details explicit in his thesis.

Bachelier was distraught, but Lévy was for a long time unrepentant. Despite the fact that Bachelier had by then published numerous papers on the subject of probability, plus the two books, Lévy was totally oblivious to him. In his memoirs of 1970 *Quelques aspects de la pensée d'un mathematicien*, Lévy reports the following revelation. '. . .in 1931, when reading Kolmogorov's fundamental paper, I came to "der Bacheliers Fall". I looked up Bachelier's works, and saw that this error, which is repeated everywhere, does not prevent him from obtaining results that would have been correct if only, instead of $v =$ constant, he had written $v = c\tau^{-1/2}$, and that prior to Einstein and prior to Wiener, he happens to have seen some important properties of the so-called Wiener or Wiener–Lévy function, namely the diffusion equation and the distribution of $\max_{0<\tau<t} X(t)$.'

Bachelier rediscovered

It was a chance rediscovery in Chicago that brought Bachelier's worldview to light

Bachelier's work influenced the influential, there is no doubt about this. In *An Introduction to Probability Theory and its Applications*, William Feller writes,

'Credit for discovering the connection between random walks and diffusion is due principally to L. Bachelier. His work is frequently of a heuristic nature, but he derived many new results. Kolmogorov's theory of stochastic processes of Markov type is based largely on Bachelier's ideas.'

Kolmogorov and Doob both referenced Bachelier, whilst Ito has acknowledged that Bachelier's work influenced him more than Wiener's. However, outside of these seekers of knowledge, due to the elitist tendencies amongst the Paris academics Bachelier's works were neglected and overlooked.

A rebirth occurred in the early 1950s when a mathematical statistician at Chicago University, Jimmy Savage, chanced upon a copy of Bachelier in the library. He was so excited by what he'd found he immediately sent off memos to around twenty academics across the States. One of the recipients of this memo was the eminent economist Paul Samuelson, who was already familiar with Bachelier's name but this time looked up the work in the MIT library.

Sixty-five years after Bachelier had assumed that prices must fluctuate randomly, Samuelson published proof that properly anticipated prices must fluctuate randomly in *Industrial Management Review* in 1965, the paper that, along with one by Fama, introduced the Efficient Markets Hypothesis. Samuelson also reiterates the assumption that prices follow a martingale — which Bachelier implicitly assumed. He later explained that Bachelier's model failed to ensure that stock prices were always positive — however, geometric Brownian motion, the cornerstone of the Black–Scholes–Merton view — solves this problem.

After this catastrophe, for which in later years Lévy did apologize, Bachelier finally was offered a permanent post at Besançon in 1927. He retired ten years later and died in 1946 aged 76.

Brilliant Bachelier

Treasurer of the Bachelier Finance Society, and a winner at the first Wilmott Awards, Peter Carr *(Bloomberg LP) looks at the works of Bachelier.*

In my humble opinion, Bachelier wrote the best doctoral dissertation in the history of both probability theory and finance. It is well known that his 1900 dissertation introduced efficient markets, Brownian motion, and option pricing theory to the world. It is less well known that in this dissertation, one can find informal discussions of stopping times, martingales, and arbitrage. One can also find a formal derivation of the probability density function (PDF) for the first passage time of Brownian motion to a given level. Although confined to the context of driftless Brownian motion, one also finds the first appearance of the Kolmogorov backward equation, the Chapman–Kolmogorov equation, and the notion of implied volatility. There are also many other seminal ideas in his dissertation, as this note will endeavor to elucidate.

To guide his assumptions and to reach his conclusions, Bachelier assumed as his fundamental pricing principle that 'the mathematical expectation of the speculator is zero'. In other words, asset prices should be such that average ex-post profit from any asset position should be neither positive nor negative. While this notion is routinely challenged in modern financial economics, I am not personally convinced that it is invalid in a setting of zero net supply and infinite trading opportunities. At any rate, we now know that this principle is equivalent to no arbitrage provided that the mathematical expectation in question is risk-neutral, i.e. the probabilities used to calculate it are implied from contemporaneous market prices, rather than assessed historically or subjectively.

Assuming zero interest rates for simplicity, Bachelier further assumed that at each future time $t > 0$, the spot price S_t of the underlying asset is normally distributed with constant known mean S_0 and increasing variance a^2t. To compactly express Bachelier's option pricing formulas, let $m_t \equiv S_t - K$ denote the moneyness at t when valuing a call of strike K and let $m_t \equiv K - S_t$ denote the moneyness at t when valuing a put of strike K.

Then a straightforward calculation yields that the probability that the final moneyness m_T exceeds a given level m is given by:

$$Pr(m_T > m) = N\left(\frac{m_0 - m}{s}\right), \tag{1}$$

where

$$N(d) \equiv \int_{-\infty}^{d} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz$$

denotes the standard normal distribution function and $s \equiv a\sqrt{T}$ is the standard deviation of S_T.

The payoff on a European option maturing at T is $m_T^+ = \int_0^\infty 1(m_T > m)\,dm$ and so integrating (1) on m from 0 to ∞ gives the expected payoff, which from Bachelier's fundamental pricing principle is the initial option price. After some straightforward manipulations, the theoretical initial value Θ of a T-maturity European option is given by:

$$\Theta(m_0, s) = m_0 N\left(\frac{m_0}{s}\right) + sN'\left(\frac{m_0}{s}\right). \tag{2}$$

Note that this formula depends only on the mean m_0 and the standard deviation s of m_T. To interpret this formula, note from setting $m = 0$ in (1) that $N(m_0/s)$ appearing in (2) is just the probability of finishing in-the-money. Although Bachelier was not focused on hedging, it is also the absolute value of the option's delta

$$\Delta(m_0, s) \equiv \frac{\partial}{\partial S_0}\Theta(m_0, s).$$

Differentiating again w.r.t. S_0 implies that

$$\frac{1}{s}N'(m_0/s)$$

is both the option's gamma

$$\Gamma(m_0, s) \equiv \frac{\partial}{\partial S_0}\Delta(m_0, s)$$

and the probability density function (PDF) of S_T at K. Hence, the initial option value satisfies:

$$\begin{aligned}\Theta(m_0, s) &= m_0 Pr\{m_T > 0\} + s^2 \frac{Pr\{S_T \in dK\}}{dK} \\ &= m_0|\Delta(m_0, s)| + s^2\Gamma(m_0, s).\end{aligned} \tag{3}$$

To obtain the needed s input, Bachelier notes that his formula (2) simplifies dramatically for an option which is initially at-the-money (ATM). Setting $m_0 = 0$ in (2) yields:

$$A \equiv \Theta(0, s) = \frac{s}{\sqrt{2\pi}}, \tag{4}$$

$$\text{since } N'(0) = \frac{1}{\sqrt{2\pi}}.$$

Inverting this relation gives an exact expression relating the Bachelier implied volatility to the market price of the ATM option:

$$s = \sqrt{2\pi}A. \tag{5}$$

Substituting (5) in (2) shows that the value of an away-from-the-money option depends only on the observable prices of the underlying asset and an ATM option:

$$\Theta = m_0 N\left(\frac{m_0}{\sqrt{2\pi}A}\right) + \sqrt{2\pi}AN'\left(\frac{m_0}{\sqrt{2\pi}A}\right). \tag{6}$$

Notice that the instantaneous volatility a and the total standard deviation s are both irrelevant given these two market prices.

Bachelier also calculates the probability that the buyer of an ATM option makes a profit. This is the probability that m_T exceeds A when $m_0 = 0$.

Substituting (4) in (1) implies that:

$$Pr(m_T > A) = N\left(-\frac{1}{\sqrt{2\pi}}\right) \approx .345. \tag{7}$$

Amazingly, this probability is a pure number independent of the underlying price, the instantaneous volatility, the option maturity, and even the price paid for the ATM option.

As a final demonstration of the convenience of the Bachelier model, suppose that barrier options were available in Bachelier's time. A down-and-out call (DOC) is just a standard call that knocks out if the underlying crosses a pre-specified barrier $H < S_0$. Although Bachelier determines the survival probability, these path-dependent options are more easily valued as follows. Suppose that we form a simple portfolio which is long a standard call of strike K and short a standard put of strike $2H - K$. Notice that the average of the two strikes is the barrier H.

If the underlying avoids the barrier by maturity, then the portfolio provides the call payoff, since the put necessarily finishes out-of-the-money.

If the underlying touches or crosses the barrier before maturity, then at the first passage time, the underlying is at the barrier due to the continuity of its price process. As a result, the long call and the short put have the same moneyness of $H - K$ at this first passage time. Since (2) applies to both put and call values, the short put has the same absolute value as the long call. It follows that the portfolio value vanishes at the first passage time, irrespective of the latter's exact realization before T.

We conclude from Bachelier's fundamental pricing principle that the initial value of the DOC is just given by the difference between the initial call premium and the initial put premium. As in (6), the normal instantaneous volatility a and the total standard deviation s are irrelevant given these prices.

Up to this point, we have been exploring Bachelier's model which assumes that the normal instantaneous volatility is constant over time, even if we didn't need to know its exact value in the presence of market prices. Using the modern language of stochastic differential equations invented later by Ito, Bachelier assumed that the stock price obeys:

$$S_t = S_0 + aW_t, \quad t \in [0, T], \tag{8}$$

where a is the constant normal volatility and W is standard Brownian motion.

In fact, all of Bachelier's results require only minor modification if the normal instantaneous volatility is time-varying, i.e.:

$$S_t = S_0 + \int_0^T a_t \, dW_t, \quad t \in [0, T]. \tag{9}$$

If the normal volatility a_t is just a deterministic function of time, then Bachelier's option pricing formula in (2) remains valid with s replaced by $s_T \equiv \sqrt{\int_0^T a_t \, dt}$. If a_t is more generally a continuous time stochastic process, then s_T becomes random, so further changes are needed to obtain deterministic option prices. For the results which follow, we will require no knowledge of the stochastic process $\{a_t, t \in [0, T]\}$ other than that it evolves independently of the spot price. In particular, the normal volatility a_t can jump and it need not be Markov in itself and time.

If we condition on the instantaneous volatility path to T, then s_T is again deterministic. As a consequence, the theoretical value of an option under stochastic (independent instantaneous normal) volatility is given by:

$$\Theta^{sv}(m_0) = \int_0^\infty \Theta(m_0, s) q(s)\, ds, \tag{10}$$

where the Bachelier pricing formula $\Theta(m_0, s)$ is defined in (2) and $q(s)$ is the probability density of s_T at s. If we treat the LHS as the observable market prices of T-maturity European options of all initial moneyness levels m_0, then one can interpret (10) as an integral equation with kernel $\Theta(m_0, s)$ multiplying the unknown function $q(s)$. Using integral transforms, then one can analytically invert (10) for the PDF $q(s)$ of s_T. This density can be used to consistently price European-style derivatives on the realized standard deviation.

In particular, setting $m_0 = 0$ in (10) implies:

$$A^{sv} \equiv \Theta^{sv}(0) = \int_0^\infty \frac{s}{\sqrt{2\pi}} q(s)\, ds = \frac{1}{\sqrt{2\pi}} E s_T. \tag{11}$$

Consider a forward contract on realized standard deviation with final payoff $s_T - s_0$ where s_0 is initially chosen so that the contract has zero cost to enter.

From Bachelier's fundamental pricing principle, the forward price of the realized standard deviation is:

$$s_0 = E s_T = \sqrt{2\pi} A^{sv}, \tag{12}$$

from (11). From (5), this is just the Bachelier implied volatility obtained from the ATM option of the same maturity as the volatility swap. What could be simpler? Using a conditioning argument, the probability that an initially ATM option finishes in-the-money is still the pure number

$$N\left(-\frac{1}{\sqrt{2\pi}}\right) \approx 0.345.$$

Similarly, the DOC is still priced by the difference between the initial premium of the call struck at K and the put struck at $2H - K$. There are yet other results for less liquid exotics such as passport options and lookback options. In the interests of brevity, let's save those for another time.

For more information on Bachelier, I recommend Mandelbrot (1987), Taqqu (2001), and Schachermayer (2003). Mandelbrot wrote an entry on Bachelier on page 86 of the Finance Volume of the *New Palgrave Dictionary of Economics*. Taqqu expanded on his conversations with a historian of probability theory named Bernard Bru in (2001). Schachermayer does a wonderful job surveying Bachelier's thesis in a survey article on derivatives pricing, which can be downloaded from `www.fam.tuwien.ac.at/~wschach/pubs`. Of course, nothing beats reading Bachelier's dissertation, which was originally written in French. His dissertation was first translated into English in 1964 and this translation appears appropriately as Chapter 1 in Cootner (1964). This classic book of readings is available at `www.riskbooks.com`, where an introduction by Andy Lo can be freely downloaded. It is part of the folklore of economics that Bachelier's works were lost to the world until being rediscovered in the 1950s.

Supposedly, Bachelier's insights had been rediscovered by the time his work was found (see e.g. Bernstein (1992), Boyle and Boyle (2001)). However, Lévy, Kolmogorov, and Ito all knew of Bachelier's work well before the 1950s. Merton brought Ito calculus to finance and used it to derive the hedging argument at the heart of this trillion-dollar industry. So one has to wonder how the world might be different if Bachelier's dissertation really did disappear more than a century ago.

Since its rediscovery, Bachelier's work on option pricing has been faulted for giving positive probabilities to negative prices. Underlying this criticism is a somewhat baseless philosophy that there is a true stochastic process governing asset prices and the goal of research is to find it. A more pragmatic view is that there is a current industry practice and the goal of research is to improve it. On the question of whether Bachelier's dissertation succeeded on this dimension in 1900, the answer can be found in the (English translation of the) text that ends it: 'It is evident that the present theory resolves the majority of problems in the study of speculation by the calculus of probability.'

REFERENCES

■ Bachelier, L. (1900) Théorie de la spéculation. *Annales des Sciences de L'Ecole Normale Superieure*, Paris, 17, 3, 21–86.
■ Bernstein, P. (1992) *Capital Ideas*. Free Press, New York.
■ Boyle, P. and Boyle, F. (2001) *Derivatives: The Tools that Changed Finance*. Risk Books, London.
■ Cootner, P., ed. (1964) *The Random Character of Stock Prices*. Cambridge: MIT.
■ Mandelbrot, B. (1987) *New Palgrave Dictionary of Economics, Finance Volume*. Eatwell, Milgate, and Newman (eds) W.W. Norton, New York.
■ Schachermayer, W. (2003) Introduction to the mathematics of financial markets. *Lecture Notes in Mathematics 1816–Lectures on Probability Theory and Statistics*, Pierre Bernard, ed. Springer Verlag, Heidelberg, 111–177.
■ Taqqu, M. (2001) Bachelier and his times: a conversation with Bernard Bru. *Finance and Stochastics*, 5, 1, 3–32.

3
The Collector: Know Your Weapon—Part 1*

Espen Gaarder Haug

Trading options is War! For an option trader a pricing or hedging formula is just like a weapon. A soldier who has perfected her pistol shooting[1] can beat a guy with a machine gun that doesn't know how to handle it. Similarly, an option trader knowing the ins and outs of the Black–Scholes–Merton (BSM) formula can beat a trader using a state-of-the-art stochastic volatility model. It comes down to two rules, just as in war. Rule number one: Know your weapon. Rule number two: Don't forget rule number one. In my ten+ years as a trader I have seen many a BSD[2] option trader getting confused with what the computer was spitting out. They often thought something was wrong with their computer system/implementation. Nothing was wrong, however, except their knowledge of their weapon. Before you move on to a more complex weapon (like a stochastic volatility model) you should make sure you know conventional equipment inside-out. In this installment I will not show the nerdy quants how to come up with the BSM formula using some new fancy mathematics—you don't need to know how to melt metal to use a gun. Neither is it a guideline on how to trade. It is meant rather like a short manual of how your weapon works in extreme situations. Real war (trading)—the pain, the pleasure, the adrenaline of winning and losing millions of dollars—can only be learned through real action. Now, the manual:

BSD trader Soldier, welcome to our trading team, this is your first day and I will instruct you about the Black–Scholes weapon.

New hired trader Hah, my Professor taught me probability theory, Itô calculus, and Malliavin calculus! I know everything about stochastic calculus and how to come up with the Black–Scholes formula.

BSD trader Soldier, you may know how to construct it, but that doesn't mean you know a shit about how it operates!

New hired trader I have used it for real trading. Before my Ph.D. I was a market maker in stock options for a year. Besides, why do you call me soldier? I was hired as an option trader.

*For this chapter I got a lot of ideas from the Wilmott forum. Thanks! And especially thanks to Alexander Adamchuk, Jørgen Haug, Hicham Mouline and James Ward for useful comments.

BSD trader Soldier, you have not been in real war. In real war you often end up in extreme situations. That's when you need to know your weapon.

New hired trader I have read *Liar's Poker*, Hull's book, *Wilmott on Wilmott*, Taleb's *Dynamic Hedging*, Haug's formula collection. I know about Delta Bleed and all that stuff. I don't think you can tell me much more. I have even read *Fooled by Ran*...

BSD trader SHUT UP, SOLDIER! If you want to survive the first six months on this trading floor you better listen to me. On this team we don't allow any mistakes. We are warriors, trained in war!

New hired trader Yes, Sir!

BSD trader Good, let's move on to our business. Today I will teach you the basics of the Black–Scholes weapon.

1 Background on the BSM formula

Let me refresh your memory of the BSM formula

$$c = Se^{(b-r)T} N(d_1) - Xe^{-rT} N(d_2)$$
$$p = Xe^{-rT} N(-d_2) - Se^{(b-r)T} N(-d_1),$$

where

$$d_1 = \frac{\ln(S/X) + (b + \sigma^2/2)T}{\sigma\sqrt{T}},$$
$$d_2 = d_1 - \sigma\sqrt{T},$$

and

S	=	stock price
X	=	strike price of option
r	=	risk-free interest rate
b	=	cost-of-carry rate of holding the underlying security
T	=	time to expiration in years
σ	=	volatility of the relative price change of the underlying stock price
N(x)	=	the cumulative normal distribution function

2 Delta Greeks

2.1 Delta

As you know, the delta is the option's sensitivity to small movements in the underlying asset price.

$$\Delta_{call} = \frac{\partial c}{\partial S} = e^{(b-r)T} N(d_1) > 0$$
$$\Delta_{put} = \frac{\partial p}{\partial S} = -e^{(b-r)T} N(-d_1) < 0$$

Delta higher than unity I have many times over the years been contacted by confused commodity traders claiming something is wrong with their BSM implementation. What they observed was a spot delta higher than one.

As we get deep-in-the-money $N(d_1)$ approaches one, but it never gets higher than one (since it's a cumulative probability function). For a European call option on a non-dividend-paying stock the delta is equal to $N(d_1)$, so the delta can never go higher than one. For other options the delta term will be multiplied by $e^{(b-r)T}$. If this term is larger than one and we are deep-in-the-money we can get deltas considerably higher than one. This occurs if the cost-of-carry is larger than the interest rate, or if interest rates are negative. Figure 1 illustrates the delta of a call option. As expected the delta reaches above unity when time to maturity is large and the option is deep-in-the-money.

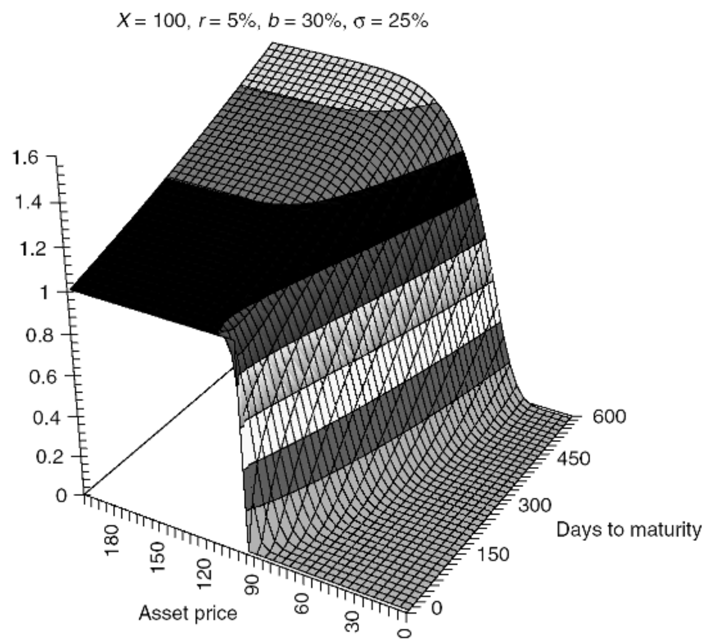


Figure 1: Spot delta

2.2 Delta mirror strikes and asset

For a put and call to have the same absolute delta value we can find the delta symmetric strikes as

$$X_p = \frac{S^2}{X_c} e^{(2b+\sigma^2)T}, \qquad X_c = \frac{S^2}{X_p} e^{(2b+\sigma^2)T}.$$

That is

$$\Delta_c(S, X_c, T, r, b, \sigma) = -\Delta_p\left(S, \frac{S^2}{X_c} e^{(2b+\sigma^2)T}, T, r, b, \sigma\right)$$

where X_c is the strike of the call and X_p is the strike of a put. These relationships are useful to determine strikes for delta neutral option strategies, especially for strangles, straddles, and butterflies. The weakness of this approach is that it works only for a symmetric volatility smile. In practice, however, you often only need an approximately delta neutral strangle. Moreover, volatility smiles are often more or less symmetric in the currency markets.

In the special case of a straddle-symmetric-delta-strike, described by Wystrup (1999), the formulas above can be simplified further to

$$X_c = X_p = Se^{(b+\sigma^2/2)T}.$$

Related to this relationship is the straddle-symmetric-asset-price. Given the identical strikes for a put and call, for what asset price will they have the same absolute delta value? The answer is

$$S = Xe^{(-b-\sigma^2/2)T}.$$

At this strike and delta-symmetric-asset-price the delta is $\frac{e^{(b-r)T}}{2}$ for a call, and $-\frac{e^{(b-r)T}}{2}$ for a put. Only for options on non-dividend paying stocks[3] ($b = r$) can we simultaneously have an absolute delta of 0.5 (50%) for a put and a call. Interestingly, the delta symmetric strike is also the strike given the asset price where the gamma and vega are at their maximums, *ceteris paribus*. The maximal gamma and vega, as well as the delta neutral strikes, are not at-the-money forward as I have noticed has been assumed by many traders. Moreover, an in-the-money put can naturally have absolute delta lower than 50% while an out-of-the-money call can have delta higher than 50%.

For an option that is at the straddle-symmetric-delta-strike the generalized BSM formula can be simplified to

$$c = \frac{Se^{(b-r)T}}{2} - Xe^{-rT}N(-\sigma\sqrt{T}),$$

and

$$p = Xe^{-rT}N(\sigma\sqrt{T}) - \frac{Se^{(b-r)T}}{2}.$$

At this point the option value will not change based on changes in cost of carry (dividend yield etc.). This is as expected as we have to adjust the strike accordingly.

2.3 Strike from delta

In several OTC (over-the-counter) markets options are quoted by delta rather than strike. This is a common quotation method in, for example, the OTC currency options market, where one typically asks for a delta and expects the sales person to return a price (in terms of volatility or pips) as well as the strike, given a spot reference. In these cases one needs to find the strike that corresponds to a given delta. Several option software systems solve this numerically using Newton–Raphson or bisection. This is actually not necessary, however. Using an inverted cumulative normal distribution $N^{-1}(\cdot)$ the strike can be derived from the delta analytically as described by Wystrup (1999). For a call option

$$X_c = S\exp[-N^{-1}(\Delta_c e^{(r-b)T})\sigma\sqrt{T} + (b+\sigma^2/2)T],$$

and for a put we have

$$X_p = S\exp[N^{-1}(-\Delta_p e^{(r-b)T})\sigma\sqrt{T} + (b+\sigma^2/2)T].$$

To get a robust and accurate implementation of this formula it is necessary to use an accurate approximation of the inverse cumulative normal distribution. I have used the algorithm of Moro (1995) with good results.

2.4 DdeltaDvol and DvegaDspot

DdeltaDvol: $\frac{\partial\Delta}{\partial\sigma}$ which mathematically is the same as DvegaDspot: $\frac{\partial\text{vega}}{\partial S}$, a.k.a. Vanna,[4] shows approximately how much your delta will change for a small change in the volatility, as well as how much your vega will change with a small change in the asset price:

$$\begin{aligned}\text{DdeltaDvol} &= \frac{\partial c}{\partial S\partial\sigma} = \frac{-e^{(b-r)T}d_2}{\sigma}n(d_1)\\ &= \frac{\partial p}{\partial S\partial\sigma} = \frac{e^{(b-r)T}d_2}{\sigma}n(d_1),\end{aligned}$$

where $n(x)$ is the standard normal density

$$n(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

One fine day in the dealing room my risk manager asked me to get into his office. He asked me why I had a big outright position in some stock index futures—I was supposed to do 'arbitrage trading'. That was strange as I believed I was delta neutral: long call options hedged with short index futures. I knew the options I had were far out-of-the-money and that their DdeltaDvol was very high. So I immediately asked what volatility the risk management used to calculate their delta. As expected, the volatility in the risk-management-system was considerably below the market and again was leading to a very low delta for the options. This example is just to illustrate how a feeling of your DdeltaDvol can be useful. If you have a high DdeltaDvol the volatility you use to compute your deltas becomes very important.[5]

Figure 2 illustrates the DdeltaDvol. As we can see the DdeltaDvol can assume positive and negative values. DdeltaDvol attains its maximal value at

$$S_L = Xe^{-bT-\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2},$$

and attains its minimal value when

$$S_U = Xe^{-bT+\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2}.$$

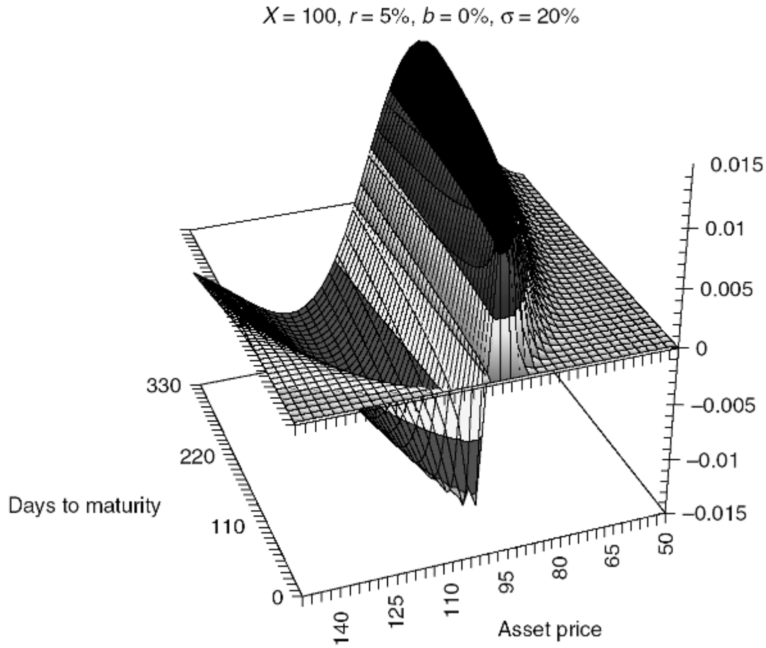


Figure 2: DdeltaDvol

Similarly, given the asset price, options with strikes X_L have maximum negative DdeltaDvol at

$$X_L = Se^{bT-\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2},$$

and options with strike X_U have maximum positive DdeltaDvol when

$$X_U = Se^{bT+\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2}.$$

One naturally can ask if these measures have any meaning. Black and Scholes assumed constant volatility, or at most deterministic volatility. Despite being theoretically inconsistent it might well be a good approximation. How good an approximation it is I leave up to you to find out or discuss at the Wilmott forum, www.wilmott.com. For more practical information about DvegaDspot or Vanna see Webb (1999).

2.5 DdeltaDtime, Charm

DdeltatDtime, a.k.a. Charm (Garman 1992) or Delta Bleed (a term used in the excellent book by Taleb 1997), is delta's sensitivity to changes in time,

$$-\frac{\partial \Delta_c}{\partial T} = -e^{(b-r)T}\left[n(d_1)\left(\frac{b}{\sigma\sqrt{T}} - \frac{d_2}{2T}\right)\right.$$
$$\left. + (b-r)N(d_1)\right] \leq\geq 0,$$

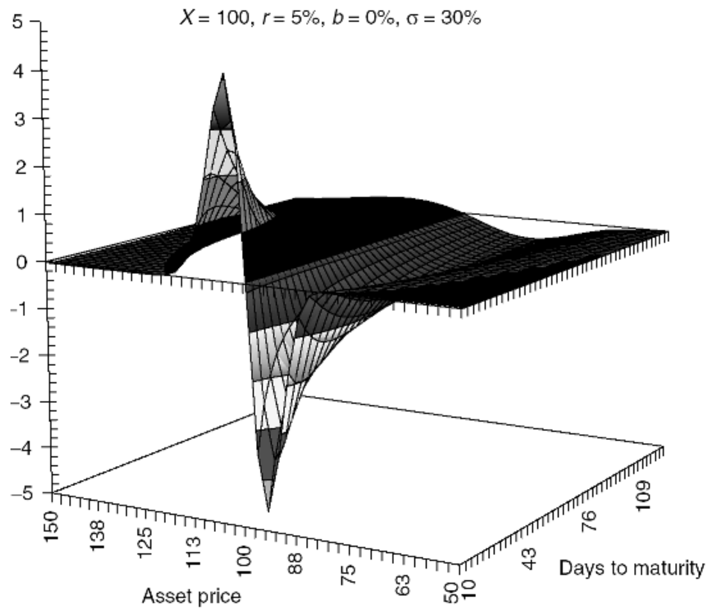


Figure 3: Charm

and

$$-\frac{\partial \Delta_p}{\partial T} = -e^{(b-r)T}\left[n(d_1)\left(\frac{b}{\sigma\sqrt{T}} - \frac{d_2}{2T}\right)\right.$$
$$\left. - (b-r)N(-d_1)\right] \leq\geq 0.$$

This Greek gives an indication of what happens with delta when we move closer to maturity. Figure 3 illustrates the Charm value for different values of the underlying asset and different times to maturity.

As Nassim Taleb points out one can have both forward and backward bleed. He also points out the importance of taking into account how expected changes in volatility over the given time period will affect delta. I am sure most readers already have his book in their collection (if not, order it now!). I will therefore not repeat all his excellent points here.

All partial derivatives with respect to time have the advantage over other Greeks in that we know which direction time will move. Moreover, we know that time moves at a constant rate. This is in contrast, for example, to the spot price, volatility, or interest rate.[6]

2.6 Elasticity

The elasticity of an option, a.k.a. the option leverage, omega, or lambda, is the sensitivity in percent to a percent movement in the underlying asset price. It is given by

$$\Lambda_{call} = \Delta_{call}\frac{S}{call} = e^{(b-r)T}N(d_1)\frac{S}{call} > 1$$
$$\Lambda_{put} = \Delta_{put}\frac{S}{put} = -e^{(b-r)T}N(-d_1)\frac{S}{put} < 0$$

The option's elasticity is a useful measure on its own, as well as to estimate the volatility, beta, and expected return from an option.

Option volatility The option volatility σ_o can be approximated using the option elasticity. The volatility of an option over a short period of time is approximately equal to the elasticity of the option multiplied by the stock volatility σ.[7]

$$\sigma_o \approx \sigma|\Lambda|.$$

Option beta The elasticity is also useful to compute the option's beta. If asset prices follow geometric Brownian motions the continuous-time capital asset pricing model of Merton (1971) holds. Expected asset returns then satisfy the CAPM equation

$$E[\text{return}] = r + E[r_m - r]\beta_i$$

where r is the risk-free rate, r_m is the return on the market portfolio, and β_i is the beta of the asset. To determine the expected return of an option we need the option's beta. The beta of a call is given by see for instance Jarrow and Rudd (1983)

$$\beta_c = \frac{S}{call}\Delta_c \beta_S,$$

where β_S is the underlying stock beta. For a put the beta is

$$\beta_p = \frac{S}{put}\Delta_p \beta_S.$$

For a beta neutral option strategy the expected return should be the same as the risk-free rate (at least in theory).

Option Sharpe ratios As the leverage does not change the Sharpe (1966) ratio, the Sharpe ratio of an option will be the same as that of the underlying stock,

$$\frac{\mu_o - r}{\sigma_o} = \frac{\mu_S - r}{\sigma}$$

where μ_o is the return of the option, and μ_S is the return of the underlying stock. This relationship indicates the limited usefulness of the Sharpe ratio as a risk-return measure for options (?). Shorting a lot of deep out-of-the-money options will likely give you a 'nice' Sharpe ratio, but you are almost guaranteed to blow up one day (with probability one if you live long enough). An interesting question here is should you use the same volatility for all strikes? For instance deep-out-of-the-money stock options typically trade for much higher implied volatility than at-the-money options. Using the volatility smile when computing Sharpe ratios for deep out-of-the-money options can also possibly make the Sharpe ratio work better for options. McDonald (2002) offers a more detailed discussion of option Sharpe ratios.

3 Gamma Greeks

3.1 Gamma

Gamma is the delta's sensitivity to small movements in the underlying asset price. Gamma is identical for put and call options, *ceteris paribus*, and is given by

$$\Gamma_{call,put} = \frac{\partial^2 c}{\partial S^2} = \frac{\partial^2 p}{\partial S^2} = \frac{n(d_1)e^{(b-r)T}}{S\sigma\sqrt{T}} > 0$$

This is the standard gamma measure given in most textbooks (Haug 1997, Hull 2000, Wilmott 2000).

3.2 Maximal gamma and the illusions of risk

One day in the trading room of a former employer of mine, one of the BSD traders suddenly got worried over his gamma. He had a long dated deep-out-of-the-money call. The stock price had been falling, and the further the out-of-the-money the option went the lower the gamma he expected. As with many option traders he believed the gamma was largest approximately at-the-money-forward. Looking at his Bloomberg screen, however, the further out-of-the-money the call went the higher his gamma got. Another BSD was coming over, and they both tried to come up with an explanation for this. Was there something wrong with Bloomberg?

In my own home-built system I was often playing around with three- and four-dimensional charts of the option Greeks, and I already knew that gamma doesn't attain its maximum at-the-money-forward (four dimensions? a dynamic three-dimensional graph). I didn't know exactly where it attained its maximum, however. Instead of joining the BSD discussion, I did a few computations in Mathematica. A few minutes later, after double checking my calculations, I handed over an equation to the BSD traders showing exactly where the BSM gamma would be at its maximum.

How good is the rule of thumb that gamma is largest for at-the-money or at-the-money-forward options? Given a strike price and time to maturity, the gamma is at maximum when the asset price is[8]

$$S_{\overline{\Gamma}} = Xe^{(-b-3\sigma^2/2)T}.$$

Given the asset price and time to maturity, gamma is maximal when the strike is

$$X_{\overline{\Gamma}} = Se^{(b+\sigma^2/2)T}.$$

Confused option traders are bad enough, confused risk management is a pain in the behind. Several large investment firms impose risk limits on how much gamma you can have. In the equity market it is common to use the standard textbook approach to compute gamma, as shown above. Putting on a long-term call (put) option that later is deep-out-of-the-money (in-the-money) can blow up the gamma risk limits, even if you actually have close to zero gamma risk. The high gamma risk for long dated deep-out-of-the-money options typically is only an illusion. This illusion of risk can be avoided by looking at percentage changes in the underlying asset (gammaP), as is typically done for FX options.

Saddle gamma Alexander (Sasha) Adamchuk was the first to make me aware of the fact that gamma has a saddle point.[9] The saddle point is attained for the time[10]

$$T_S = \frac{1}{2(\sigma^2 + 2b - r)},$$

and at asset price

$$S_{\overline{\Gamma}} = Xe^{(-b-3\sigma^2/2)T_S}.$$

The gamma at this point is given by

$$\Gamma_S = \Gamma(S_{\overline{\Gamma}}, T_S) = \frac{\sqrt{\frac{e}{\pi}}\sqrt{\frac{2b-r}{\sigma^2}+1}}{X}$$

Many traders get surprised by this feature of gamma—that gamma is not necessarily decreasing with longer time to maturity. The maximum gamma for a given strike price is first decreasing until the saddle gamma point, then increasing again, given that we follow the edge of the maximal gamma asset price.

Figure 4 shows the saddle gamma. The saddle point is between the two gamma 'mountain' tops. This graph also illustrates one of the big limitations in the textbook gamma definition, which is actually in use by many option systems and traders. The gamma increases dramatically when we have long time to maturity and the asset price is close to zero. How can the gamma be larger than for an option closer to at-the-money? Is the real gamma risk that big? No, this is in most

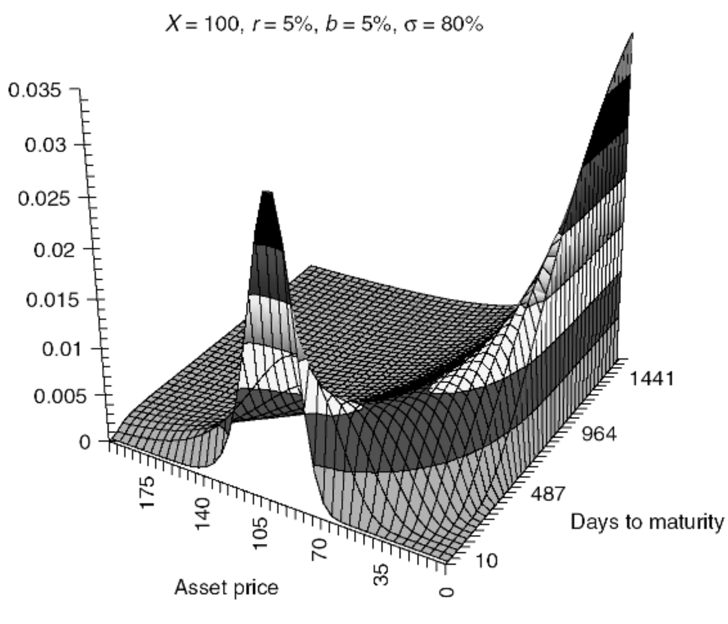


Figure 4: Saddle gamma

cases simply an illusion, due to the above unmotivated definition of gamma. Gamma is typically defined as the change in delta for a one unit change in the asset price. When the asset price is close to zero a one unit change is naturally enormous in percent of the asset price. In this case it is also highly unlikely that the asset price will increase by one dollar in an instant. In other words, the gamma measurement should be reformulated, as many option systems already have done. It makes far more sense to look at percentage moves in the underlying than unit moves. To compare gamma risk from different underlyings one should also adjust for the volatility in the underlying.

3.3 GammaP

As already mentioned, there are several problems with the traditional gamma definition. A better measure is to look at percentage changes in delta for percentage changes in the underlying,[11] for example a 1% point change in underlying. With this definition we get for both puts and calls (gamma percent)

$$\Gamma_P = \frac{S\Gamma}{100} > 0. \tag{3.1}$$

GammaP attains a maximum at an asset price of

$$S_{\overline{\Gamma}_P} = Xe^{(-b-\sigma^2/2)T}.$$

Alternatively, given the asset price the maximal Γ_P occurs at strike

$$X_{\overline{\Gamma}_P} = Se^{(b+\sigma^2/2)T}.$$

Interestingly, this is also where we have a straddle symmetric asset price as well as maximal gamma. This implies that a delta neutral straddle has maximal Γ_P. In most circumstances going from measuring the gamma risk as Γ_P instead of gamma we avoid the illusion of a high gamma risk when the option is far out-of-the-money and the asset price is low. Figure 5 is an illustration of this, using the same parameters as in Figure 4.

If the cost-of-carry is very high it is still possible to experience high Γ_P for deep-out-of-the-money call options with a low asset price and a long time to maturity. This is because a high cost-of-carry can make the ratio of a deep-out-of-the money call to the spot close to the at-the-money-forward. At this point the spot-delta will be close to 50% and so the Γ_P will be large. This is not an illusion of gamma risk, but a reality. Figure 6 shows Γ_P with the same parameters as in Figure 5, with cost-of-carry of 60%.

To makes things even more complicated, the high Γ_P we can have for deep-out-of-the-money calls (in-the-money puts) is the only case when we are dealing with spot gammaP (change in spot delta). We can avoid this by looking at future/forward gammaP. However, if you hedge with spot, then spot gammaP is the relevant metric. Only if you hedge with the future/forward the forward gammaP is the relevant metric. The forward gammaP we have when the cost-of-carry is set to zero, and the underlying asset is the futures price.

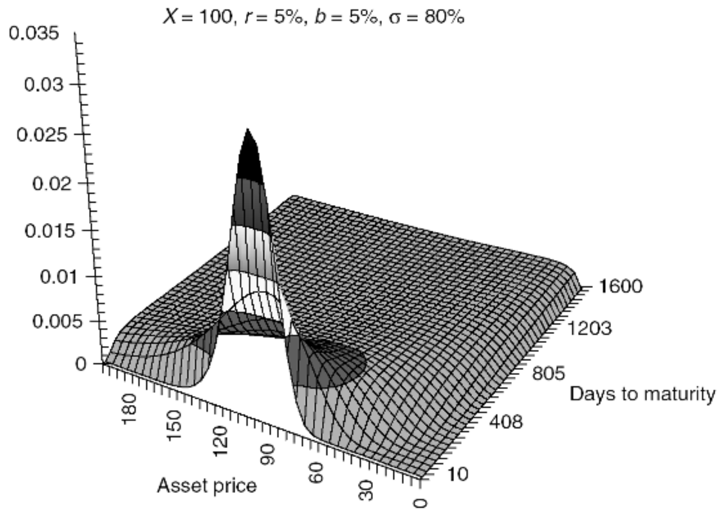


Figure 5: GammaP

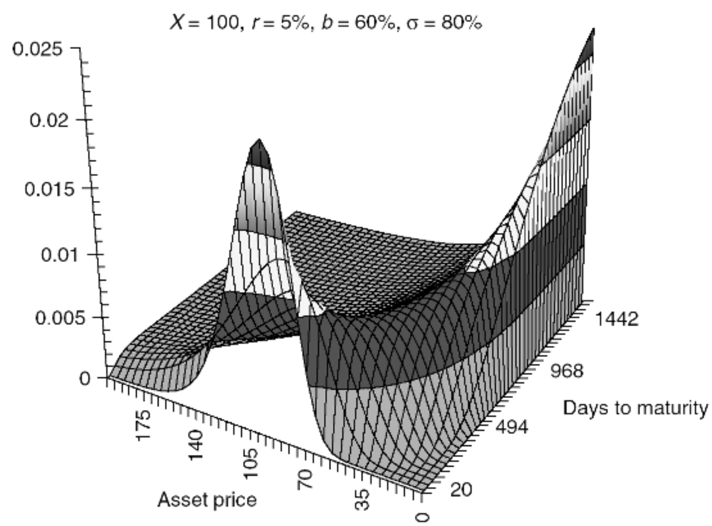


Figure 6: Saddle gammaP

3.4 Gamma symmetry

Given the same strike the gamma is identical for both put and call options. Although this equality breaks down when the strikes differ, there is a useful put and call gamma symmetry. The put-call symmetry of Bates (1991) and Carr and Bowie (1994) is given by

$$c(S, X, T, r, b, \sigma) = \frac{X}{Se^{bT}} p\left(S, \frac{(Se^{bT})^2}{X}, T, r, b, \sigma\right).$$

This put-call value symmetry yields the gamma symmetry; however, the gamma symmetry is more general as it is independent of whether the option is a put or call, for example it could be two calls, two puts, or a put and a call.

$$\Gamma(S, X, T, r, b, \sigma) = \frac{X}{Se^{bT}}\Gamma(S, \frac{(Se^{bT})^2}{X}, T, r, b, \sigma).$$

Interestingly, the put-call symmetry also gives us vega and cost-of-carry symmetries, and in the case of zero cost-of-carry also theta and rho symmetry. Delta symmetry, however, is not obtained.

3.5 DgammaDvol, Zomma

DgammaDvol, a.k.a. Zomma, is the sensitivity of gamma with respect to changes in implied volatility. In my view, DgammaDvol is one of the more important Greeks for options trading. It is given by

$$\begin{aligned}\text{DgammaDvol}_{call,put} &= \frac{\partial\Gamma}{\partial\sigma} \\ &= \Gamma\left(\frac{d_1 d_2 - 1}{\sigma}\right) \leq\geq 0.\end{aligned}$$

For the gammaP we have DgammaPDvol

$$\text{DgammaPDvol}_{call,put} = \Gamma_P\left(\frac{d_1 d_2 - 1}{\sigma}\right) \leq\geq 0$$

where Γ is the textbook Gamma of the option.

For practical purposes, where one typically wants to look at DgammaDvol for a one unit volatility change, for example from 30% to 31%, one should divide the DGammaDVol by 100. Moreover, DgammaDvol and DgammaPDvol are negative for asset prices between S_L and S_U and positive outside this interval, where

$$\begin{aligned}S_L &= Xe^{-bT-\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2}, \\ S_U &= Xe^{-bT+\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2}.\end{aligned}$$

For a given asset price the DgammaDvol and DgammaPDvol are negative for strikes between

$$X_L = Se^{bT-\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2}$$

and

$$X_U = Se^{bT+\sigma\sqrt{T}\sqrt{4+T\sigma^2}/2},$$

and positive for strikes above X_U or below X_L, *ceteris paribus*. In practice, these points will change with other variables and parameters. These levels should, therefore, be considered good approximations at best.

In general you want positive DgammaDvol—especially if you don't need to pay for it (flat volatility smile). In this respect DgammaDvol actually offers a lot of intuition for how stochastic volatility should affect the BSM values (?). Figure 7 illustrates this point. The DgammaDvol is positive for deep-out-of-the-money options, outside the S_L and S_U interval. For at-the-money options and slightly in- or out-of-the-money options the DgammaDvol is negative. If the volatility is stochastic and uncorrelated with the asset price then this offers a good indication for which strikes you should use higher/lower volatility when deciding on your volatility smile. In the case of volatility correlated with the asset price this naturally becomes more complicated.

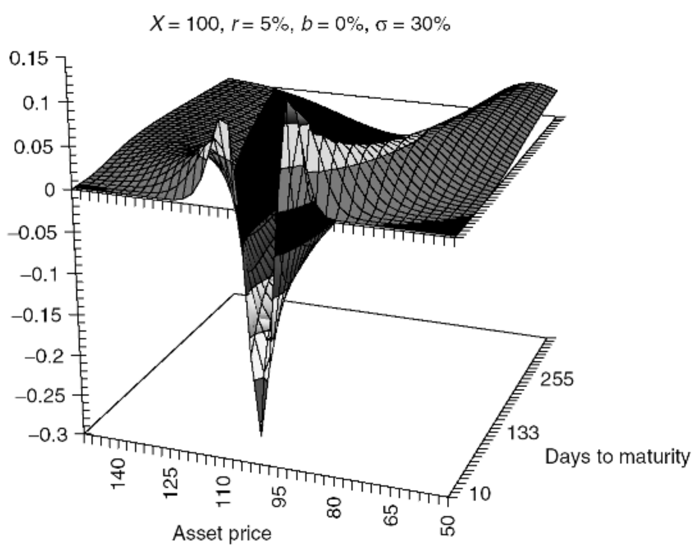


Figure 7: DgammaDvol

3.6 DgammaDspot, Speed

I have heard rumors about how being on speed can help see higher dimensions that are ignored or hidden for most people. It should be of little surprise that in the world of options the third derivative of the option price with respect to spot, known as Speed, is ignored by most people. Judging from his book, Nassim Taleb is also a fan of higher-order Greeks. There he mentions Greeks of up to seventh order.

Speed was probably first mentioned by Garman (1992),[12] for the generalized BSM formula we get

$$\frac{\partial^3 c}{\partial S^3} = -\frac{\Gamma\left(1 + \frac{d_1}{\sigma\sqrt{T}}\right)}{S}.$$

A high Speed value indicates that the gamma is very sensitive to moves in the underlying asset. Academics typically claim that third- or higher-order 'Greeks' are of no use. For an option trader, on the other hand, it can definitely make sense to have a sense of an option's Speed. Interestingly, Speed is used by Fouque *et al*. (2000) as a part of a stochastic volatility model adjustment. More to the point, Speed is useful when gamma is at its maximum with respect to the asset price. Figure 8 shows the graph of Speed with respect to the asset price and time to maturity.

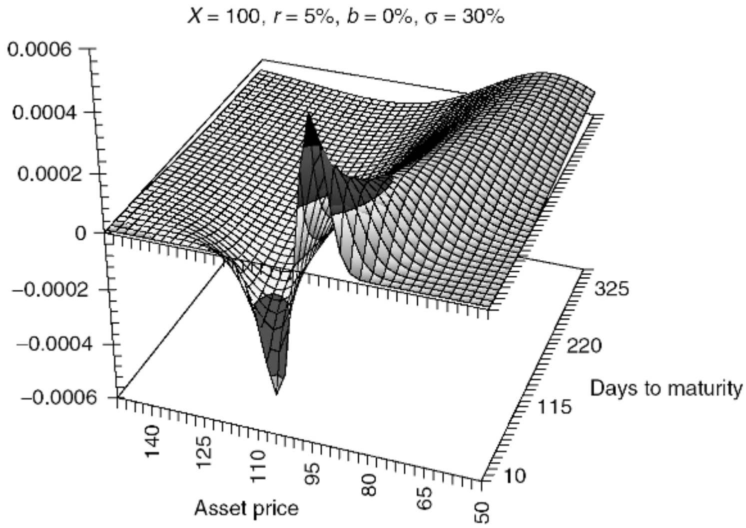


Figure 8: Speed

For Γ_P we have an even simpler expression for Speed, that is SpeedP (Speed for percentage gamma)

$$\text{SpeedP} = -\Gamma \frac{d_1}{100\sigma\sqrt{T}}.$$

3.7 DgammaDtime, Colour

The change in gamma with respect to small changes in time to maturity, DGammaDtime, a.k.a. GammaTheta or Colour (Garman 1992), is given by (assuming we get closer to maturity):

$$-\frac{\partial \Gamma}{\partial T} = \frac{e^{(b-r)T} n(d_1)}{S\sigma\sqrt{T}} \left(r - b + \frac{bd_1}{\sigma\sqrt{T}} + \frac{1 - d_1 d_2}{2T} \right)$$
$$= \Gamma \left(r - b + \frac{bd_1}{\sigma\sqrt{T}} + \frac{1 - d_1 d_2}{2T} \right) \leq\geq 0.$$

Divide by 365 to get the sensitivity for a one day move. In practice one typically also takes into account the expected change in volatility with respect to time. If you, for example on Friday are wondering how your gamma will be on Monday you typically will also assume a higher implied volatility on Monday morning. For Γ_P we have DgammaPDtime

$$-\frac{\partial \Gamma_P}{\partial T} = \Gamma_P \left(r - b + \frac{bd_1}{\sigma\sqrt{T}} + \frac{1 - d_1 d_2}{2T} \right) \leq\geq 0.$$

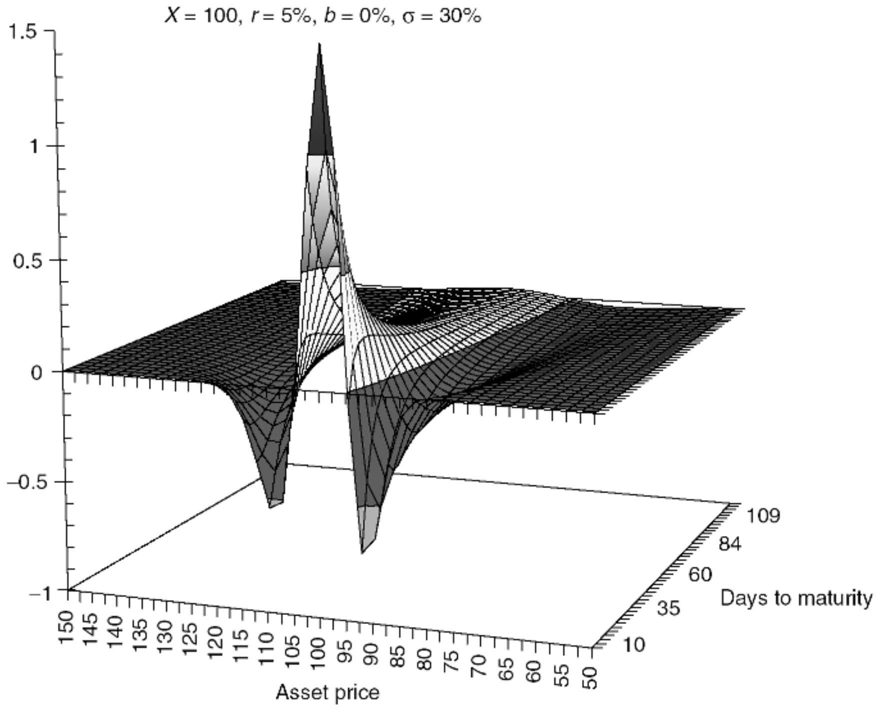


Figure 9: DgammaDtime

Figure 9 illustrates the DgammaDtime of an option with respect to varying asset price and time to maturity.

4 Numerical Greeks

So far we have looked only at analytical Greeks. A frequently used alternative is to use numerical Greeks. Most first-order partial derivatives can be computed by the two-sided finite difference method

$$\frac{c(S+\Delta S, X, T, r, b, \sigma) - c(S-\Delta S, X, T, r, b, \sigma)}{2\Delta S}.$$

In the case of derivatives with respect to time, we know what direction time will move and it is more accurate (for what is happening in the 'real' world) to use a backward derivative

$$\Theta \approx \frac{c(S, X, T, r, b, \sigma) - c(S, X, T-\Delta T, r, b, \sigma)}{\Delta T}.$$

Numerical Greeks have several advantages over analytical ones. If for instance we have a sticky delta volatility smile then we can also change the volatilities accordingly when calculating the numerical delta. (We have a sticky delta volatility smile when the shape of the volatility smile sticks to the deltas but not to the strike; in other words the volatility for a given strike will move as the underlying moves.)

$$\Delta_c \approx \frac{c(S+\Delta S, X, T, r, b, \sigma_1) - c(S-\Delta S, X, T, r, b, \sigma_2)}{2\Delta S}.$$

Numerical Greeks are moreover model independent, while the analytical Greeks presented above are specific to the BSM model.

For gamma and other second derivatives, $\frac{\partial^2 f}{\partial x^2}$, for example DvegaDvol, we can use the central finite difference method

$$\Gamma \approx \frac{c(S+\Delta S, \ldots) - 2c(S, \ldots) + c(S-\Delta S, \ldots)}{\Delta S^2}.$$

If you are very close to maturity (a few hours) and you are approximately at-the-money the analytical gamma can approach infinity, which is naturally an illusion of your real risk. The reason is simply that analytical partial derivatives are accurate only for infinite small changes, while in practice one sees only discrete changes. The numerical gamma solves this problem and offers a more accurate gamma in these cases. This is particularly true when it comes to barrier options (Taleb 1997).

For Speed and other third-order derivatives, $\frac{\partial^3 f}{\partial x^3}$, we can, for example, use the following approximation

$$\begin{aligned}\text{Speed} \approx \frac{1}{\Delta S^3}&[c(S+2\Delta S, \ldots) - 3c(S+\Delta S, \ldots)\\ &+3c(S, \ldots) - c(S-\Delta S, \ldots)].\end{aligned}$$

What about mixed derivatives, $\frac{\partial f}{\partial x \partial y}$, for example DdeltaDvol and Charm. This can be calculated numerical by DdeltaDvol

$$\begin{aligned}&\approx \frac{1}{4\Delta S\Delta\sigma}[c(S+\Delta S, \ldots, \sigma+\Delta\sigma)\\ &-c(S+\Delta S, \ldots, \sigma-\Delta\sigma) - c(S-\Delta S, \ldots, \sigma+\Delta\sigma)\\ &+c(S-\Delta S, \ldots, \sigma-\Delta\sigma)].\end{aligned}$$

In the case of DdeltaDvol one would 'typically' divide it by 100 to get the 'right' notation.

End Part 1

BSD trader That is enough for today soldier.
New hired trader Sir, I learned a few things today. Can I start trading now?
BSD trader We don't let fresh soldiers play around with ammunition (capital) before they know the basics of a conventional weapon like the Black–Scholes formula.
New hired trader Understood, Sir!

BSD trader Next time I will tell you about Vega-kappa, probability Greeks and some other stuff. Until then you are dismissed! Now bring me a double cheeseburger with a lot of fries!

New hired trader Yes, Sir!

FOOTNOTES & REFERENCES

1. The author was among the best pistol shooters in Norway.
2. If you don't know the meaning of this expression, BSD, then it's high time you read Michael Lewis' *Liar's Poker*.
3. And naturally also for commodity options in the special case where cost-of-carry equals r.
4. I wrote about the importance of this Greek variable back in 1992. It was my second paper about options, and my first written in English. Well, it got rejected. What could I expect? Most people totally ignored DdeltaDvol at that time and the paper has collected dust since then.
5. An important question naturally is what volatility you should use to compute your deltas. I will not give you an answer to that here, but there has been discussions on this topic at www.wilmott.com.
6. This is true only because everybody trading options at Mother Earth moves at about the same speed, and is affected by approximately the same gravity. In the future, with huge space stations moving with speeds significant to that of the speed of light, this will no longer hold true. See Haug (2003a and b) for some possible consequences.
7. This approximation is used by Bensoussan *et al.* (1995) for an approximate valuation of compound options.
8. Rubinstein (1990) indicates in a footnote that this maximum curvature point possibly can explain why the greatest demand for calls tend to be just slightly out-of-the-money.
9. Described by Adamchuk at the Wilmott forum www.wilmott.com February 6, 2002, http://www.wilmott.com/310/messageview.cfm?catid=4&threadid=664&highlight_key=y&keyword1=vanna and even earlier on his page http://finmath.com/Chicago/NAFTCORP/Saddle_Gamma.html
10. It is worth mentioning that T_S must be larger than zero for the gamma to have a saddle point, that means b must be larger then $\frac{r-\sigma^2}{2}$, and r must be smaller than $\sigma^2 + 2b$.
11. Wystrup (1999) also describes how this redefinition of gamma removes the dependence on the spot level S. He calls it 'traders gamma'. This measure of gamma has for a long time been popular, particularly in the FX market, but is still absent in options textbooks.
12. However, he was too 'lazy' to give us the formula so I had to do the boring derivation myself.

■ Bates, D. S. (1991) The crash of '87: Was it expected? The evidence from options markets. *Journal of Finance*, 46(3), 1009–1044.
■ Bensoussan, A., Crouhy, M. and Galai, D. (1995) Black–Scholes approximation of warrant prices. *Advances in Futures and Options Research*, 8, 1–14.
■ Black, F. (1976) The pricing of commodity contracts. *Journal of Financial Economics*, 3, 167–179.

■ Black, F. and Scholes M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–654.
■ Carr, P. and Bowie, J. (1994) Static simplicity. *Risk Magazine*, 7(8).
■ Fouque, J., Papanicolaou, G. and Sircar, K. R. (2000) *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press.
■ Garman, M. (1992) Charm school. *Risk Magazine*, 5(7), 53–56.
■ Haug, E. G. (1997) *The Complete Guide to Option Pricing Formulas*. McGraw-Hill, New York.
■ Haug, E. G. (2003a) Frozen time arbitrage. *Wilmott Magazine*, January.
■ Haug, E. G. (2003b) The special and general relativity's implications on mathematical finance. *Working paper*, January.
■ Hull, J. (2000) *Option, Futures, and Other Derivatives*. Prentice Hall.
■ Jarrow, R. and Rudd, A. (1983) *Option Pricing*. Irwin.
■ Lewis, M. (1992) *Liar's Poker*. Penguin.
■ McDonald, R. L. (2002) *Derivatives Markets*. Addison Wesley.
■ Merton, R. C. (1971) Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3, 373–413.
■ Merton, R. C. (1973) Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4, 141–183.
■ Moro, B. (1995) The full Monte. *Risk Magazine*, February.
■ Rubinstein, M. (1990) The super trust. *Working paper*, www.in-the-money.com.
■ Sharpe, W. (1966) Mutual fund performance. *Journal of Business*, 119–138.
■ Taleb, N. (1997) *Dynamic Hedging*. John Wiley & Sons.
■ Webb, A. (1999) The sensitivity of Vega. *Derivatives Strategy*, http://www.derivativesstrategy.com/magazine/archive/1999/1199fea1.asp, November, 16–19.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance*. John Wiley & Sons.
■ Wystrup, U. (1999) Aspects of symmetry and duality of the Black-Scholes pricing formula for European style put and call options. *Working paper, Sal. Oppenheim Jr. & Cie*.

4

The Collector: Know Your Weapon—Part 2*

Espen Gaarder Haug

BSD trader Soldier, last time I told you about delta and gamma Greeks. Today I'll enlighten you in on Vega, theta, and probability Greeks.
New trader Sir, I already know Vega.
BSD trader Soldier, if you want to speculate on an increase in implied volatility what type of options offer the most bang for the bucks?
New trader At-the-money options with long time to maturity.
BSD trader Soldier, you are possibly wrong on strikes and time! Now start with 20 push-ups while I start to tell you about Vega.
New trader Yes, Sir!

1 Refreshing notation on the BSM formula

Let me also this time refresh your memory of the Black–Scholes–Merton (BSM) formula

$$c = Se^{(b-r)T}N(d_1) - Xe^{-rT}N(d_2)$$
$$p = Xe^{-rT}N(-d_2) - Se^{(b-r)T}N(-d_1),$$

where

$$d_1 = \frac{\ln(S/X) + (b + \sigma^2/2)T}{\sigma\sqrt{T}},$$
$$d_2 = d_1 - \sigma\sqrt{T},$$

*Thanks to Jørgen Haug for useful comments.

and

$$\begin{aligned} S &= \text{asset price} \\ X &= \text{strike price} \\ r &= \text{risk-free interest rate} \\ b &= \text{cost-of-carry rate of holding the underlying security} \\ T &= \text{time to expiration in years} \\ \sigma &= \text{volatility of the relative price change of the underlying asset price} \\ N(x) &= \text{the cumulative normal distribution function} \end{aligned}$$

2 Vega Greeks

2.1 Vega

Vega,[1] also known as kappa, is the option's sensitivity to a small change in the implied volatility. Vega is equal for put and call options.

$$\text{Vega} = \frac{\partial c}{\partial \sigma} = \frac{\partial p}{\partial \sigma} = Se^{(b-r)T} n(d_1)\sqrt{T} > 0.$$

Implied volatility is often considered the market's best estimate of expected volatility for the duration of the option. It can also be interpreted as a basket of adjustments to the BSM formula, for factors that the formula doesn't take into account; demand and supply for that particular strike and maturity, stochastic volatility, jumps, and more. For instance a sudden increase in the Black–Scholes implied volatility for an out-of-the-money strike does not necessary imply that investors expect higher volatility. The increase can just as well be due to an option 'arbitrageur' expecting higher volatility of volatility.

Vega local maximum When trying to profit from moves in implied volatility it is useful to know where the option has the maximum Vega value for a given time to maturity. For a given strike price Vega attains its maximum when the asset price is

$$S = Xe^{(-b+\sigma^2/2)T}.$$

At this asset price we also have in-the-money risk neutral probability symmetry (which I come back to later). Moreover, at this asset price the generalized Black–Scholes–Merton (BSM) formula simplifies to

$$c = Se^{(b-r)T} N(\sigma\sqrt{T}) - \frac{Xe^{-rT}}{2},$$

$$p = \frac{Xe^{-rT}}{2} - Se^{(b-r)T} N(-\sigma\sqrt{T}).$$

Similarly, the strike that maximizes Vega given the asset price is

$$X = Se^{(b+\sigma^2/2)T}.$$

Vega global maximum Some years back a BSD trader called me late one evening, close to freaking out. He had shorted long-term options, which he hedged by going long short-term options. To his surprise the long-term options' Vega increased as time went by. After looking at my 3D Vega chart I confirmed that this was indeed the expected behavior. For options with long term to maturity the maximum Vega is not necessarily increasing with longer time to maturity, as many traders believe. Indeed, Vega has a global maximum at time

$$T_{\overline{V}} = \frac{1}{2r},$$

and asset price

$$S_{\overline{V}} = Xe^{(-b+\sigma^2/2)T_{\overline{V}}} = Xe^{\frac{-b+\sigma^2/2}{2r}}.$$

At this global maximum, Vega itself, described by Alexander (Sasha) Adamchuk,[2] is equal to the following simple expression

$$\text{Vega}(S_{\overline{V}}, T_{\overline{V}}) = \frac{X}{2\sqrt{re\pi}}.$$

Figure 1 shows the graph of Vega with respect to the asset price and time. The intuition behind the Vega-top (Vega-mountain) is that the effect of discounting at some point in time dominates volatility (Vega): the lower the interest rate, the lower the effect of discounting, and the higher the relative effect of volatility on the option price. As the risk-free-rate goes to zero the time for the global maximum goes to infinity, that is we will have no global maximum when the risk-free

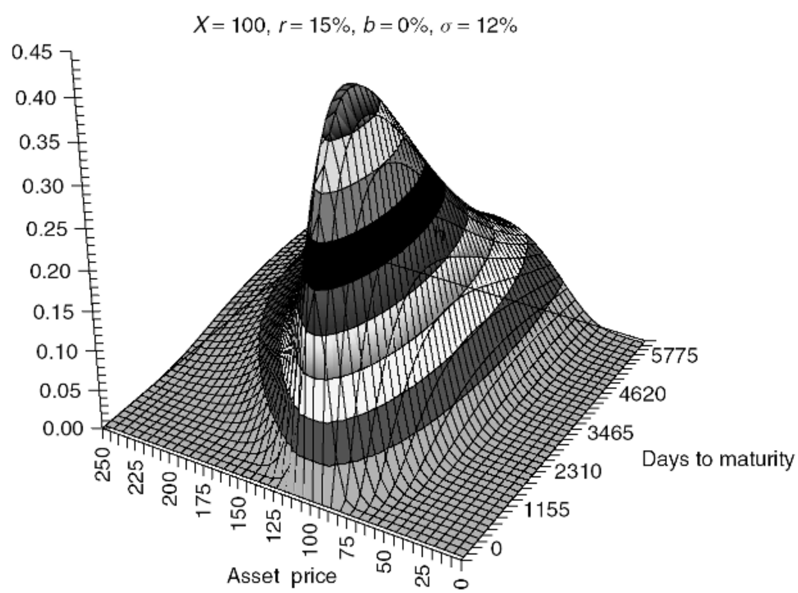


Figure 1: Vega

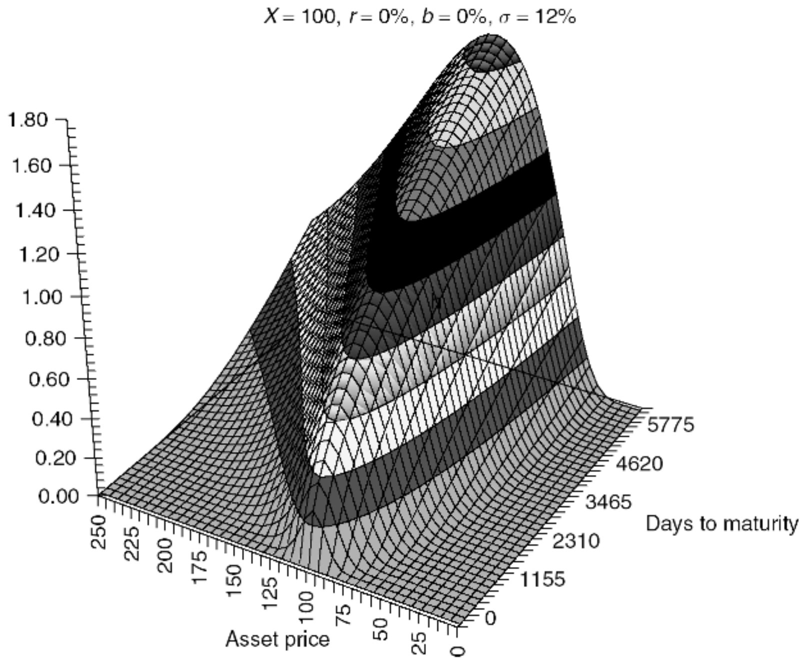


Figure 2: Vega

rate is zero. Figure 2 is the same as Figure 1 but with zero interest rate. The effect of Vega being a decreasing function of time to maturity typically kicks in only for options with very long times to maturity—unless the interest rate is very high. It is not, however, uncommon for caps and floors traders to use the Black-76 formula to compute Vegas for options with 10 to 15 years to expiration (caplets).

2.2 Vega symmetry

For options with different strikes we have the following Vega symmetry

$$\text{Vega}(S, X, T, r, b, \sigma) = \frac{X}{Se^{bT}}\text{Vega}\left(S, \frac{(Se^{bT})^2}{X}, T, r, b, \sigma\right).$$

As for the gamma symmetry, see Haug (2003), this symmetry is independent of the options being calls or puts—at least in theory.

2.3 Vega–gamma relationship

The following is a simple and useful relationship between Vega and gamma, described by Taleb (1997) amongst others:

$$\text{Vega} = \Gamma\sigma S^2 T.$$

2.4 Vega from delta

Given that we know the delta, what is the Vega? Vega and delta are related by a simple formula described by Wystrup (2002):

$$\text{Vega} = Se^{(b-r)T}\sqrt{T}n\left[N^{-1}(e^{(r-b)T}|\Delta|)\right],$$

where $N^{-1}(\cdot)$ is the inverted cumulative normal distribution, $n()$ is the normal density function, and Δ is the delta of a call or put option. Using the Vega–gamma relationship we can rewrite this relationship to express gamma as a function of the delta

$$\Gamma = \frac{e^{(b-r)T}n[N^{-1}(e^{(r-b)T}|\Delta|)]}{S\sigma\sqrt{T}}.$$

Relationships, such as the above ones, between delta and other option sensitivities are particularly useful in the FX options markets, where one often considers a particular delta rather than strike.

2.5 VegaP

The traditional textbook Vega gives the dollar change in option price for a percentage *point* change in volatility. When comparing the Vega risk of options on different assets it makes more sense to look at percentage changes in volatility. This metric can be constructed simply by multiplying the standard Vega with $\frac{\sigma}{10}$, which gives what is known as VegaP (percentage change in option price for a 10% change in volatility):

$$\text{VegaP} = \frac{\sigma}{10}Se^{(b-r)T}n(d_1)\sqrt{T} \geq 0.$$

VegaP attains its local and global maximum at the same asset price and time as for Vega. Some options systems use traditional textbook Vega, while others use VegaP.

When comparing Vegas for options with different maturities (calendar spreads) it makes more sense to look at some kind of weighted Vega, or alternatively Vega bucketing,[3] because short-term implied volatilities are typically more volatile than long-term implied volatilities. Several options systems implement some type of Vega weighting or Vega bucketing (see Haug 1993 and Taleb 1997 for more details).

2.6 Vega leverage, Vega elasticity

The percentage change in option value with respect to percentage point change in volatility is given by

$$\text{VegaLeverage}_{\text{call}} = \text{Vega}\frac{\sigma}{\text{call}} \geq 0,$$

$$\text{VegaLeverage}_{\text{put}} = \text{Vega}\frac{\sigma}{\text{put}} \geq 0.$$

The Vega elasticity is highest for out-of-the-money options. If you believe in an increase in implied volatility you will therefore get maximum bang for your bucks by buying out-of-the-money

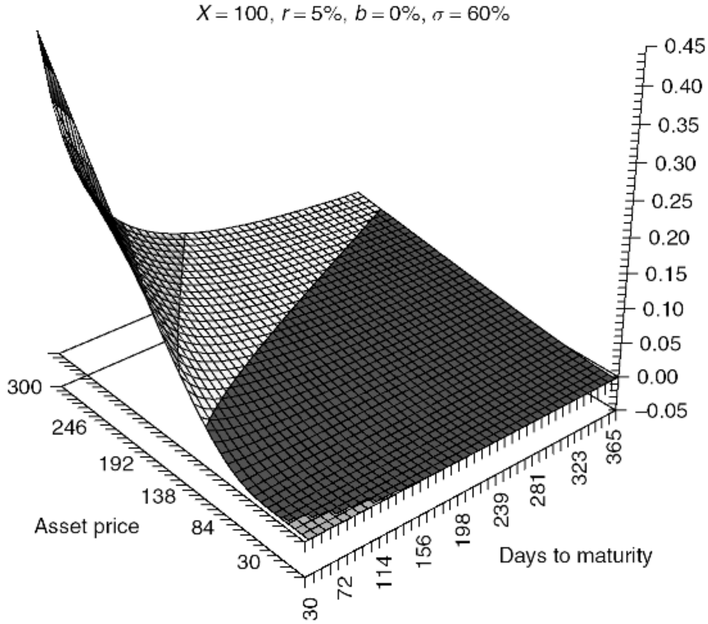


Figure 3: Vega leverage

options. Several traders I have met will typically tell you to buy at-the-money options when they want to speculate on higher implied volatility, to maximize Vega. There are several advantages to buying out-of-the-money options in such a scenario. One is the higher Vega-leverage. Another advantage is that you often also get a positive DvegaDvol (and also DgammaDvol), a measure we will have a closer look at below. The drawbacks of deep-out-of-the-money options are faster time decay (in percent of premium), and typically lower liquidity. Figure 3 illustrates the Vega leverage of a put option.

2.7 DvegaDvol, Vomma

DvegaDvol, also known as Vega convexity, Vomma (see Webb 1999), or Volga, is the sensitivity of Vega to changes in implied volatility. Together with DgammaDvol, see Haug (2003), Vomma is in my view one of the most important Greeks. DvegaDvol is given by

$$\text{DvegaDvol} = \frac{\partial^2 c}{\partial \sigma^2} = \frac{\partial^2 p}{\partial \sigma^2} = \text{Vega}\left(\frac{d_1 d_2}{\sigma}\right) \leq\geq 0.$$

For practical purposes, where one ‘typically’ wants to look at Vomma for the change of one percentage *point* in the volatility, one should divide Vomma by 10 000.

In case of DvegaPDvol we have

$$\text{DvegaPDvol} = \text{VegaP}\left(\frac{d_1 d_2}{\sigma}\right) \leq\geq 0.$$

Options far out-of-the money have the highest Vomma. More precisely given the strike price, Vomma is positive outside the interval

$$(S_L = Xe^{(-b-\sigma^2/2)T}, S_U = Xe^{(-b+\sigma^2/2)T}).$$

Given the asset price the Vomma is positive outside the interval (relevant only before conducting the trade)

$$(X_L = Se^{(b-\sigma^2/2)T}, X_U = Se^{(b+\sigma^2/2)T}).$$

If you are long options you typically want to have as high positive DvegaDvol as possible. If short options, you typically want negative DvegaDvol. Positive DvegaDvol tells you that you will earn more for every percentage point increase in volatility, and if implied volatility is falling you will lose less and less—that is, you have positive Vega convexity.

While DgammaDvol is most relevant for the volatility of the actual volatility of the underlying asset, DvegaDvol is more relevant for the volatility of the implied volatility. Although the volatility of implied volatility and the volatility of actual volatility will typically have high correlation, this is not always the case. DgammaDvol is relevant for traditional dynamic delta hedging under stochastic volatility. DvegaDvol trading has little to do with traditional dynamic delta hedging. DvegaDvol trading is a bet on changes on the price (changes in implied vol) for uncertainty in:

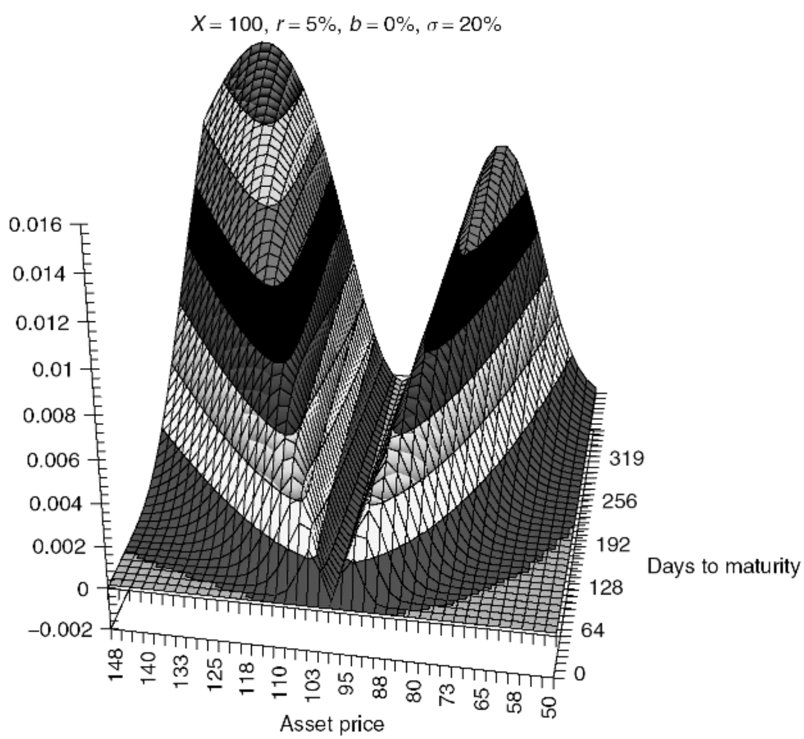


Figure 4: DvegaDvol

supply and demand, stochastic actual volatility (remember this is correlated to implied volatility), jumps and any other model risk: factors that affect the option price, but that are not taken into account in the Black–Scholes formula. A DvegaDvol trader does not necessarily need to identify the exact reason for the implied volatility to change. If you think the implied volatility will be volatile in the short term you should typically try to find options with high DvegaDvol. Figure 4 shows the graph of DvegaDvol for changes in asset price and time to maturity.

2.8 DvegaDtime

DvegaDtime is the change in Vega with respect to changes in time. Since we typically are looking at decreasing time to maturity we express this as minus the partial derivative

$$\text{DvegaDtime} = -\frac{\partial \text{Vega}}{\partial T} = \text{Vega}\left(r - b + \frac{bd_1}{\sigma\sqrt{T}} - \frac{1 + d_1 d_2}{2T}\right) \leq\geq 0$$

For practical purposes, where one 'typically' wants to express the sensitivity for a one percentage point change in volatility to a one day change in time, one should divide the DvegaDtime by 36 500, or 25 200 if you look at trading days only. Figure 5 illustrates DvegaDtime. Figure 6 shows DvegaDtime for a wider range of parameters and a lower implied volatility, as expected from Figure 1 we can see here that DvegaDtime actually can be positive.

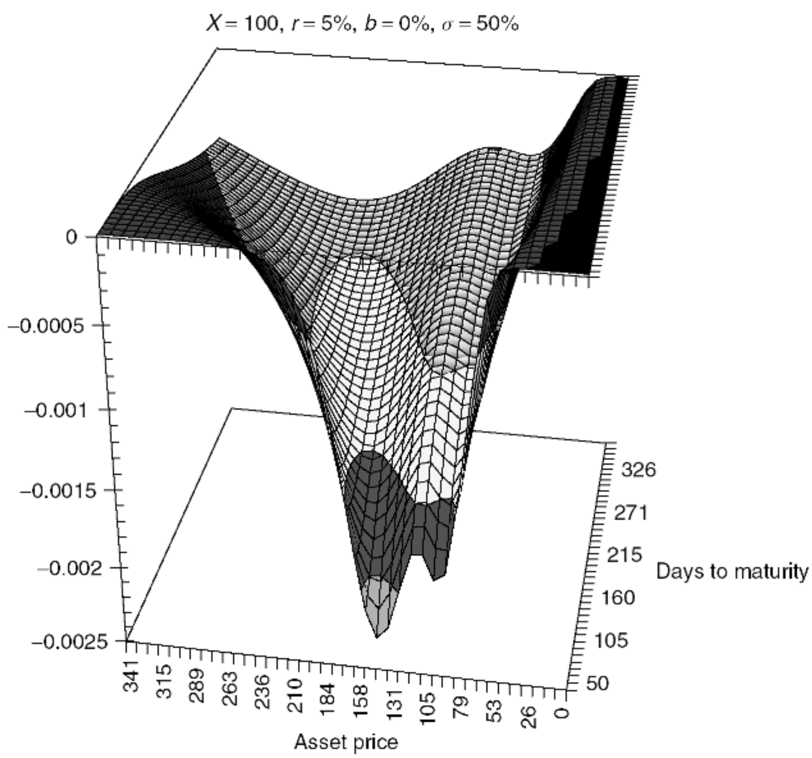


Figure 5: DvegaDtime

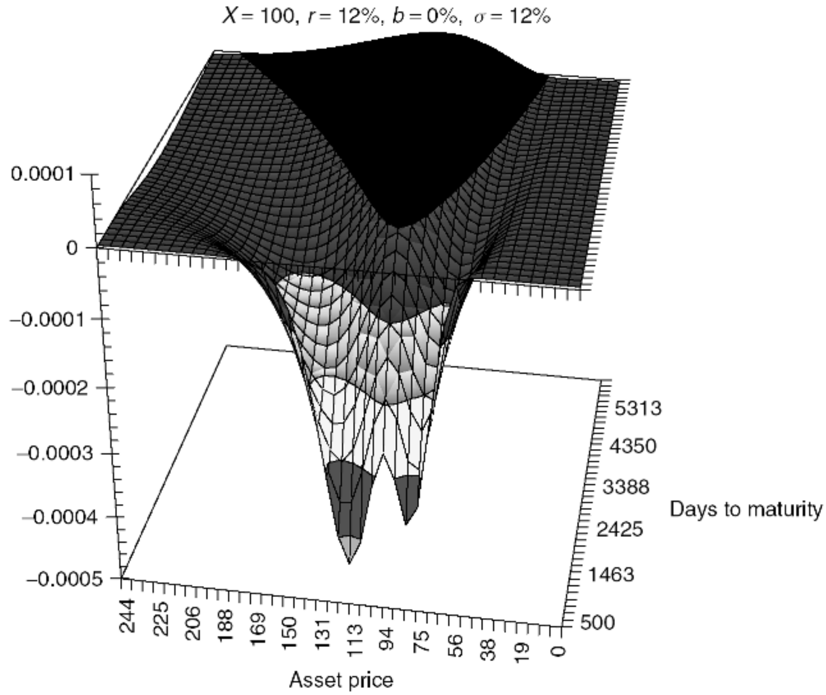


Figure 6: DvegaDtime (Vanna)

3 Theta Greeks

3.1 Theta

Theta is the option's sensitivity to a small change in time to maturity. As time to maturity decreases, it is normal to express theta as minus the partial derivative with respect to time.

Call

$$\Theta_{\text{call}} = -\frac{\partial c}{\partial T} = -\frac{Se^{(b-r)T}n(d_1)\sigma}{2\sqrt{T}} - (b-r)Se^{(b-r)T}N(d_1)$$
$$- rXe^{-rT}N(d_2) \lesseqgtr 0.$$

Put

$$\Theta_{\text{put}} = -\frac{\partial p}{\partial T} = -\frac{Se^{(b-r)T}n(d_1)\sigma}{2\sqrt{T}} + (b-r)Se^{(b-r)T}N(-d_1)$$
$$+ rXe^{-rT}N(-d_2) \lesseqgtr 0.$$

Drift-less theta In practice it is often also of interest to know the drift-less theta, θ, which measures time decay without taking into account the drift of the underlying or discounting. In other words the drift-less theta isolates the effect time-decay has on uncertainty, assuming unchanged volatility. The uncertainty or volatilities effect on the option consists of time and volatility. In that case we have

$$\theta_{\text{call}} = \theta_{\text{put}} = \theta = -\frac{Sn(d_1)\sigma}{2\sqrt{T}} \leq 0.$$

3.2 Theta symmetry

In the case of drift-less theta for options with different strikes we have the following symmetry, for both puts and calls,

$$\theta(S, X, T, 0, 0, \sigma) = \frac{X}{S}\theta\left(S, \frac{S^2}{X}, T, 0, 0, \sigma\right)$$

Theta–Vega relationship There is a simple relationship between Vega and drift-less theta

$$\theta = -\frac{\text{Vega} \times \sigma}{2T}.$$

Bleed-offset volatility A more practical relationship between theta and Vega is what is known as bleed-offset vol. It measures how much the volatility must increase to offset the theta-bleed/time decay. Bleed-offset vol can be found simply by dividing the one-day theta by Vega, $\frac{\Theta}{\text{Vega}}$. In the case of positive theta you can actually have negative offset vol. Deep-in-the-money European options can have positive theta, in this case the offset-vol will be negative.

Theta–gamma relationship There is a simple relationship between drift-less gamma and drift-less theta

$$\Gamma = \frac{-2\theta}{S^2\sigma^2}.$$

4 Rho Greeks

4.1 Rho

Rho is the option's sensitivity to a small change in the risk-free interest rate.

Call

$$\rho_{\text{call}} = \frac{\partial c}{\partial r} = TXe^{-rT}N(d_2) > 0,$$

in the case the option is on a future or forward (that is b will always stay 0) the rho is given by

$$\rho_{\text{call}} = \frac{\partial c}{\partial r} = -Tc < 0.$$

Put

$$\rho_{\text{put}} = \frac{\partial p}{\partial r} = -TXe^{-rT}N(-d_2) < 0$$

in the case the option is on a future or forward (that is b will always stay 0) the rho is given by

$$\rho_{\text{put}} = \frac{\partial c}{\partial r} = -Tp < 0.$$

4.2 Cost-of-carry

This is the option's sensitivity to a marginal change in the cost-of-carry rate.

Cost-of-carry call

$$\frac{\partial c}{\partial b} = TSe^{(b-r)T}N(d_1) > 0.$$

Cost-of-carry put

$$\frac{\partial p}{\partial b} = -TSe^{(b-r)T}N(-d_1) < 0.$$

5 Probability Greeks

In this section we will look at risk neutral probabilities in relation to the BSM formula. Keep in mind that such risk adjusted probabilities could be very different from real world probabilities.[4]

5.1 In-the-money probability

In the (Black and Scholes 1973, Merton 1973) model, the risk neutral probability for a call option finishing in-the-money is

$$\zeta_c = N(d_2) > 0,$$

and for a put option

$$\zeta_p = N(-d_2) > 0.$$

This is the risk neutral probability of ending up in-the-money at maturity. It is not identical to the real world probability of ending up in-the-money. The real probability we simply cannot extract from options prices alone. A related sensitivity is the strike-delta, which is the partial derivatives of the option formula with respect to the strike price

$$\frac{\partial c}{\partial X} = -e^{-rT}N(d_2) > 0,$$
$$\frac{\partial p}{\partial X} = e^{-rT}N(-d_2) > 0.$$

This can be interpreted as the discounted risk neutral probability of ending up in-the-money (assuming you take the absolute value of the call strike-delta).

Probability mirror strikes For a put and a call to have the same risk neutral probability of finishing in-the-money, we can find the probability symmetric strikes

$$X_p = \frac{S^2}{X_c} e^{(2b-\sigma^2)T}, \qquad X_c = \frac{S^2}{X_p} e^{(2b-\sigma^2)T},$$

where X_p is the put strike, and X_c is the call strike. This naturally reduces to $N[d_2(X_c)] = N[d_2(X_p)]$. A special case is $X_c = X_p$, a probability mirror straddle (probability-neutral straddle). We have this at

$$X_c = X_p = Se^{(b-\sigma^2/2)T}.$$

At this point the risk neutral probability of ending up in-the-money is 0.5 for both the put and the call. Standard puts and calls will not have the same value at this point. The same value for a put and a call occurs when the options are at-the-money forward, $X = S^{bT}$. However, for a cash-or-nothing option (see Reiner and Rubinstein 1991b, Haug 1997) we will also have value-symmetry for puts and calls at the risk neutral probability strike. Moreover, at the probability-neutral straddle we will also have Vega symmetry as well as zero Vomma.

Strikes from probability Another interesting formula returns the strike of an option, given the risk neutral probability p_i of ending up in-the-money. The strike of a call is given by

$$X_c = S\exp[-N^{-1}(p_i)\sigma\sqrt{T} + (b - \sigma^2/2)T],$$

where $N^{-1}(x)$ is the inverse cumulative normal distribution. The strike for a put is given by

$$X_p = S\exp[N^{-1}(p_i)\sigma\sqrt{T} + (b - \sigma^2/2)T].$$

5.2 DzetaDvol

Zeta's sensitivity to change in the implied volatility is given by

$$\frac{\partial \zeta_c}{\partial \sigma} = \frac{\partial \zeta_p}{\partial \sigma} = -n(d_2)\left(\frac{d_1}{\sigma}\right) \le\ge 0$$

and for a put

$$\frac{\partial \zeta_p}{\partial \sigma} = \frac{\partial \zeta_p}{\partial \sigma} = n(d_2)\left(\frac{d_1}{\sigma}\right) \le\ge 0.$$

Divide by 100 to get the associated measure for percentage point volatility changes.

5.3 DzetaDtime

The in-the-money risk neutral probability's sensitivity to moving closer to maturity is given by

$$-\frac{\partial \zeta_c}{\partial T} = n(d_2)\left(\frac{b}{\sigma\sqrt{T}} - \frac{d_1}{2T}\right) \leq\geq 0,$$

and for a put

$$-\frac{\partial \zeta_p}{\partial T} = -n(d_2)\left(\frac{b}{\sigma\sqrt{T}} - \frac{d_1}{2T}\right) \leq\geq 0.$$

Divide by 365 to get the sensitivity for a one-day move.

5.4 Risk neutral probability density

BSM second partial derivatives with respect to the strike price yield the risk neutral probability density of the underlying asset, see Breeden and Litzenberger (1978) (this is also known as the strike gamma)

$$\text{RND} = \frac{\partial^2 c}{\partial X^2} = \frac{\partial^2 p}{\partial X^2} = \frac{n(d_2)e^{-rT}}{X\sigma\sqrt{T}} \geq 0.$$

Figure 7 illustrates the risk neutral probability density with respect to variable time and asset price. With the same volatility for any asset price this is naturally the log-normal distribution of the asset price, as evident from the graph.

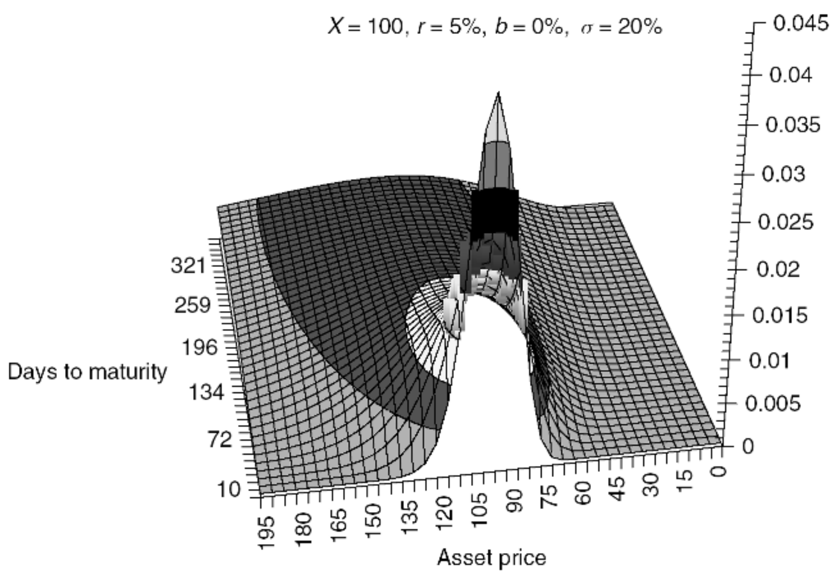


Figure 7: Risk neutral density

5.5 From in-the-money probability to density

Given the in-the-money risk-neutral probability, p_i, the risk neutral probability density is given by

$$\text{RND} = \frac{e^{-rT} n[N^{-1}(p_i)]}{X\sigma\sqrt{T}},$$

where $n()$ is the normal density function.

5.6 Probability of ever getting in-the-money

For in-the-money options the probability of ever getting in-the-money (hitting the strike) before maturity naturally equals unity, since we are already in-the-money. The risk neutral probability for an out-of-the-money call ever getting in-the-money is[5]

$$p_c = (X/S)^{\mu+\lambda} N(-z) + (X/S)^{\mu-\lambda} N(-z + 2\lambda\sigma\sqrt{T}).$$

Similarly, the risk neutral probability for an out-of-the-money put ever getting in-the-money (hitting the strike) before maturity is

$$p_p = (X/S)^{\mu+\lambda} N(z) + (X/S)^{\mu-\lambda} N(z - 2\lambda\sigma\sqrt{T}),$$

where

$$z = \frac{\ln(X/S)}{\sigma\sqrt{T}} + \lambda\sigma\sqrt{T}, \qquad \mu = \frac{b - \sigma^2/2}{\sigma^2}, \qquad \lambda = \sqrt{\mu^2 + \frac{2r}{\sigma^2}}.$$

This is equal to the barrier hit probability used for computing the value of a rebate, developed by Reiner and Rubinstein (1991a). Alternatively, the probability of ever getting in-the-money before maturity can be calculated in a very simple way in a binomial tree, using Brownian bridge probabilities.

End of Part 2

BSD trader Sergeant, that is all for now. You now know the basic operation of the Black–Scholes weapon.
New trader Did I hear you right? 'Sergeant'?
BSD trader Yes. Now that you know the basics of the Black–Scholes weapon, I have decided to promote you.
New trader Thank you, Sir, for teaching me all your tricks.
BSD trader Here's a three million loss limit. Time for you to start trading.
New trader Only three million?

FOOTNOTES & REFERENCES

1. While the other sensitivities have names that correspond to Greek letters Vega is the name of a star.

2. Described by Adamchuk on the Wilmott forum www.wilmott.com on February 6, 2002.
3. Vega bucketing simply refers to dividing the Vega risk into time buckets.
4. Risk neutral probabilities are simply real world probabilities that have been adjusted for risk. It is therefore not necessary to adjust for risk also in the discount factor for cash flows. This makes it valid to compute market prices as simple expectations of cash flows, with the *risk adjusted probabilities*, discounted at the *risk less interest rate*—hence the common name 'risk neutral' probabilities, which is somewhat of a misnomer.
5.This analytical probability was first published by Reiner and Rubinstein (1991a) in the context of barrier hit probability.

■ Black, F. (1976) The pricing of commodity contracts. *Journal of Financial Economics*, 3, 167–179.
■ Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–654.
■ Breeden, D. T. and Litzenberger, R. H. (1978) Price of state-contingent claims implicit in option prices. *Journal of Business*, 51, 621–651.
■ Haug, E. G. (1993) Opportunities and perils of using option sensitivities. *Journal of Financial Engineering*, 2(3), 253–269.
■ Haug, E. G. (1997) *The Complete Guide to Option Pricing Formulas*. McGraw-Hill, New York.
■ Haug, E. G. (2003) Know your weapon, Part 1. *Wilmott Magazine*, May.
■ Merton, R. C. (1973) Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4, 141–183.
■ Reiner, E. and Rubinstein, M. (1991a) Breaking down the barriers. *Risk Magazine*, 4(8).
■ Reiner, E. and Rubinstein, M. (1991b) Unscrambling the binary code. *Risk Magazine*, 4(9).
■ Taleb, N. (1997) *Dynamic Hedging*. John Wiley & Sons.
■ Webb, A. (1999) The sensitivity of Vega. *Derivatives Strategy,* http://www.derivativesstrategy.com/magazine/archive/1999/1199fea1.asp, November, 16–19.
■ Wystrup, U. (2002) Vanilla options, in the book *Foreign Exchange Risk* by Hakala, J. and Wystrup, U. Risk Books.

5

Take a Chance

Bill Ziemba

Gambling and investment practices are not so far removed from one another.

There is a fine but distinct line between the public's and the law's distinction between investing and legalized gambling. Stocks and bonds, bank accounts and real estate are traditional investments. Poker, blackjack, lotteries and horseracing are popular gambling games. Gold and silver, commodity and financial futures and stock and index options are somewhat in between but are generally thought to be on the investment side of the line. English spread betting is a good example where legal bets can be made without tax liability on sports events and financial investments such as stock index futures. The higher transaction costs are compensated by the absence of taxes. See Tables 1 and 2.

In all of these situations, one is making decisions whose resulting outcome has some degree of uncertainty. The outcome may also depend upon the actions of others. For example, consider buying shares of Qualcom. The stock is one of the few high flying, US, internet, high tech stocks that did not completely crash in 2000. The stock was around 100 in December versus a high of 200 in January 2000. The company has signed deals with the Chinese government and others for their pioneering wireless technology. Despite its 90 plus price earnings ratio, its future prospects looked excellent. While its niche in digital wireless communication is fairly unique and future demand growth looks outstanding, others could possibly market successful and cheaper alternatives or the marketing deals could unravel. What looks good now has frequently turned into disaster in the late 1990's technology market place since enormous growth is needed in the future to justify today's high prices. Qualcom has continued to grow but at a slower rate and its stock price fell to a third of its December 2000 value in mid-2002. Such is a typical price experience of high PE stocks.

Economic effects that manifest themselves into general market trends are important also in a stock's price. Most stocks are going up in a rising market and vice versa. Indeed in the most popular stock market pricing theory—the so-called capital asset market equilibrium beta model—securities are compared via their relative price movement up or down and at what rate when the general market average (e.g. S&P500 index) rises or falls. Over time, stocks have greatly outperformed bonds, T-bills, inflation and gold. For example, $1 invested in 1802 in gold was only worth $11 in 1997, CPI inflation was $13, T-bills $3679, bonds $10,744 and stocks a whopping $7.47 million. And the gains are pretty steady over time; see Table 3.

So for success in stocks, one has two crucial elements: general uncertainty about the economy and the product's acceptance and the effect of competition.

TABLE 1: ESSENTIALS OF INVESTMENT AND GAMBLING

Investing	Leveraged investing	Gambling
Bank accounts and term deposits Gold and silver Mutual funds Real estate Stocks and bonds	Commodity and financial futures Options Spread betting Hedge funds	Blackjack Dice games Horseracing Jai Alai Lotteries Roulette Sports betting
• Positive sum game usually • SLOW — 'buy and hold' • To preserve your capital • Gains usually exceed transaction costs for the average person • Path dependence is not extremely crucial	• ZERO SUM game • Many winners and many losers • Low transactions costs • Risk control is important • Path dependence is crucial	• Negative sum game, average person LOSES • FAST play • Entertainment • High transactions costs • Winners share net pool: house cannot lose if payoffs are pari-mutuel, percent of play • Edge on each play: each play is either won or lost; house cannot lose except in fixed odds cases where they do not diversify

TABLE 2: ASPECTS OF SEVERAL GAMBLING/INVESTMENT SITUATIONS WITH WINNING SYSTEMS

	Average edge	Probability of winning	Wagers	Does the wager affect the odds?
Blackjack	1.5%	45–55%	Large	No
Financial futures	10%+	2–98%	Extremely large	Yes
Horseracing	10%+	2–98%	Medium to large	Yes
Lotteries	25%+	Less than 1%	Very small	Yes

An analogous situation is found in a gambling context such as sports betting on the super bowl. The general uncertainty affects the outcome of the game whereas the competition from other players shows up in higher or lower odds.

What then is the difference between investing and gambling? In investing one buys some item be it a stock, a bar of platinum or a waterfront house, pays the commission to the seller, and goes off possibly for a long time. Nothing prevents all participants from gaining. In fact they usually do. The essence of an investment is this: it is possible for *every* person buying the item to gain and it is generally expected that most people will in fact reap profits.

TABLE 3: AVERAGE CONTINUOUSLY COMPOUNDED YEARLY RATES OF RETURN, 1802–1997

	Average edge
Gold	1.2%
CPI	1.3%+
Bills	4.1%+
Bonds	4.6%
Stocks	7.9%

More interesting and profitable is the construction of hedges involving combinations of long and short risky situations where one makes a moderate profit most of the time with little risk. This is the basis of some successful hedge fund and bank trading department strategies.

The situation is different with a gambling game. There is usually a house or some type of negative or zero sum game, be it a casino, racetrack management or provincial lottery that takes a predetermined (minimum or average) commission. On the surface, it seems that the house cannot lose except in rare instances and certainly not in the aggregate. Surprisingly, lottery organizations around the world make many conceptual mistakes in game design that lead to situations in which winning player strategies exist. How about the players? On average, they must lose since the house always makes its commission. So all players *cannot* win. Some may win but many or most will lose. In fact estimates show that very few persons (about one in a hundred) actually make profits in gambling over extended periods of time. Most people talk about their wins and are much more quiet about their losses. As my colleague Mr B says, they want ‘bragging rights’. For most people, gambling is a form of entertainment and although they would like to win, their losses seem to be adequately compensated by the enjoyment of the play. The game also does not take very long. When it is played, the management takes its commission and distributes the prizes or winnings in a quick and orderly fashion. Then the game is repeated.

A gambling situation can be of two types. In fixed payment games the players wager against the house. In any particular play, the house and the player either win or lose. What one wins the other loses so both parties have risk. However, by having an edge and by diversifying over many players, the house remains profitable. But the players cannot do this. In pari-mutuel games, the house is passive and takes a fixed piece of the action. The rest is then split among the winners. In this way the house cannot lose and takes no risk.

In these games, the players are really wagering against each other. In both types of gambling the average player loses. Some players may win, but the players as a group have a net loss. The vigorish (transactions costs) is essentially the payment for the pleasure of playing. It is an important result in the mathematics of gambling that, faced with a sequence of unfavorable games, no gambling system can be devised that will yield a profit on average after one, two, three or any number of plays. You simple cannot change an unfavorable (negative expected value) game into a favorable game with a clever mathematical betting scheme.

The colocation of money and math

Be especially wary of advice of the doubling up strategies: martingales, pyramids, etc. While such systems may allow you to make small profits most of the time, the gigantic losses you suffer once in a while yield losses in the long run.

The most useful result that we have for unfavorable games is that if you want to maximize your chances of achieving a goal before falling to some lower wealth level, you should use 'bold play'. With bold play you do not let the casino defeat you by grinding out small profits from you along the way. Rather, you bet amounts that get you to your goal as soon as possible.

Consider roulette, which is an unfavorable game with an edge of minus 2/38 or minus 5.26. Assuming that you are not able to predict the numbers that will occur any better than random, you should bet on only one number with a wager that if you win you will either reach your goal or a wealth level from which you can reach it on one or more subsequent plays. If your fortune is \$10 and your goal is \$1000, then it is optimal to bet the entire \$10 on only one number. If you lose you are out. If you win you have \$360 (with the 35-1 payoff) and then you bet \$19, which takes you to \$1006 if you win and \$341 if you lose. Upon losing you would bet the smallest amount—\$19 again—so that if you win you reach your goal of \$1000, etc. This bold play strategy always gives you the highest chance of achieving your goal.

On the other hand, when you—the casino in the case of roulette—have the edge and your goal is to reach some higher level of wealth before falling to a lower level with as high a probability as possible, then 'timid play' is optimal. With timid play, you wager small amounts to make sure some small sample random result does not hurt you. Then, after a moderate number of plays you are virtually sure of winning. This is precisely what casinos do. With even a small edge, all they need to do to be practically guaranteed of large and steady profits is to diversify the wagers so that the percentage wagered by each gambler is small. With crowded casinos, this is usually easy to accomplish. A simple example of this idea, non-diversification, shows up in many if not most or all financial disasters.

Changing a gamble into an investment

The point of all this is that if you are to have any chance at all of winning, you must develop a playing strategy so that at least some of the time, and preferably most or all of the time, you are betting when you are getting on average more than a dollar for each dollar wagered. We call this changing a gamble into an investment. This is possible in roulette, see Tom Bass' *The Eudomonic Pie*. Also, for the simpler game of the wheel of fortune, see Ed Thorp's article in the March 1982 issue of *Gambling Times*.

In this chapter and Chapter 6, I discuss topics in the mathematics of gambling and investment. The basic goal is to turn gambles into investments with the development of good playing strategies and then to wager intelligently. The strategy development follows general principles but is somewhat different for each particular situation. The wagering or money management concepts apply to all games. The difference in application depends upon the edge and the probability of winning. The size of the wager depends on the edge but much more so on the probability of winning if one takes a long run rate of growth of profit approach, as I discuss in Chapter 6. We will look at situations where the player has an edge and develop playing strategies to exploit that edge. These are situations where, on average, the player can win using a workable system. The analyses utilize concepts from modern financial economics investment theory and related mathematical optimization, psychological, statistical and computer techniques and apply them to the gambling situations to yield profitable systems. This frequently involves the identification of a security market imperfection, anomaly or partially predictable prices. Naturally in gambling situations all players cannot win so the potential gain will depend upon how good the system is, how well it is played and how many are using it or other profitable systems—also, most crucially, on the risk control system in use. Nor will every game have a useful favorable system where one can make profits on average. Baccarat or Chemin de Fer is one such example. However, virtually every financial market will have strategies that lead to winning investment situations.

There are two aspects of the analysis of each situation: when should one bet and how much should be bet? These may be referred to as strategy development and money management. They are equally important. While the strategy development aspect is fairly well understood by many, the money management (risk control) element is more subtle and it is errors here that lead to financial disasters. In the next chapter we will discuss the basic theory of gambling/investing over time using the capital growth/Kelly and fractional Kelly betting systems and apply this to futures trading. Chapters 6 and 7 discuss hedge funds and focus on strategy development and risk control failures such as Long Term Capital Management in 1998 and models of lottery, horserace and other betting situations, pension, insurance company and individual investment planning over time.

I have been fortunate to have worked and consulted with four individuals who have used these ideas in four separate areas: ‘market neutral’ hedge funds, private futures trading hedge funds, mispriced options and racetrack betting to turn essentially zero into more than $300 million. All four, while different in many ways, began with a gambling focus and retain this in their trading. They are true investors with heavy emphasis on computerized mathematical investing and risk control. They understand downside risk well. They are even more focused on not losing their capital than on having more winnings. They have their losses but rarely do they overbet or non-diversify enough to have a major blowout like the hedge fund occurrences.

6
Good and Bad Properties of the Kelly Criterion

Bill Ziemba

> If your outlook is well extended, the Kelly criterion is the approach best suited to generating a fortune.

In this chapter I would like to discuss good and bad properties of the Kelly expected log capital growth criterion and in the process lead into the next chapter on hedge funds by discussing two of the great traders who ran unofficial hedge funds. The main advantages are that if your horizon is long enough then the Kelly criterion is the road, however bumpy, to the most wealth at the end and the fastest path to a given rather large fortune.

Thorp (1997) has shown that the great investor Warren Buffett's Berkshire Hathaway actually has had a growth path quite similar to full Kelly betting. Figure 1 shows this performance from 1985 to 2000 in comparison with other great funds. Buffett also had a great record from 1977 to 1985 turning 100 into 1429.87, and 65,852.40 in April 2000.

Keynes was another Kelly-type bettor. His record running King's College Cambridge's Chest Fund is shown in Figure 2 versus the British market index for 1927 to 1945, data from Chua and Woodward (1983). Notice how much Keynes lost the first few years; obviously his academic brilliance and the recognition that he was facing a rather tough market kept him in this job. In total his geometric mean return beat the index by 10.01%. Keynes was an aggressive investor with a beta of 1.78 versus the benchmark United Kingdom market return, a Sharpe ratio of 0.385, geometric mean returns of 9.12% per year versus −0.89% for the benchmark. Keynes had a yearly standard deviation of 29.28% versus 12.55% for the benchmark. These returns do not include Keynes' (or the benchmark's) dividends and interest, which he used to pay the college expenses. These were 3% per year. Kelly cowboys have their great returns and losses and embarrassments. Not covering a grain contract in time led to Keynes taking delivery and filling up the famous chapel. Fortunately it was big enough to fit in the grain and store it safely until it could be sold.

Keynes emphasized three principles of successful investments in his 1933 report:

1. A careful selection of a few investments (or a few types of investment) having regard to their cheapness in relation to their probable actual and potential intrinsic value over a period of years ahead and in relation to alternative investments at the time;

Mr Keynes believed God to be a large chicken, the Reverend surmised

2. A steadfast holding of these in fairly large units through thick and thin, perhaps for several years until either they have fulfilled their promise or it is evident that they were purchased on a mistake; and
3. A balanced investment position, i.e. a variety of risks in spite of individual holdings being large, and if possible, opposed risks.

He really was a lot like Buffett with an emphasis on value, large holdings and patience.

In November 1919 Keynes was appointed second bursar. Up to this time King's College investments were only in fixed income trustee securities plus their own land and buildings. By June 1920 Keynes convinced the college to start a separate fund containing stocks, currency and commodity futures. Keynes became first bursar in 1924 and held this post which had final authority on investment decisions until his death in 1945.

And Keynes did not believe in market timing as he said:

> We have not proved able to take much advantage of a general systematic movement out of and into ordinary shares as a whole at different phases of the trade cycle. As a result of these experiences I am clear that the idea of wholesale shifts is for various reasons impracticable and indeed undesirable. Most of those who attempt to, sell too

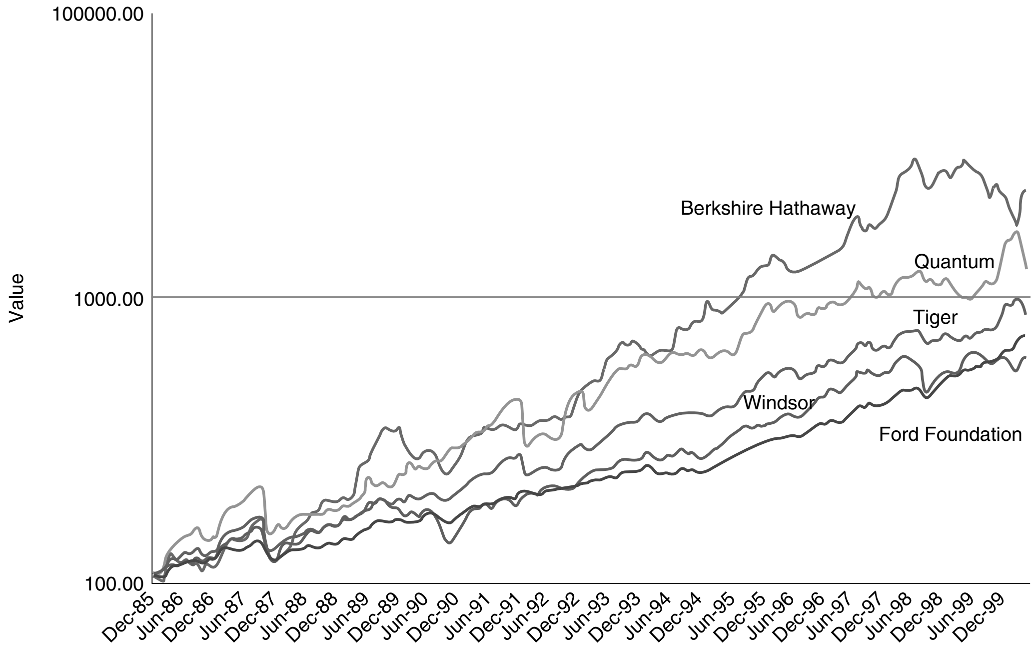


Figure 1: Growth of assets, log scale, various high performing funds, 1985–2000. Source: Ziemba (2003)

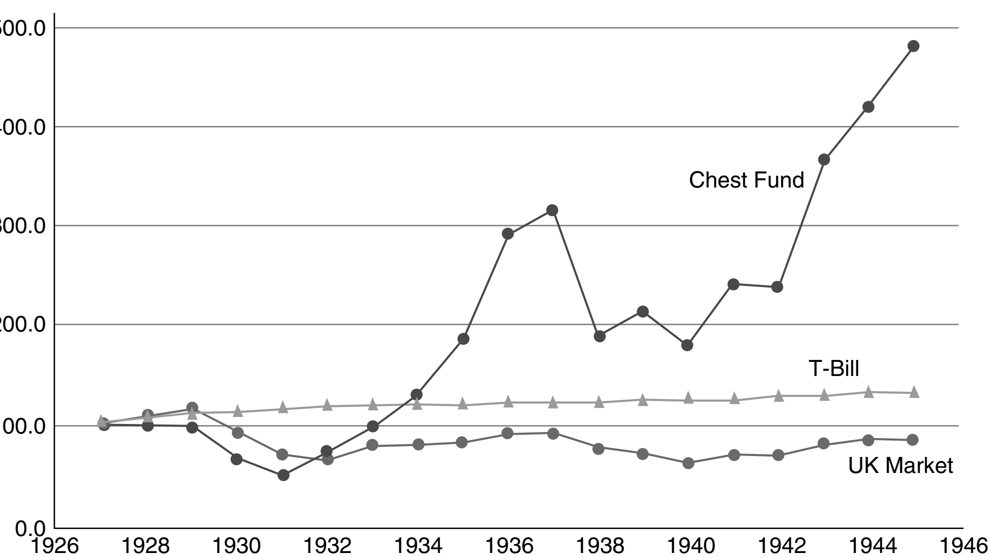


Figure 2: Graph of the performance of the Chest Fund, 1927–1945. Source: Ziemba (2003)

late and buy too late, and do both too often, incurring heavy expenses and developing too unsettled and speculative a state of mind, which, if it is widespread, has besides the grave social disadvantage of aggravating the scale of the fluctuations.

The main disadvantages result because the Kelly strategy is very very aggressive with huge bets that are larger and larger as the situations are most attractive: recall that the bet is mean return divided by the odds of winning. As I repeatedly argue it's the mean that counts by far the most. There is about a 20–2:1 ratio of expected utility loss from similarly sized errors of means, variances and covariances, respectively. See Table 1 and Figure 3; see Kallberg and Ziemba (1984) and Chopra and Ziemba (1993) for details. Returning to Buffett who gets the mean right, better than almost all, notice that the other funds he outperformed are not shabby ones at all.

TABLE 1: AVERAGE RATIO OF CERTAINTY EQUIVALENT LOSS FOR ERRORS IN MEANS, VARIANCES AND COVARIANCES

t Risk tolerance	Errors in means vs covariances	Errors in means vs variances	Errors in variances vs covariances
25	5.38	3.22	1.67
50	22.50	10.98	2.05
75	56.84	21.42	2.68
	↓	↓	↓
	20	10	2
	Error Mean	Error Var	Error Covar
	20	2	1

Source: Chopra and Ziemba (1993)

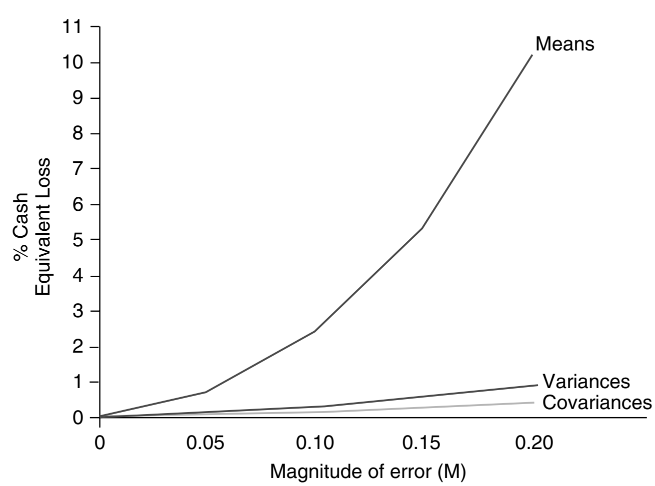


Figure 3: Mean percentage cash equivalent loss due to errors in inputs. Source: Chopra and Ziemba (1993)

TABLE 2: KELLY CRITERION PROPERTIES

Good Maximizing $E[\log X]$ asymptotically maximizes the rate of asset growth. See Breiman (1961), Algoet and Cover (1988).

Good The expected time to reach a pre-assigned goal is asymptotical as X increases least with a strategy maximizing $E[\log X_N]$. See Breiman (1961), Algoet and Cover (1988), Browne (1997).

Good Maximizing median $\log X$. See Ethier (1987).

Bad False property: If maximizing $E[\log X_N]$ almost certainly leads to a better outcome then the expected utility of its outcome exceeds that of any other rule provided N is sufficiently large. Counter example: $u(x) = x,\ 1/2 < p < 1$, Bernoulli trials $f = 1$ maximizes $E[U(x)]$ but $f = 2p - 1 < 1$ maximizes $E[\log X_N]$. See Samuelson (1971), Thorp (1975).

Good The $E[\log X]$ bettor never risks ruin. See Hakansson and Miller (1975).

Bad If the $E[\log X_N]$ bettor wins then loses or loses then wins, he is behind. The order of win and loss is immaterial for one, two, . . ., sets of trials. $(1 + \gamma)(1 - \gamma)X_0 = (1 - \gamma^2)X_0 \le X_0$.

Good The absolute amount bet is monotone in wealth. $(\delta E[\log X])/\delta W_0 > 0$.

Bad The bets are extremely large when the wager is favorable and the risk is very low. For single investment worlds, the optimal wager is proportional to the edge divided by the odds. Hence for low risk situations and corresponding low odds, the wager can be extremely large. For one such example, see Ziemba and Hausch (1986; 159–160). There, in the inaugural 1984 Breeders' Cup Classic \$3 million race, the optimal fractional wager on the 3–5 shot Slew of Gold was 64%. Thorp and I actually made this place and show bet and won with a low fractional Kelly wager. Slew actually finished third but the second place horse Gate Dancer was disqualified and placed third. Luck (a good scenario) is also nice to have in betting markets. Wild Again won this race; the first great victory by the masterful jockey Pat Day.

Bad One overinvests when the problem data is uncertain. Investing more than the optimal capital growth wager is dominated in a growth-security sense. Hence, if the problem data provides probabilities, edges and odds that may be in error, then the suggested wager will be too large.

Bad The total amount wagered swamps the winnings—that is, there is much churning. Ethier and Tavaré (1983) and Griffin (1985) show that the Expected Gain/E Bet is arbitrarily small and converges to zero in a Bernoulli game where one wins the expected fraction p of games.

Bad The unweighted average rate of return converges to half the arithmetic rate of return. Related to property 5 this indicates that you do not seem to win as much as you expect. See Ethier and Tavaré (1983) and Griffin (1985).

Bad Betting double the optimal Kelly bet reduces the growth rate of wealth to zero plus the risk-free rate. See Janacek (1998) and Ziemba (1993) for proofs.

Good The $E[\log X]$ bettor is never behind any other bettor on average in 1, 2, . . . trials. See Finkelstein and Whitley (1981).

Good The $E[\log X]$ bettor has an optimal myopic policy. He does not have to consider prior to subsequent investment opportunities. This is a crucially important result for practical use. Hakansson (1971) proved that the myopic policy obtains for dependent investments with the log utility function. For independent investments and power utility a myopic policy is optimal, see Mossin (1968).

Good The chance that an $E[\log X]$ wagerer will be ahead of any other wagerer after the first play is at least 50%. See Bell and Cover (1980).

(*continued overleaf*)

TABLE 2: (*continued*)

Good Simulation studies show that the E[log]X bettor's fortune pulls way ahead of other strategies, wealth for reasonable-sized samples. The key again is risk. See Ziemba and Hausch (1986). General formulas are in Aucamp (1993).

Good If you wish to have higher security by trading it off for lower growth, then use a negative power utility function or fractional Kelly strategy. See MacLean *et al*. (2005). MacLean *et al*. (2004) show how to compute the coefficient to stay above a growth path with given probability.

Bad Despite its superior long-run growth properties, it is possible to have very poor return outcome. For example, making 700 wagers all of which have a 14% advantage, the least of which had a 19% chance of winning, can turn \$1000 into \$18. But \$1000 turns into \$100 000 plus 16.6% of the time, see Ziemba and Hausch (1996).

Bad It can take a long time for a Kelly bettor to dominate an essentially different strategy. In fact this time may be without limit. Suppose $\mu_A = 20\%$, $\mu_B = 10\%$, $\sigma_A = \sigma_B = 10\%$. Then in five years A is ahead of B with 95% confidence. But if $\sigma_A = 20\%$, $\sigma_B = 10\%$ with the same means, it takes 157 years for A to beat B with 95% confidence. In coin tossing suppose game A has an edge of 1.0% and game B 1.1%. It takes two million trials to have an 84% chance that game A dominates game B, see Thorp (1997).

Indeed they are George Soros' Quantum, John Neff's Windsor, Julian Robertson's Tiger and the Ford Foundation, all of whom had great records as measured by the Sharpe ratio. Buffett made 32.07% per year net from July 1977 to March 2000 versus 16.71% for the S&P500. Wow! Those of us who like wealth prefer Warren's path but his higher standard deviation path (mostly winnings) leads to a lower Sharpe (normal distribution based) measure; see Clifford *et al*. (2001). See Ziemba (2005) for a modification of the Sharpe ratio only considering losses. This new measure Sharpe improved only Buffett in this group but still Ford is preferred because Buffett has large losses as well as large gains.

Kelly has essentially zero risk aversion since its Arrow–Pratt risk aversion index is $u''(w)/u'(w) = 1/w$, which is essentially zero. Hence it never pays to bet more than the Kelly strategy because then risk increases (lower security) and growth decreases so is stochastically dominated. As you bet more and more above the Kelly bet, its properties become worse and worse. When you bet exactly twice the Kelly bet, then the growth rate is zero plus the risk free rate.

If you bet more than double the Kelly criterion, then you will have a negative growth rate. With derivative positions one's bet changes continuously so a set of positions amounting to a small bet can turn into a large bet very quickly with market moves. Long Term Capital is a prime example of this overbetting leading to disaster but the phenomenon occurs all the time all over the world. Overbetting plus a bad scenario leads invariably to disaster.

Thus you must either bet Kelly or less. We call less than Kelly 'fractional Kelly', which is simply a blend of Kelly and cash. Consider the negative power utility function $\delta\omega^{\delta}$ for $\delta<0$. This utility function is concave and when $\delta \to 0$ it converges to log utility. As δ gets larger negatively, the investor is less aggressive since his Arrow–Pratt risk aversion is also higher. For a given δ an $\alpha = 1/(1-\delta)$ between 0 and 1, will provide the same portfolio when α is invested in the Kelly portfolio and $1-\alpha$ is invested in cash.

This result is correct for log-normal investments and approximately correct for other distributed assets; see MacLean, Ziemba and Li (2005). For example, half Kelly is $\delta = -1$ and quarter Kelly is $\delta = -3$. So if you want a less aggressive path than Kelly, then pick an appropriate δ. MacLean *et al*.(2004) discuss a way to pick δ continuously in time so that wealth will stay above a desired wealth growth path with high given probability.

I have listed these and other important Kelly criterion properties in Table 2 which was updated from MacLean, Ziemba and Blazenko (1992) and MacLean and Ziemba (1999).

REFERENCES

■ Chopra, V. and Ziemba, W. T. (1993) The effect of errors in mean, variance and covariance estimates on optimal portfolio choice. *Journal of Portfolio Management*, 6–11.

■ Chua, J. H. and Woodward, R. S. (1983) J. M. Keynes's investment performance; a note. *Journal of Finance*, 38(1): 232–235.

■ Clifford, S. W., Kroner, K. F. and Siegel, L. B. (2001) In pursuit of performance: the greatest return stories ever told. *Investment Insights, Barclays Global Investor*, 4(1): 1–25.

■ Kallberg, J. G. and. Ziemba, W. T. (1984) Mis-specifications in portfolio selection problems. In G. Bamberg and A. Spremann (eds.), *Risk and Capital*, pp. 74–87. Springer Verlag, New York.

■ MacLean, L. C. and. Ziemba, W. T. (1999) Growth versus security tradeoffs in dynamic investment analysis. *Annals of Operations Research*, 85, 193–227.

■ MacLean, L. C., Ziemba, W. T. and Blazenko, G. (1992) Growth versus security in dynamic investment analysis. *Management Science*, 38, 1562–1585.
■ MacLean, L. C., Ziemba, W. T. and Li, Y. (2005) Time to wealth goals in capital accumulation and the optimal trade off of growth versus security. *Quantitative Finance*, in press.
■ Thorp, E. O. (1997) The Kelly criterion in blackjack, sports betting, and the stock market, Presented at the 10th International Conference on Gambling and Risk Taking, Montreal, June.
■ Ziemba, W. T. (2003) *The Stochastic Programming Approach to Asset, Liability and Wealth Management*. AIMR, Charlottesville, Virginia.
■ Ziemba, W. T. (2005) The symmetric downside risk Sharpe ratio and the evaluation of great investors and speculators. *Journal of Portfolio Management*, forthcoming, Fall, 15pp.

In the table

■ Algoet, P. H. and Cover, T. M. (1988) Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Annals of Probability*, 16(2): 876–898.
■ Aucamp, D. (1993) On the extensive number of plays to achieve superior performance with the geometric mean strategy. *Management Science*, 39, 1163–1172.
■ Bell, R. M. and Cover, T. M. (1980) Competitive optimality of logarithmic investment. *Math of Operations Research*, 5: 161–166.
■ Breiman, L. (1961) Optimal gambling system for favorable games, in Proceedings 4th Berkeley Symposium on Mathematics. *Statistics and Probability*, 1, 63–68.
■ Browne, S. (1997) Survival and growth with a fixed liability: optimal portfolios in continuous time. *Math of Operations Research*, 22, 468–493.
■ Ethier, S. N. (1987) The proportional bettor's fortune. Proceedings 7th International Conference on Gambling and Risk Taking. Department of Economics, University of Nevada, Reno.
■ Ethier, S. N. and Tavaré, S. (1983) The proportional bettor's return on investment. *Journal of Applied Probability*, 20, 563–573.
■ Finkelstein, M. and Whitley, R. (1981) Optimal strategies for repeated games. *Advanced Applied Probability*, 13, 415–428.
■ Griffin, P. (1985) Different measures of win rates for optimal proportional betting. *Management Science*, 30, 1540–1547.
■ Hakansson, N. H. (1971) On myopic portfolio policies: with and without serial correlation of yields. *Journal of Business*, 44(3): 324–334.
■ Hakansson, N. H. and Miller, B. L. (1975) Compound-return mean-variance efficient portfolios never risk ruin. *Management Science*, 22, 391–400.
■ Janacek, K. (1998) Optimal growth in gambling and investing. MSc thesis, Charles University.
■ MacLean, L. C., Sanegre, R., Zhao, Y. and Ziemba, W. T. (2004) Capital growth with security. *Journal of Economic Dynamics and Control*, 28(4): 937–954.
■ Mossin, J. (1968) Optimal multiperiod portfolio policies. *Journal of Business*, 41, 215–229.
■ Samuelson, P. A. (1971) The "fallacy" of maximizing the geometric mean in long sequences of investing or gambling. *Proceedings National Academy of Science*, 68, 2493–2496.
■ Thorp, E. O. (1975). Portfolio choice and the Kelly criterion. In W.T. Ziemba and R.G. Vickson (eds), *Stochastic Optimization Models in Finance*, Academic Press, New York.
■ Ziemba, W. T. and Hausch, D. B. (1986) *Betting at the Racetrack*. Dr. Z. Investments, Los Angeles.

7
Algorithms: Mathematics of Gambling and Investment. The Stochastic Programming Approach to Managing Hedge and Pension Fund Risk, Disasters and their Prevention

Bill Ziemba

Hedge fund and pension fund disasters occur with different speeds. With a hedge fund, it is usually immediate in one or two days or over a month or so. That is because their positions are usually highly levered. The action is quick and furious when things go wrong. A pension fund on the other hand does not make decisions on an hourly, daily, or weekly basis like a hedge fund. Rather, their decisions are how to allocate their funds into broad investment classes over longer periods of time. Decision review periods are typically yearly or possibly quarterly after meetings with their fund managers.

There have been many hedge fund disasters such as Long Term Capital Management and Niederhoffer (1997); see Ziemba (2003). They almost invariably have three ingredients: the fund is overbet, that is, too highly levered; the positions are not really diversified; and then a bad scenario occurs. Once the trouble starts, it is hard to get out of it without excess cash. So it is better to have the cash in advance, that is, to be less levered in the first place.

Pension funds have had their share of disasters as well. And the sums are much greater. The University of Toronto announced that their pension fund lost some $450 million in 2002. The British universities pension system was in a shortfall of about 18% (5 billion pounds) in early 2005. Worldwide pensions had a shortfall of $2.5 trillion in January 2003, according to Watson Wyatt.

Pension funds of the defined benefit variety, which owe a fixed stream of money, are the source of the trouble. Many governments such as those in France, Italy, Israel and many US states have such problems. On the other hand, defined contribution plans like that of my university where you put the money in, get contributions from the university, manage the assets and have what you have experience far less trouble. Losses and gains are the property of the retirees not the plan sponsor. So these have no macro problem, though for individuals their retirement prospects can be bleak if the funds have not been well managed.

The key issue for pension funds is their strategic asset allocation to stocks, bonds, cash, real estate and possibly other assets.

Stochastic programming models provide a good way to deal with the risk control of both pension and hedge fund portfolios using an overall approach to position size taking into account various possible scenarios that may be beyond the range of previous historical data. Since correlations are scenario dependent, this approach is useful to model the overall position size. The model will not allow the hedge fund to maintain positions so large and so underdiversified that a major disaster can occur. Also the model will force consideration of how the fund will attempt to deal with the bad scenario because once there is a derivative disaster, it is very difficult to resolve the problem. More cash is immediately needed and there are liquidity and other considerations. For pension funds, the problem is a shortfall to its retirees and the political fallout from that.

Let's first discuss fixed mix versus strategic asset allocation.

1 Fixed mix and strategic asset allocation

Fixed mix strategies, in which the asset allocation weights are fixed and at each decision point the assets are rebalanced to the initial weights, are very common and yield good results. An attractive feature is an effective form of volatility pumping since they rebalance by selling assets high and buying them low. Fixed mix strategies compare well with buy and hold strategies: see for example Figure 1 which shows the 1982 to 1994 performance of a number of asset categories including mixtures of EAFE (Europe, Australia and the Far East) index, S&P500, bonds, the Russell 2000 small cap index and cash.

Theoretical properties of fixed mix strategies are discussed by Dempster *et al*. (2003) and Merton (1990) who show their advantages. In stationary markets where the return distributions are the same each year, the long run growth of wealth is exponential with probability one. The stationary assumption is fine for long run behavior but for short time horizons, even up to 10 to 30 years, using scenarios to represent the future will generally give better results.

Hensel *et al*. (1991) showed the value of strategic asset allocation. They evaluated the results of seven representative Frank Russell US clients who were having their assets managed by approved

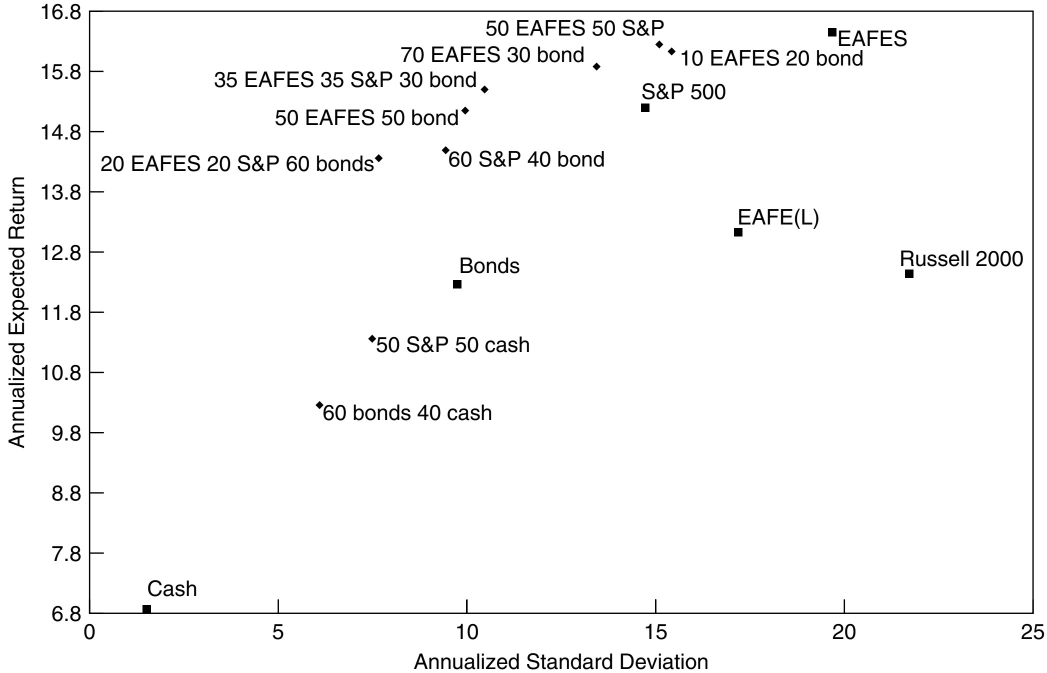


Figure 1: Historical performance of some asset categories, 1/1/1982 to 12/31/1994. Source: Ziemba and Mulvey (1998)

professional managers who are supposed to beat their benchmarks with lower risk. The study was over 16 quarters from January 1985 to December 1988. A fixed mix *naive* benchmark was: US equity (50%), non-US equity (5%), US fixed (30%), real estate (5%), cash (10%). Table 1 shows the results concerning the mean quarterly returns and the variation explained. Most of the volatility (94.35% of the total) is explained by the naive policy allocation. This is similar to the 93.6% in Brinson *et al*. (1986). T-bill returns (1.62%) and the fixed mix strategy (2.13%) explain most of the mean returns. The managers returned 3.86% versus 3.75% for T-bills plus fixed mix so they added value. This added value was from their superior strategic asset allocation into stocks, bonds and cash. The managers were unable to market time or to pick securities better than the fixed mix strategy.

Further evidence that strategic asset allocation accounts for most of the time series variation in portfolio returns while market timing and asset selection are far less important has been given by Blake *et al*. (1999). They used a nine-year (1986–1994) monthly data set on 306 UK pension funds having eight asset classes. They find also a slow mean reversion in the funds' portfolio weights toward a common, time varying strategic asset allocation.

The UK pension industry is concentrated in very few management companies. Indeed four companies control 80% of the market. This differs from the US where the largest company in 1992 had a 3.7% share according to Lakonishok *et al*. (1992). During the 1980s, the pensions were about 50% overfunded. Fees are related to performance usually relative to a benchmark or peer group. They concluded that:

TABLE 1: AVERAGE RETURN AND RETURN VARIATION EXPLAINED (QUARTERLY BY THE SEVEN CLIENTS), PERCENT

Decision level	Average contribution, %	Additional variation explained by this level (volatility), %
Minimum risk (T-bills)	1.62	2.66
Naive allocation (fixed mix)	2.13	94.35
Specific policy allocation	0.49	0.50
Market timing	(0.10)	0.14
Security selection	(0.23)	0.40
Interaction and activity	(0.005)	1.95
Total	3.86	100.00
T-bills and fixed mix	3.75	

Source: Hensel *et al.* (1991)

1. UK pension fund managers have a weak incentive to add value and face constraints on how they try to do it. Though strategic asset allocation may be set by the trustees these are flexible and have wide tolerance for short-run deviations and can be renegotiated.
2. Fund managers know that relative rather than absolute performance determines their long-term survival in the industry.
3. Fund managers earn fees related to the value of assets under management not to their relative performance against a benchmark or their peers with no specific penalty for underperforming nor reward for outperforming.
4. The concentration in the industry leads to portfolios being dominated by a small number of similar *house positions* for asset allocation to reduce the risk of relative underperformance.

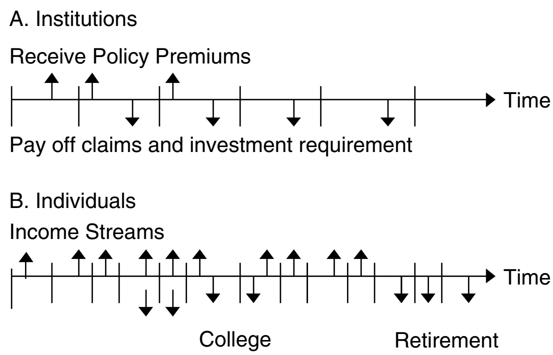
The asset classes from WM Company data were UK equities, international equities, UK bonds, international bonds, cash, UK property and international property. UK portfolios are heavily equity weighted. For example, the 1994 weights for these eight asset classes over the 306 pension funds were 53.6, 22.5, 5.3, 2.8, 3.6, 4.2, 7.6 and 0.4%, respectively. In contrast, US pension funds had 44.8, 8.3, 34.2, 2.0, 0.0, 7.5, 3.2 and 0.0%, respectively.

Most of the 306 funds had very similar returns year by year. The semi-interquartile range was 11.47 to 12.59% and the 5th and 95th percentiles were less than 3% apart.

The returns on different asset classes were not very great except for international property. The eight classes averaged value weighted 12.97, 11.23, 10.76, 10.03, 8.12, 9.01, 9.52 and −8.13 (for the international property) and overall 11.73% per year. Bonds and cash kept up with equities quite well in this period. They found, similar to the previous studies, that for UK equities, a very high percent (91.13) of the variance in differential returns across funds because of strategic asset allocation. For the other asset classes, this is lower: 60.31% (international equities), 39.82% (UK bonds), 16.10% (international bonds), 40.06% (UK index bonds), 15.18% (cash), 76.31% (UK property) and 50.91% (international property). For these other asset classes, variations in net cash flow differentials and covariance relationships explain the rest of the variation.

2 Stochastic programming models applied to hedge and pension fund problems

Let's now discuss how stochastic programming models may be applied to hedge fund pension fund problems as well as the asset-liability commitments for other institutions such as insurance companies, banks, pension funds and savings and loans and individuals. These problems evolve over time as follows:



The stochastic programming approach considers the following aspects:

- Multiple discrete time periods; possible use of end effects−steady state after decision horizon adds one more decision period to the model; the tradeoff is an end effects period or a larger model with one less period.
- Consistency with economic and financial theory for interest rates, bond prices etc.
- Discrete scenarios for random elements−returns, liabilities, currencies; these are the possible evolutions of the future; since they are discrete, they do not need to be lognormal and/or any other parametric form.
- Scenario dependent correlation matrices so that correlations change for extreme scenarios.
- Utilize various forecasting models that handle fat tails and other parts of the return distributions.
- Include institutional, legal and policy constraints.
- Model derivatives, illiquid assets and transactions costs.
- Expressions of risk in terms understandable to decision makers based on targets to be achieved and convex penalties for their non-attainment.
- This yields simple, easy to understand, risk averse utility functions that maximize long run expected profits net of expected discounted penalty costs for shortfalls; that pay more and more penalty for shortfalls as they increase (highly preferable to VaR).
- Model various goals as constraints or penalty costs in the objective.
- Maintain adequate reserves and cash levels and meet regularity requirements.
- We can now solve very realistic multiperiod problems on modern work-stations and PCs using large-scale linear programming and stochastic programming algorithms.

- The model makes you *diversify*—the key for keeping out of trouble.

I would like to focus on a model I designed for the Siemens' Austrian pension fund which was implemented in 2000. Alois Geyer of the University of Vienna built the model with me. The model is described in Geyer *et al*. (2003).

3 InnoALM, The Innovest Austrian Pension Fund Financial Planning Model

Siemens AG Österreich, part of the global Siemens Corporation, is the largest privately owned industrial company in Austria. Its businesses with revenues of €2.4 billion in 1999, include information and communication networks, information and communication products, business services, energy and traveling technology, and medical equipment. Their pension fund, established in 1998, is the largest corporate pension plan in Austria and is a defined contribution plan. Over 15 000 employees and 5000 pensioners are members of the pension plan with €510 million in assets under management as of December 1999.

Innovest Finanzdienstleistungs AG founded in 1998 is the investment manager for Siemens AG Österreich, the Siemens Pension Plan and other institutional investors in Austria. With €2.2 billion in assets under management, Innovest focuses on asset management for institutional money and pension funds. This pension plan was rated the best in Austria of 17 analyzed in the 1999/2000 period. The motivation to build InnoALM, which is described in Geyer *et al*. (2003), is part of their desire to have superior performance and good decision aids to help achieve this.

Various uncertain aspects, possible future economic scenarios, stock, bond and other investments, transactions costs, liquidity, currency aspects, liability commitments over time, Austrian pension fund law and company policy suggested that a good way to approach this was via a multiperiod stochastic linear programming model. These models evolve from Kusy and Ziemba (1986), Cariño and Ziemba *et al*. (1994, 1998a, b), Ziemba and Mulvey (1998) and Ziemba (2003). This model has innovative features such as state dependent correlation matrices, fat tailed asset return distributions, simple computational schemes and output.

InnoALM was produced in six months during 2000 with Geyer and Ziemba serving as consultants and Herold and Kontriner being Innovest employees. InnoALM demonstrates that a small team of researchers with a limited budget can quickly produce a valuable modeling system that can easily be operated by non-stochastic programming specialists on a single PC. The IBM OSL stochastic programming software provides a good solver. The solver was interfaced with user friendly input and output capabilities. Calculation times on the PC are such that different modeling situations can be easily developed and the implications of policy, scenario, and other changes seen quickly. The graphical output provides pension fund management with essential information to aid in the making of informed investment decisions and understand the probable outcomes and risk involved with these actions. The model can be used to explore possible European, Austrian and Innovest policy alternatives.

The liability side of the Siemens Pension Plan consists of employees, for whom Siemens is contributing DCP payments, and retired employees who receive pension payments. Contributions are based on a fixed fraction of salaries, which varies across employees. Active employees are assumed to be in steady state; so employees are replaced by a new employee with the same qualification and sex so there is a constant number of similar employees. Newly employed staff start with less salary than retired staff, which implies that total contributions grow less rapidly than individual salaries. The set of retired employees is modeled using Austrian mortality and marital tables. Widows receive 60% of the pension payments. Retired employees receive pension payments after reaching age 65 for men and 60 for women. Payments to retired employees are based upon the individually accumulated contribution and the fund performance during active employment. The annual pension payments are based on a discount rate of 6% and the remaining life expectancy at the time of retirement. These annuities grow by 1.5% annually to compensate for inflation. Hence, the wealth of the pension fund must grow by 7.5% per year to match liability commitments. Another output of the computations is the expected annual net cash flow of plan contributions minus payments. Since the number of pensioners is rising faster than plan contributions, these cash flows are negative so the plan is declining in size.

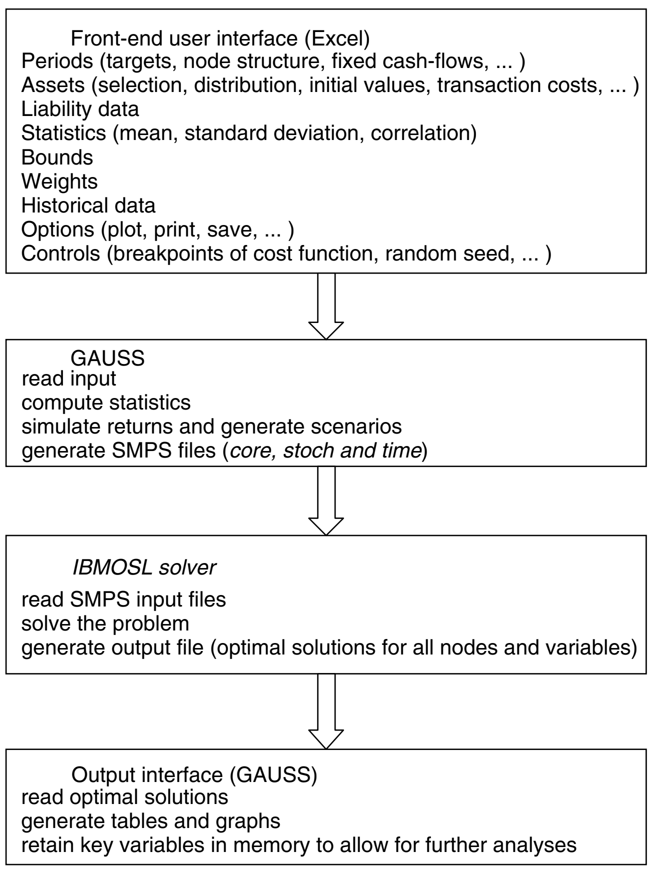


Figure 2: Elements of InnoALM. Source: Geyer *et al*. (2003)

The model determines the optimal purchases and sales for each of N assets in each of T planning periods. Typical asset classes used at Innovest are US, Pacific, European, and Emerging Market equities and US, UK, Japanese and European bonds. The objective is to maximize the concave risk averse utility function expected terminal wealth less convex penalty costs subject to various linear constraints. The effect of such constraints is evaluated in the examples that follow, including Austria's limits of 40% maximum in equities, 45% maximum in foreign securities, and 40% minimum in Eurobonds. The convex risk measure is approximated by a piecewise linear function so the model is a multiperiod stochastic linear program. Typical targets that the model tries to achieve, and if not is penalized for, are wealth (the fund's assets) to grow by 7.5% per year and for portfolio performance returns to exceed benchmarks. Excess wealth is placed into surplus reserves and a portion of that is paid out in succeeding years.

The elements of InnoALM are described in Figure 2. The interface to read in data and problem elements uses Excel. Statistical calculations use the program Gauss and this data is fed into the IBM0SL solver which generates the optimal solution to the stochastic program. The output used Gauss to generate various tables and graphs and retains key variables in memory to allow for future modeling calculations. Details of the model formulation are in Geyer *et al*. (2003).

3.1 Some typical applications

To illustrate the model's use we present results for a problem with four asset classes (Stocks Europe, Stocks US, Bonds Europe, and Bonds US) with five periods (six stages). The periods are twice 1 year, twice 2 years and 4 years (10 years in total). We assume discrete compounding which implies that the mean return for asset i (μ_i) used in simulations is $\mu_i = exp(\overline{y})_i - 1$ where $\overline{y}_i$ is the mean based on log returns. We generate 10 000 scenarios using a 100-5-5-2-2 node structure. Initial wealth equals 100 units and the wealth target is assumed to grow at an annual rate of 7.5%. No benchmark target and no cash in- and outflows are considered in this sample application to make its results more general. We use risk aversion $R_A = 4$ and the discount factor equals 5%, which corresponds roughly with a simple static mean-variance model to a standard 60-40 stock-bond pension fund mix; see Kallberg and Ziemba (1983).

Assumptions about the statistical properties of returns measured in nominal Euros are based on a sample of monthly data from January 1970 for stocks and 1986 for bonds to September 2000. Summary statistics for monthly and annual log returns are in Table 2. The US and European equity means for the longer period 1970–2000 are much lower than for 1986–2000 and slightly less volatile. The monthly stock returns are non-normal and negatively skewed. Monthly stock returns are fat tailed whereas monthly bond returns are close to normal (the critical value of the Jarque–Bera test for a = 0.01 is 9.2).

However, for long-term planning models such as InnoALM with its one year review period, properties of monthly returns are less relevant. The bottom panel of Table 2 contains statistics for annual returns. While average returns and volatilities remain about the same (we lose one year of data when we compute annual returns), the distributional properties change dramatically. While we still find negative skewness, there is no evidence for fat tails in annual returns except for European stocks (1970–2000) and US bonds.

The mean returns from this sample are comparable to the 1900–2000 one hundred and one year mean returns estimated by Dimson *et al*. (2002). Their estimate of the nominal mean equity return for the US is 12.0% and that for Germany and UK is 13.6% (the simple average of the two countries' means). The mean of bond returns is 5.1% for US and 5.4% for Germany and UK.

Assumptions about means, standard deviations and correlations for the applications of InnoALM appear in Table 4 and are based on the sample statistics presented in Table 3. Projecting future rates of returns from past data is difficult. We use the equity means from the period

TABLE 2: STATISTICAL PROPERTIES OF ASSET RETURNS

	Stocks Eur		Stocks US		Bonds Eur	Bonds US
Monthly returns	1/70 –9/00	1/86 –9/00	1/70 –9/0	1/86 –9/00	1/86 –9/00	1/86 –9/00
Mean (% p.a.)	10.6	13.3	10.7	14.8	6.5	7.2
Std.dev (% p.a.)	16.1	17.4	19.0	20.2	3.7	11.3
Skewness	−0.90	−1.43	−0.72	−1.04	−0.50	0.52
Kurtosis	7.05	8.43	5.79	7.09	3.25	3.30
Jarque–Bera test	302.6	277.3	151.9	155.6	7.7	8.5
Annual returns						
Mean (%)	11.1	13.3	11.0	15.2	6.5	6.9
Std.dev (%)	17.2	16.2	20.1	18.4	4.8	12.1
Skewness	−0.53	−0.10	−0.23	−0.28	−0.20	−0.42
Kurtosis	3.23	2.28	2.56	2.45	2.25	2.26
Jarque–Bera test	17.4	3.9	6.2	4.2	5.0	8.7

Source: Geyer *et al*. (2003)

1970–2000 since 1986–2000 had exceptionally good performance of stocks which is not assumed to prevail in the long run.

TABLE 3: REGRESSION EQUATIONS RELATING ASSET CORRELATIONS AND US STOCK RETURN VOLATILITY (MONTHLY RETURNS; JAN 1989-SEP 2000; 141 OBSERVATIONS)

Correlation between	Constant	Slope w.r.t. US stock volatility	t-Statistic of slope	R
Stocks Europe–Stocks US	0.62	2.7	6.5	0.23
Stocks Europe–Bonds Europe	1.05	−14.4	−16.9	0.67
Stocks Europe–Bonds US	0.86	−7.0	−9.7	0.40
Stocks US–Bonds Europe	1.11	−16.5	−25.2	0.82
Stocks US–Bonds US	1.07	−5.7	−11.2	0.48
Bonds Europe–Bonds US	1.10	−15.4	−12.8	0.54

Source: Geyer *et al*. (2003)

The correlation matrices in Table 4 for the three different regimes are based on the regression approach of Solnik *et al*. (1996). Moving average estimates of correlations among all assets are functions of standard deviations of US equity returns. The estimated regression equations are then used to predict the correlations in the three regimes shown in Table 4. Results for the estimated regression equations appear in Table 3. Three regimes are considered and it is assumed that 10% of the time, equity markets are extremely volatile, 20% of the time markets are characterized

TABLE 4: MEANS, STANDARD DEVIATIONS AND CORRELATIONS ASSUMPTIONS

		Stocks Europe	Stocks US	Bonds Europe	Bonds US
normal periods (70% of the time)	Stocks US	0.755			
	Bonds Europe	0.334	0.286		
	Bonds US	0.514	0.780	0.333	
	Standard deviation	14.6	17.3	3.3	10.9
high volatility (20% of the time)	Stocks US	0.786			
	Bonds Europe	171	0.100		
	Bonds US	0.435	0.715	0.159	
	Standard deviation	19.2	21.1	4.1	12.4
extreme periods (10% of the time)	Stocks US	0.832			
	Bonds Europe	−0.075	−0.182		
	Bonds US	0.315	0.618	−0.104	
	Standard deviation	21.7	27.1	4.4	12.9
average period	Stocks US	0.769			
	Bonds Europe	0.261	0.202		
	Bonds US	0.478	0.751	0.255	
	Standard deviation	16.4	19.3	3.6	11.4
all periods	Mean	10.6	10.7	6.5	7.2

Source: Geyer *et al.* (2003)

by high volatility and 70% of the time, markets are normal. The 35% quantile of US equity return volatility defines *normal* periods. *Highly volatile* periods are based on the 80% volatility quantile and *extreme* periods on the 95% quartile. The associated correlations reflect the return relationships that typically prevailed during those market conditions. The correlations in Table 4 show a distinct pattern across the three regimes. Correlations among stocks increase as stock return volatility rises, whereas the correlations between stocks and bonds tend to decrease. European bonds may serve as a hedge for equities during extremely volatile periods since bonds and stocks returns, which are usually positively correlated, are then negatively correlated. The latter is a major reason why using scenario dependent correlation matrices is a major advance over sensitivity of one correlation matrix.

Optimal portfolios were calculated for seven cases—with and without mixing of correlations and with normal, t- and historical distributions. Cases NM, HM and TM use mixing correlations. Case NM assumes normal distributions for all assets. Case HM uses the historical distributions of each asset. Case TM assumes t-distributions with five degrees of freedom for stock returns, whereas bond returns are assumed to have normal distributions. The cases NA, HA and TA use the same distribution assumptions with no mixing of correlations matrices. Instead the correlations and standard deviations used in these cases correspond to an ‘average’ period where 10%, 20% and 70% weights are used to compute averages of correlations and standard deviations used in the three different regimes. Comparisons of the average (A) cases and mixing (M) cases are mainly intended to investigate the effect of mixing correlations. TMC maintains all assumptions of case TM but uses Austria’s constraints on asset weights that Eurobonds must be at least 40% and equity at most 40%, and these constraints are binding.

3.2 Some test results

Table 5 shows the optimal initial asset weights at stage 1 for the various cases. Table 6 shows results for the final stage (expected weights, expected terminal wealth, expected reserves and shortfall probabilities). These tables show that the mixing correlation cases initially assign a much lower weight to European bonds than the average period cases. Single-period, mean-variance optimization and the average period cases (NA, HA and TA) suggest an approximate 45–55 mix between equities and bonds. The mixing correlation cases (NM, HM and TM) imply a 65-35 mix. Investing in US Bonds is not optimal at stage 1 in any of the cases which seems due to the relatively high volatility of US bonds.

TABLE 5: OPTIMAL INITIAL ASSET WEIGHTS AT STAGE 1 BY CASE (PERCENTAGE)

	Stocks Europe	Stocks US	Bonds Europe	Bonds US
Single-period, mean-variance optimal weights (average periods)	34.8	9.6	55.6	0.0
Case NA: no mixing (average periods) normal distributions	27.2	10.5	62.3	0.0
Case HA: no mixing (average periods) historical distributions	40.0	4.1	55.9	0.0
Case TA: no mixing (average periods) t-distributions for stocks	44.2	1.1	54.7	0.0
Case NM: mixing correlations normal distributions	47.0	27.6	25.4	0.0
Case HM: mixing correlations historical distributions	37.9	25.2	36.8	0.0
Case TM: mixing correlations t-distributions for stocks	53.4	11.1	35.5	0.0
Case TMC: mixing correlations historical distributions; constraints on asset weights	35.1	4.9	60.0	0.0

Source: Geyer *et al*. (2003)

Table 6 shows that the distinction between the A and M cases becomes less pronounced over time. However, European equities still have a consistently higher weight in the mixing cases than in no-mixing cases. This higher weight is mainly at the expense of Eurobonds. In general the proportion of equities at the final stage is much higher than in the first stage. This may be explained by the fact that the expected portfolio wealth at later stages is far above the target wealth level (206.1 at stage 6) and the higher risk associated with stocks is less important. The constraints in case TMC lead to lower expected portfolio wealth throughout the horizon and to a higher shortfall probability than any other case. Calculations show that initial wealth would have

TABLE 6: EXPECTED PORTFOLIO WEIGHTS AT THE FINAL STAGE BY CASE (PERCENTAGE), EXPECTED TERMINAL WEALTH, EXPECTED RESERVES, AND THE PROBABILITY FOR WEALTH TARGET SHORTFALLS (PERCENTAGE) AT THE FINAL STAGE

	Stocks Europe	Stocks US	Bonds Europe	Bonds US	Expected terminal wealth	Expected reserves at stage 6	Probability of target shortfall
NA	34.3	49.6	11.7	4.4	328.9	202.8	11.2
HA	33.5	48.1	13.6	4.8	328.9	205.2	13.7
TA	35.5	50.2	11.4	2.9	327.9	202.2	10.9
NM	38.0	49.7	8.3	4.0	349.8	240.1	9.3
HM	39.3	46.9	10.1	3.7	349.1	235.2	10.0
TM	38.1	51.5	7.4	2.9	342.8	226.6	8.3
TMC	20.4	20.8	46.3	12.4	253.1	86.9	16.1

Source: Geyer *et al*. (2003)

to be 35% higher to compensate for the loss in terminal expected wealth due to those constraints. In all cases the optimal weight of equities is much higher than the historical 4.1% in Austria.

The expected terminal wealth levels and the shortfall probabilities at the final stage shown in Table 6 make the difference between mixing and no-mixing cases even clearer. Mixing correlations yields higher levels of terminal wealth and lower shortfall probabilities.

If the level of portfolio wealth exceeds the target, the surplus $\widetilde{D}_j$ is allocated to a reserve account. The reserves in t are computed from $\sum_{j=1}^{t} \widetilde{D}_j$ and as shown in Table 6 for the final stage. These values are in monetary units given an initial wealth level of 100. They can be compared to the wealth target 206.1 at stage 6. Expected reserves exceed the target level at the final stage by up to 16%. Depending on the scenario the reserves can be as high as 1800. Their standard deviation (across scenarios) ranges from 5 at the first stage to 200 at the final stage. The constraints in case TMC lead to a much lower level of reserves compared to the other cases which implies, in fact, less security against future increases of pension payments.

Summarizing we find that optimal allocations, expected wealth and shortfall probabilities are mainly affected by considering mixing correlations while the type of distribution chosen has a smaller impact. This distinction is mainly due to the higher proportion allocated to equities if different market conditions are taken into account by mixing correlations.

The results of any asset allocation strategy crucially depend upon the mean returns. This effect is now investigated by parametrizing the forecasted future means of equity returns. Assume that an econometric model forecasts that the future mean return for US equities is some value between 5 and 15%. The mean of European equities is adjusted accordingly so that the ratio of equity means and the mean bond returns as in Table 4 are maintained. We retain all other assumptions of case NM (normal distribution and mixing correlations). Figure 3 summarizes the effects of these mean changes in terms of the optimal initial weights. As expected, see Chopra and Ziemba (1993) and Kallberg and Ziemba (1981, 1984), the results are very sensitive to the choice of the mean return. If the mean return for US stocks is assumed to equal the long run mean of 12% as estimated by Dimson *et al*. (2002), the model yields an optimal weight for equities of 100%. However, a mean return for US stocks of 9% implies less than 30% optimal weight for equities.

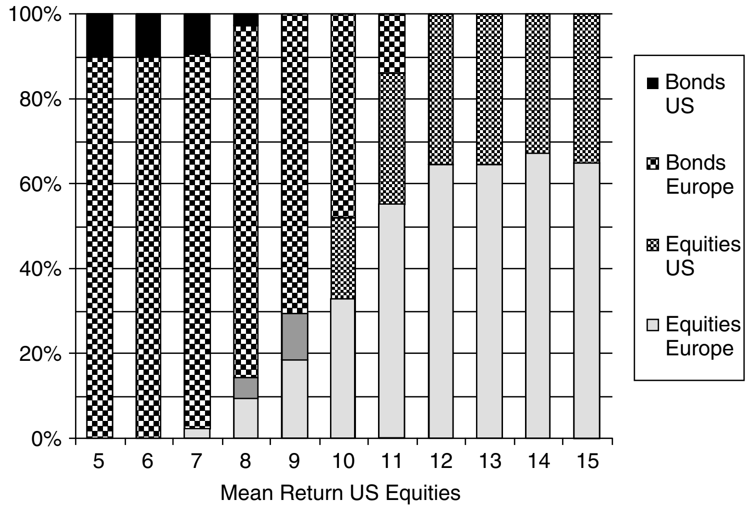


Figure 3: Optimal asset weights at stage 1 for varying levels of US equity means. Source: Geyer *et al*. (2003)

3.3 Model tests

Since state dependent correlations have a significant impact on allocation decisions it is worthwhile to further investigate their nature and implications from the perspective of testing the model. Positive effects on the pension fund performance induced by the stochastic, multiperiod planning approach will only be realized if the portfolio is dynamically rebalanced as implied by the optimal scenario tree. The performance of the model is tested considering this aspect. As a starting point it is instructive to break down the rebalancing decisions at later stages into groups of achieved wealth levels. This reveals the 'decision rule' implied by the model depending on the current state. Consider case TM. Quintiles of wealth are formed at stage 2 and the average optimal weights assigned to each quintile are computed. The same is done using quintiles of wealth at stage 5.

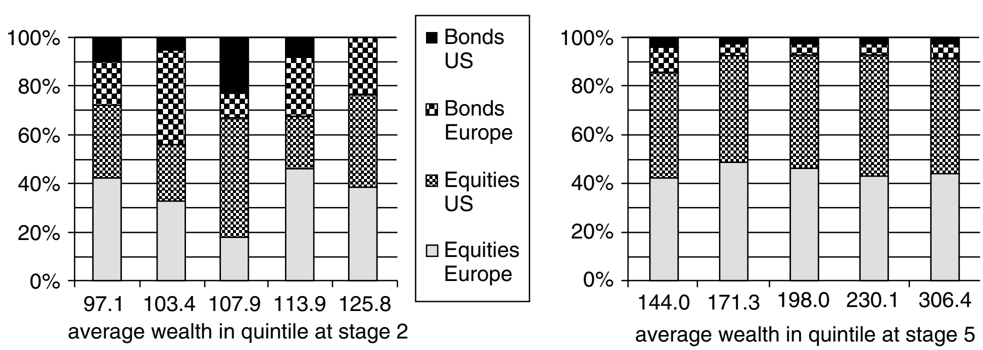


Figure 4: Optimal weights conditional on quintiles of portfolio wealth at stage 2 and 5. Source: Geyer *et al*. (2003)

Figure 4 shows the distribution of weights for each of the five average levels of wealth at the two stages. While the average allocation at stage 5 is essentially independent of the wealth level achieved (the target wealth at stage 5 is 154.3), the distribution at stage 2 depends on the wealth level in a specific way. If average attained wealth is 103.4, which is slightly below the target, a very cautious strategy is chosen. Bonds have the highest weight in this case (almost 50%). In this situation the model implies that the risk of even stronger underachievement of the target is to be minimized. The model relies on the low but more certain expected returns of bonds to move back to the target level. If attained wealth is far below the target (97.1) the model implies more than 70% equities and a high share (10.9%) of relatively risky US bonds. With such strong underachievement there is no room for a cautious strategy to attain the target level again. If average attained wealth equals 107.9, which is close to the target wealth of 107.5, the highest proportion is invested in US assets with 49.6% invested in equities and 22.8% in bonds. The US assets are more risky than the corresponding European assets which is acceptable because portfolio wealth is very close to the target and risk does not play a big role. For wealth levels above the target most of the portfolio is switched to European assets which are safer than US assets. This 'decision' may be interpreted as an attempt to preserve the high levels of attained wealth. The decision rules implied by the optimal solution can be used to perform a test of the model using the following rebalancing strategy. Consider the ten-year period from January 1992 to January 2002. In the first month of this period we assume that wealth is allocated according to the optimal solution for stage 1 given in Table 5. In each of the subsequent months the portfolio is rebalanced as follows: identify the current volatility regime (extreme, highly volatile, or normal) based on the observed US stock return volatility. Then search the scenario tree to find a node that corresponds to the current volatility regime and has the same or a similar level of wealth. The optimal weights from that node determine the rebalancing decision. For the no-mixing cases (NA, TA and HA) the information about the current volatility regime cannot be used to identify optimal weights. In those cases use the weights from a node with a level of wealth as close as possible to the current level of wealth. Table 7 presents summary statistics for the complete sample and the out-of-sample period October 2000 to January 2002. The mixing correlation solutions assuming normal and t-distributions (cases NM and TM) provide a higher average return with lower standard deviation than the corresponding non-mixing cases (NA and TA). The advantage may be substantial as indicated by the 14.9% average return of TM compared to 10.0% for TA. The t-statistic for this difference is 1.7 and is significant at the 5% level (one-sided test). Using the historical distribution and mixing correlations (HM) yields a lower average return than no-mixing (HA). In the constrained case (TMC) the average return for the complete sample is in the same range as for the unconstrained cases. This is mainly due to relatively high weights assigned to US bonds which performed very well during the test period, whereas stocks performed poorly. The standard deviation of returns is much lower because the constraints imply a lower degree of rebalancing.

To emphasize the difference between the cases TM and TA Figure 5 compares the cumulated monthly returns obtained from the rebalancing strategy for the two cases as well as a buy and hold strategy which assumes that the portfolio weights on January 1992 are fixed at the optimal TM weights throughout the test period. Rebalancing on the basis of the optimal TM scenario tree provides a substantial gain when compared to the buy and hold strategy or the performance using TA results, where rebalancing does not account for different correlation and volatility regimes.

Such in- and out-of-sample comparisons depend on the asset returns and test period. To isolate the potential benefits from considering state dependent correlations the following controlled

TABLE 7: RESULTS OF ASSET ALLOCATION STRATEGIES USING THE *DECISION RULE* IMPLIED BY THE OPTIMAL SCENARIO TREE

	Complete sample 01/92–01/02		Out-of-sample 10/00–01/02	
	Mean	Std.dev.	Mean	Std.dev.
NA	11.6	16.1	−17.1	18.6
NM	13.1	15.5	−9.6	16.9
HA	12.6	16.5	−15.7	21.1
HM	11.8	16.5	−15.8	19.3
TA	10.0	16.0	−14.6	18.9
TM	14.9	15.9	−10.8	17.6
TMC	12.4	8.5	0.6	9.9

Source: Geyer *et al*. (2003)

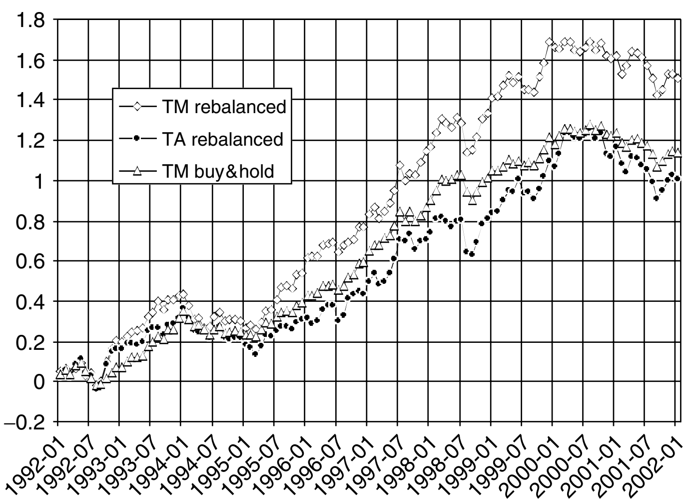


Figure 5: Cumulative monthly returns for different strategies. Source: Geyer *et al*. (2003)

simulation experiment was performed. Consider 1000 ten-year periods where simulated annual returns of the four assets are assumed to have the statistical properties summarized in Table 4. One of the ten years is assumed to be an 'extreme' year, two years correspond to 'highly volatile' markets and seven years are 'normal' years. We compare the average annual return of two strategies: (a) a buy and hold strategy using the optimal TM weights from Table 5 throughout the ten-year period, and (b) a rebalancing strategy that uses the implied decision rules of the optimal scenario tree as explained in the in- and out-of-sample tests above. For simplicity it was assumed that the current volatility regime is known in each period. The average annual returns over 1000 repetitions of the two strategies are 9.8% (rebalancing) and 9.2% (buy and hold). The t-statistic for the mean difference is 5.4 and indicates a highly significant advantage of the rebalancing strategy which exploits the information about state dependent correlations. For comparison the same experiment was repeated using the optimal weights from the constrained case TMC. We obtain the same average mean of 8.1% for both strategies. This indicates that the constraints imply insufficient rebalancing capacity. Therefore knowledge about the volatility regime cannot be sufficiently exploited to achieve superior performance relative to buy and hold. This result also shows that the relatively good performance of the TMC rebalancing strategy in the sample period 1992–2002 is positively biased by the favorable conditions during that time.

REFERENCES

■ Blake, D., Lehmann, B. N. and Timmermann, A. (1999) Asset allocation dynamics and pension fund performance. *Journal of Business*, 72, 429–461.

■ Brinson, B., Hood, L. R. and Beebower, G. L. (1986) Determinants of portfolio performance. *Financial Analysts Journal*, 42, 39–45.

■ Cariño, D., Myers, R. and Ziemba, W. T. (1998a) Concepts, technical issues and uses of the Russell–Yasuda Kasai financial planning model. *Operations Research*, 46, 450–462.

■ Cariño, D. and Ziemba, W. T. (1998b) Formulation of Russell–Yasuda Kasai financial planning model. *Operations Research*, 46, 433–449.

■ Cariño, D. R., Kent, T., Myers, D. H., Stacy, C., Sylvanus, M., Turner, A. L., Watanabe, K. and Ziemba, W. T. (1994) The Russell-Yasuda Kasai model: an asset/liability model for a Japanese insurance company using multistage stochastic programming. *Interfaces*, 24, 29–49.

■ Chopra, V. K. and Ziemba, W. T. (1993) The effect of errors in mean, variance and co-variance estimates on optimal portfolio choice. *Journal of Portfolio Management*, 19, 6–11.

■ Dempster, M. A. H., Evstigneev, I. V. and Schenk-Hoppé, K. R. (2003) Exponential growth of fixed mix assets in stationary markets. *Finance and Stochastics*, 7(2), 263–276.

■ Dimson, E., Marsh, P. and Staunton, M. (2002) *Triumph of the Optimists: 101 Years of Global Investment Returns*. Princeton University Press, Princeton.

■ Geyer, A., Herold, W., Kontriner, K. and Ziemba, W. T. (2003) The Innovest Austrian Pension fund Financial Planning Model InnoALM. Working paper, UBC.

■ Hensel, C. R., Ezra, D. D. and Ilkiw, J. H. (1991) The importance of the asset allocation decision. *Financial Analysts Journal*, July/August, 65–72.

■ Kallberg, J. G., White, R. and Ziemba, W. T. (1982) Short term financial planning under uncertainty. *Management Science*, XXVIII, 670–682.

■ Kallberg, J. G. and Ziemba, W. T. (1984) Mis-specifications in portfolio selection problems. In G. Bamberg and A. Spremann (eds), *Risk and Capital*, pp. 74–87. Springer Verlag, New York.

■ Kallberg, J. G. and Ziemba, W. T. (1983) Comparison of alternative utility functions in portfolio selection problems. *Management Science*, 29(11), 1257–1276.

■ Kusy, M. I. and Ziemba, W. T. (1986) A bank asset and liability management model. *Operations Research*, 34(3), 356–376.

■ Lakonishok, J., Shleifer, A. and Vishny, R. (1992) The structure and performance of the money management industry. *Brookings Papers on Economic Activity Microeconomics*, 229–91.

■ Merton, R. C. (1990). *Continuous-Time Finance*. Blackwell Publishers, Cambridge, Mass.

■ Solnik, B., Boucrelle, C. and Le Fur, Y. (1996) International market correlation and volatility. *Financial Analysts Journal*, 52, 17–34.
■ Ziemba, W. T. (2003) *The Stochastic Programming Approach to Asset Liability and Wealth Management*. AIMR, Charlottesville, VA.
■ Ziemba, W. T. and Mulvey, J. M. (eds) (1998) *World Wide Asset and Liability Modeling*. Cambridge University Press.

8
Efficient Estimates for Valuing American Options

Mike Staunton*

Although the search for an exact solution to valuing American options under Black-Scholes dynamics continues, recent developments in analytic approximations and numerical methods now allow errors of a tiny order of magnitude to be achieved.

We compare the two best base approximations—the analytic approximation of Ju and Zhong (1999) and the combination of a curved exercise boundary and the integral equation of Ju (1998)—against a wide range of grid and lattice methods. Where possible, the comparison uses both Richardson extrapolation (following Leisen 1998) and curtailed ranges (following Andricopoulos *et al*. 2004). Efficiency is judged by the trade-off between accuracy and calculation speed. Some, though not all, choices of parameters for binomial trees in addition to finite difference methods display uniform convergence with increasing numbers of time steps and this can allow the use of Richardson extrapolation to improve accuracy. The more familiar choices of binomial parameters such as Cox, Ross and Rubinstein or Jarrow and Rudd exhibit oscillation and hence cannot take advantage of extrapolation. The close connection between trinomial trees and explicit finite differences is well known but the key discriminator between the various grid and lattice methods should be the avoidance of calculations that contribute almost nothing to option value. By curtailing the range of binomial trees, we make its shape resemble that of a grid and the resulting reductions in calculation speed can be substantial while preserving the accuracy of the full tree.

1 Richardson extrapolation

Extrapolation has been a common theme in improving analytic approximations for American option values ever since Geske and Johnson (1984) used a sequence of European then Bermudan option values with 1, 2, 3 and 4 exercise points. In the most recent use, Ju combined option values

*E-mail: mstaunton@london.edu

assuming that the early exercise boundary could be estimated as a multipiece exponential function with 1, 2, 3 and 4 pieces and the resulting sequence of option values then used for extrapolation. Leisen was the first to combine extrapolation with binomial trees and used 2 points based on $2n$ steps (with weight 2) and n steps (with weight -1). The proper weights depend on the order of convergence, which for American puts we assume to be 1 (the regression slope coefficients for log error on log steps are -1.01 for LR trees, -1.12 for EFD and -0.95 for IFD).

2 Methodology

We follow the set of 16 options chosen by Andricopoulos *et al*. ($S = 100$, $q = 0\%$, $K = 95/105$, $r = 6\%/20\%$, $t = 0.5/1.0$ and $\sigma = 20\%/40\%$) and use the root mean squared error (RMSE) as our measure of accuracy. Our experience is that the rankings of RMSE between the different numerical methods is sufficiently close to those from using far more option values (such as the 1250 deterministic or random options typically chosen in other comparisons), helped by the relatively large variation in volatility for the 16 options.

The choice of 'true' value should have an error some orders of magnitude lower than that of the numerical methods used in any comparison. An all-too common choice is that of the CRR binomial tree with 10 000 time steps. However, this choice is insufficiently accurate for our purposes, and so we prefer to extrapolate from the estimates of LR trees with 4999 and 9999 steps. When measured against this benchmark, the CRR tree with 10 000 steps had an RMSE of 1.28E-4. It should be understood that all RMSE estimates quoted in this chapter assume that the 'true' value is that obtained from the LR2 4999 binomial tree method.

3 The base methods

The chart in Figure 1 confirms why we have chosen JZ and Ju (EXP2, EXP3 and EXP4) as our base methods. We have taken the RMSE and calculation speed for the 16 put options of the

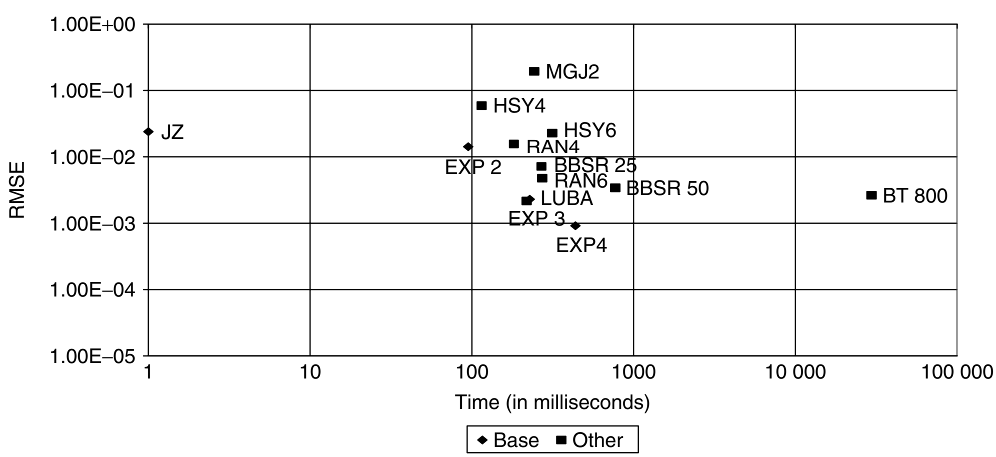


Figure 1: Previous comparisons

base methods using VBA code on a Pentium 4 2.66 GHz with 512 MB of RAM, and then used EXP3 to scale the previous results from the paper by Ju. The chart is a simple X-Y plot where both axes are given a logarithmic scale. Points at the top left of the chart are very quick but less accurate while movements along the diagonal towards the bottom right achieve greater accuracy but are getting progressively slower. A perfect method would plot in the bottom left corner but it is more worthwhile instead to seek undominated methods that have no better methods to their left or below them. You can see that the JZ, EXP2, EXP3 and EXP4 dominate all the other methods chosen by Ju and hence form our base methods. Ju chose comparator methods with a similar accuracy and speed to his three approaches apart from the sole CRR binomial tree with 800 steps. His methods are clustered with computation speed between 100 and 1000 milliseconds combined with RMSE almost always above 1.00E-3. We have added the much quicker JZ approach from their subsequent paper and, for ease of comparison, used the same scales for all the charts in this chapter.

4 Explicit finite differences

The chart in Figure 2 compares the explicit finite difference method, without (EFD) and with two-point extrapolation using n and $2n$ time steps (EFD2). We have chosen the values of λ^2 (where $\Delta x = \sigma\lambda\sqrt{\Delta t}$) to minimise the errors with 1499 time steps and these were 1.71 for EFD and 1.87 for EFD2. Despite this optimisation, the explicit finite difference method performs relatively poorly against the base methods. The EFD with 99 steps is on a par with EXP2 but the EFD with 999 steps has a slightly lower error than EXP4 but is seven times slower. There is no gain from extrapolation as the extra calculation time outweighs any improvement in accuracy and the EFD2 with 1499 steps has a slightly lower error than EXP4 but is 25 times slower.

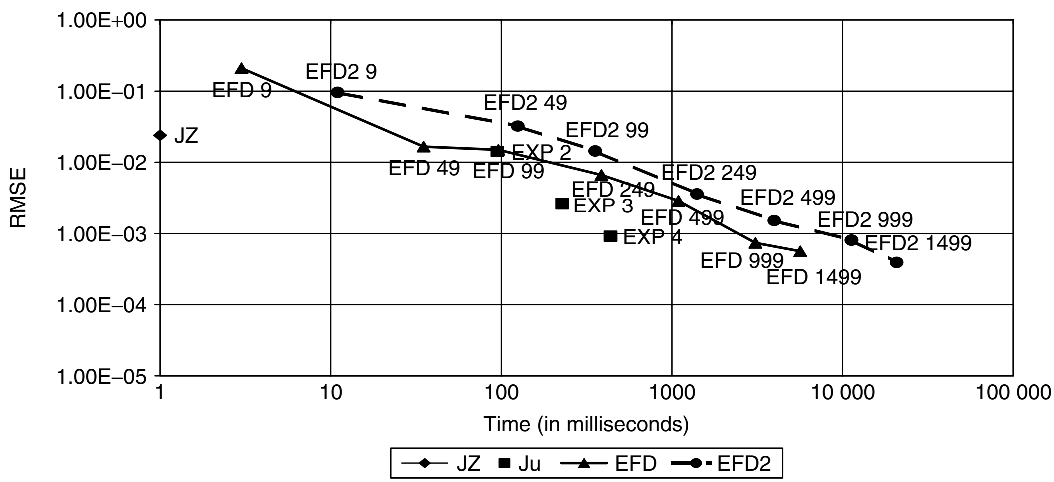


Figure 2: Explicit finite difference

5 Implicit finite differences

The chart in Figure 3 compares the implicit finite difference method, again without (IFD) and with two-point extrapolation (IFD2). The values of λ^2 that minimised the errors with 1499 time steps were 2.17 for IFD and 2.08 for IFD2. The implicit finite difference methods perform worse than their explicit finite difference counterparts, though there are gains from extrapolation. The IFD with 499 steps has a slightly lower error than EXP2 but is 25 times slower. The IFD2 with 999 steps has a slightly lower error than EXP4 but is 59 times slower.

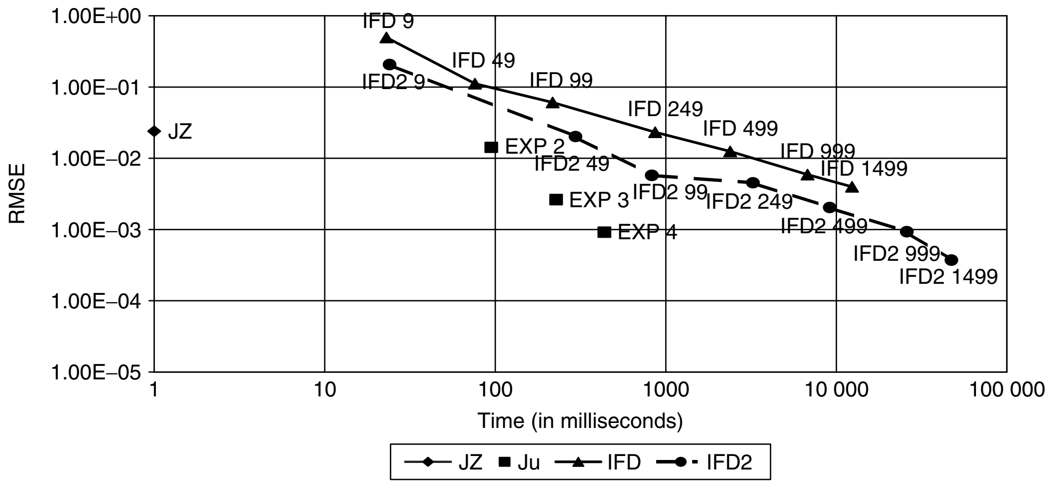


Figure 3: Implicit finite difference

6 The Leisen and Reimer binomial tree

The chart in Figure 4 compares the LR binomial tree method, again without (LR) and with two-point extrapolation (LR2). The LR tree is competitive over a much wider spectrum of computation speeds and is better than EXP2 of the base methods but slightly worse than EXP3.

Extrapolation always improves efficiency and, even excluding the outlier for LR2 with 99 steps, the LR2 method is probably on a par with the EXP3 but worse than EXP4 of the base methods. The LR2 tree with nine steps is only twice as slow as the very quick JZ method but, at the other end of the speed spectrum, the LR2 tree with 1499 steps reached previously uncharted territory with an RMSE of only 2.30E-5.

7 Curtailing the range for binomial trees

The chart in Figure 5 compares the curtailed range LR tree method, again without (LRC) and with two-point extrapolation (LRC2). By curtailing the range (here to within six standard deviations of

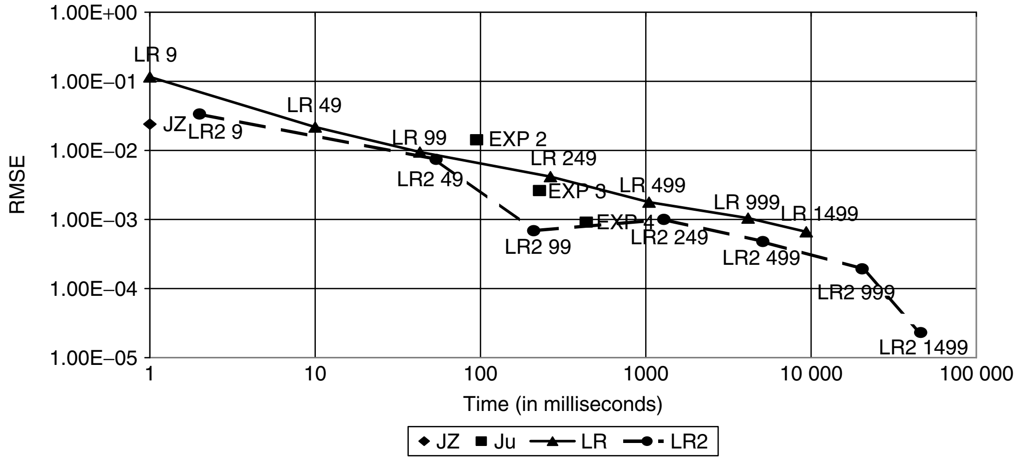


Figure 4: Binomial tree extrapolation

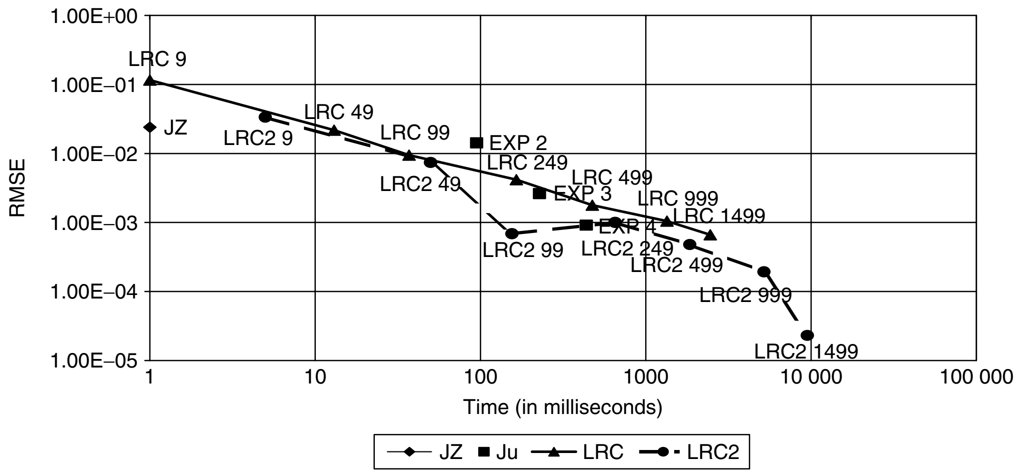


Figure 5: Binomial tree curtailed range

the mean) we preserve the accuracy but reduce calculation speed by between 26% with 99 steps and 79% with 1499 steps. This puts the LRC2 method with 249 steps within striking distance of the EXP4 base method (slightly worse accuracy but taking 1.49 times as long). But we preserve the dominance for methods with the lowest errors, achieving the RMSE of 2.30E-5 in less than 10 000 milliseconds.

8 Conclusions

The drawback of all the methods apart from the grid methods is that they can only value a single option at a time whereas the finite difference methods can value options with a wide range of strike price using only the single grid. In addition there are potentially superior methods (such as Crank-Nicholson) that could improve the efficiency of the finite difference methods. But the LRC2 binomial trees are still likely to be very close to the most efficient methods for valuing American puts, even allowing for this. For example, the LRC2 binomial tree with 499 steps is 11 times quicker than the EFD2 method with 1499 steps and 25 times quicker than the IFD2 method with 1499 steps. What is certainly clear is that all binomial trees should use the LR choice of parameters and be complemented with curtailed ranges and Richardson extrapolation.

REFERENCES

■ Andricopoulos, A. D., Widdicks, M., Duck, P. W. and Newton, D. P. (2003) Universal option valuation using quadrature methods. *Journal of Financial Economics*, 67, 447–471.
■ Andricopoulos, A. D., Widdicks, M., Duck, P. W. and Newton, D. P. (2004) Curtailing the range for lattice and grid methods. *Journal of Derivatives*, 12, Summer, 55–61.
■ Geske, R. and Johnson, H. E. (1984) The American put valued analytically. *Journal of Finance*, 39, 1511–1524.
■ Ju, N. (1998) Pricing an American option by approximating its early exercise boundary as a multipiece exponential function. *Review of Financial Studies*, 11, 627–646.
■ Ju, N. and Zhong, R. (1999) An approximate formula for pricing American options. *Journal of Derivatives*, 7, Winter, 31–34.
■ Leisen, D. P. J. (1998) Pricing the American put option: a detailed convergence analysis for binomial models. *Journal of Economic Dynamics and Control*, 22, 1419–1444.
■ Leisen, D. P. J. and Reimer, M. (1996) Binomial models for option valuation–examining and improving convergence. *Applied Mathematical Finance*, 3, 319–346.

9 Appendix: VBA code for Leisen and Reimer binomial tree

```
Option Base 0

Function VAAmerPutLR#(S#, K#, r#, q#, tyr#, sigma#, nstep%)
’   Returns LR Bin Amer Put Value
    Dim delt#, erdelt#, sigt#, M1#, d2#, c1#, pu#, pd#, c2#
    Dim u#, d#, du#, lnu#, lndu#, lnS#, pudt#, pddt#, Si#
    Dim i%, j%
    Dim vvec#()
    If nstep Mod 2 = 0 Then nstep = nstep + 1
    ReDim vvec(nstep)

    delt = tyr / nstep
    erdelt = Exp(-r * delt)
    sigt = sigma * Sqr(tyr)
    M1 = Exp((r - q) * delt)
    d2 = (Log(S / K) + (r - q - 0.5 * sigma * sigma) * tyr) / sigt
    c1 = d2 / (nstep + 1 / 3 + 0.1 / (nstep + 1))
    pu = 0.5 * (1 + Sgn(d2) * Sqr(1 - Exp(-c1 * c1 * (nstep + 1 / 6))))
    c2 = (d2 + sigt) / (nstep + 1 / 3 + 0.1 / (nstep + 1))
```

```
    pd = 0.5 * (1 + Sgn(d2 + sigt) * Sqr(1 - Exp(-c2 * c2 * (nstep + 1 / 6))))
    u = M1 * pd / pu
    d = M1 * (1 - pd) / (1 - pu)
    du = d / u
    lnu = Log(u)
    lndu = Log(du)
    lnS = Log(S)
    pudt = erdelt * pu
    pddt = erdelt * (1 - pu)

    Si = Exp(lnS + nstep * lnu - lndu)
    For i = 0 To nstep
      Si = Si * du
      vvec(i) = Max(K - Si, 0)
    Next i

    For j = nstep - 1 To 0 Step -1
      Si = Exp(lnS + j * lnu - lndu)
      For i = 0 To j
        Si = Si * du
        vvec(i) = Max(pudt * vvec(i) + pddt * vvec(i + 1), K - Si)
      Next i
    Next j

    VAAmerPutLR = vvec(0)
End Function
```

9
The Relative Valuation of an Equity Price Index[1]

Ruben D. Cohen

A new approach for the relative valuation of an equity price index is presented. The method is based on a coordinate transformation or mapping, which enables one to weigh the index against the aggregated earnings and GDP. This, therefore, gives rise to the notion of relative valuation between the index, the earnings and the GDP. A practical demonstration of this is then provided for the US, UK and Japan economies and some of their major equity indices, namely the S&P500, FTSE100 and TOPIX, respectively.

Another potential application of the above is also discussed, which relates to forecasting the GDP. This stems from the assumption that the expected GDP, one year ahead from today, is readily priced in today's interest rates. The method is further applied to computing duration. This is shown to circumvent the difficulties that are generally associated with calculating the parameter.

1 Introduction

Relative valuation is a generic term that refers to the notion of comparing the price of an asset to the market value of similar assets. In the field of securities investment, the idea has led to important practical tools, which could presumably spot pricing anomalies. Over time, these tools have become instrumental in enabling analysts and investors to make vital decisions on asset allocation.

In equities, the concept separates into two areas—one pertaining to individual equities and the other to indices. The most common methodology for the former is based on comparing certain financial ratios or multiples, such as the price to book value, price to earnings, EBITDA to enterprise value, etc., of the equity in question to those of its peers (see, for instance, Barth *et al*. 1998, D'Mello *et al*. 1991 and Peters 1991). This type of approach, which is largely popular as a strategic tool in the financial industry, is mainly statistical and based on historical data.

For an equity index, however, the above fails mainly because it is difficult to group indices into peer groups. Consequently, relative valuation here is generally carried out by comparing the index's performance to economic and market fundamentals, which may include GDP growth, interest rate and inflation forecasts, as well as earnings growth, among others. This style of comparison is popular among practising economists in their attempt to rationalise the connections between the equity markets and the economy.

The above approach also has its faults, however—one being that, even if the fundamentals were known, there appears to be no consensus methodology, as the procedures that are generally implemented tend to be subjective, ad hoc and dependent on personal style. Thus, it would be useful to devise a new approach to enable one to add some objectivity to the process.

In constructing such a framework here, the classical equity valuation models are first summarised, after which the role of the equity risk premium and how it fits in are clarified. A couple or so simple propositions are then brought in to help facilitate the process. The use of this new method is later demonstrated by (1) suggesting other potential applications, such as forecasting the GDP and calculating duration and (2) incorporating it as a relative-valuation tool. It should be noted that, owing to the nature of the approach, there is no need for any detailed statistical testing, as conclusions can be drawn simply by visual examination of graphs and charts alone.

2 A background on equity valuation

Since the classical models of equity valuation are covered well in the literature, it would be repetitive to discuss them here in any depth. Nevertheless, it is still necessary to go over some of the assumptions and limitations that underlie these models, as they comprise part of the foundation upon which the new model for relative valuation is based.

2.1 The classical models of equity valuation

In the classical theory of equity valuation, three relationships dominate. They are:

$$\frac{S_f(t) - S(t)}{S(t)} + \frac{\delta_f(t)}{S(t)} = R_M(t) \tag{2.1}$$

$$\frac{\delta_f(t) - \delta(t)}{\delta(t)} + \frac{\delta_f(t)}{S(t)} = R_I(t) \tag{2.2}$$

and

$$\frac{E_f(t)}{S(t)} = R_F(t) \tag{2.3}$$

where $S(t)$ and $\delta(t)$, respectively, are the price and dividends at time t, while $S_f(t)$, $\delta_f(t)$ and $E_f(t)$ signify the 'expected' price, dividends and earnings (after interest and tax, but before dividends). These are yearly expectations, generated for one year ahead from today.

With regards to the above, note that, while Equation 2.1 is an identity, with $R_M(t)$ denoting the expected total rate of return, Equations 2.2 and 2.3 represent valuation models, namely, Gordon's Growth Model[2] and the discounted-cash-flow (DCF) relationship,[3,4] respectively, with $R_I(t)$ and $R_F(t)$ being their expected discount rates. The equity risk premium is discussed briefly in the next section, after which the derivation of the relative valuation model will be carried out.

2.2 The equity risk premium

Owing to its importance in the area of equity investment, the equity risk premium has always attracted attention from academics and practitioners. Countless papers have been written so far on the subject, each proposing a reason for why the risk premium should exist, what it depends on and/or how large it should be. Although many of these works present conflicting theories and/or conclusions, all concur unanimously that the risk premium is a result of uncertainties. It is not the concern here to discuss what causes these uncertainties. These uncertainties simply exist, have always been and will remain to be around as long as no one can predict accurately what the future—near-term or far—holds for the economy and markets.

What is relevant here is how does the equity risk premium, as a parameter, get integrated into valuation? By definition, the risk premium is the difference between the rate of return or discount rate, which could be any of the ones appearing in Equations 2.1–2.3 above, and some 'risk-free' rate.[5] As to what discount rate and risk-free rate one should use is another matter, which, again, shall be left out here. Rather, what is important is that under total and unconditional absence of all uncertainty—past, present and future—the risk premium would not exist, so that all the rates of return that appear in Equations 2.1–2.3 become equal to the 'true' risk-free rate, which itself would remain constant and free of volatility.[6] This, therefore, leads to Proposition 1, which may be expressed as:

Proposition 1 In the absence of all uncertainty and change—past, present and going forward—all risk premiums become zero.

Thus, what entails the above proposition is that all arbitrage opportunities between different types of securities disappear. For instance, equity and fixed income instruments will yield the same, as the yield curve flattens and becomes horizontal. In this instance, therefore, all yields will equal b^*, where b^* symbolises the 'true' risk-free rate. Moreover, in the absence of the risk premium, all rates of return (or discount rates) in Equations 2.1–2.3 will also equal b^*.

In addition to the above, the golden rule of economics enters also, so that

$$\frac{d\ln G}{dt} = \left(\frac{\partial \ln G}{\partial t}\right)_{b=b^*=constant} = b^* \tag{2.4}$$

where G is the level of the nominal GDP and b is the interest rate, which is set constant at b^*. Finally, all forecasts in 2.1–2.3 above—i.e. $S_f(t)$, $\delta_f(t)$ and $E_f(t)$—become identical to their real-time counterparts, $S(t+1)$, $\delta(t+1)$ and $E(t+1)$, respectively, realised a year later at $t+1$.

With Proposition 1 in place, Proposition 2 may now be stated as:

Proposition 2 Under Proposition 1, the golden rule applies also to the rate of growth in equity earnings.

Proposition 2 basically unites the golden rule, as it relates to the GDP in Equation 2.4, to equity earnings as well. This is possible under the above circumstances because equity earnings, or profits, comprise a subset of the GDP and, in the absence of arbitrage, all subdivisions within the GDP must yield at the same rate.

Quantitatively, this is expressible by

$$\left(\frac{\partial \ln E}{\partial t}\right)_{b=b^*=constant} = b^* \tag{2.5}$$

where E is the equity earnings. Thus, under Propositions 1 and 2, with all rates of return in 2.1–2.3 being equal to b^*, as well as the forecasts of S, δ and E remaining identical to their real-time counterparts a year later, Equation 2.5 may be applied to 2.3 to give:

$$\left(\frac{\partial \ln E}{\partial t}\right)_{b=b^*} = \left(\frac{\partial \ln S}{\partial t}\right)_{b=b^*} = b^* \tag{2.6}$$

since, in this case, the discount rate, R_F, also equals b^*.

The implication of Equation 2.6, which states that, subject to the conditions imposed above, the golden rule applies as well to the equity price, $S(t)$, is significant. This is because, upon first using the approximation[7]

$$\left(\frac{\partial \ln S}{\partial t}\right)_{b=b^*} \approx \frac{S(t+1) - S(t)}{S(t)} \tag{2.7}$$

then substituting 2.6 and 2.7 into 2.1 and, finally, setting the rate of return, $R_M(t)$, equal to b^*, all in the absence of the risk premium, the dividend yield, $\delta(t+1)/S(t)$, tends to zero. This simply suggests that, in a world with no uncertainty and change, and, hence, no risk premiums, the investor will not demand any dividend yield.[8]

Therefore, do markets pay and/or investors demand a positive dividend yield because of uncertainties? This, inevitably, points to the much debated issue of the dividend puzzle, along with its link to the equity risk premium, both of which will be left out here as they are not relevant to this work, but, nonetheless, whose details may be found elsewhere (Cohen 2002). Notwithstanding, the above conclusions do lead to the next step, which is to develop a model for the relative valuation of an equity price index.

3 A model for the relative valuation of an equity price index

The new model for relative valuation is constructed here in two ways—one focusing on equity (Section 3.1) and the other on the fundamentals, namely GDP and equity earnings (Section 3.2). The latter two occupy the same section because their underlying principles happen to be the same. The final results will then be united to present the relative valuation measures.

3.1 The equity model

Beginning here with Equation 2.6, which states

$$\left(\frac{\partial \ln S}{\partial t}\right)_{b=b^*} = b^* \tag{2.6}$$

it follows that $\ln S$ could be written as a function of time, t, as well as b^*—i.e.:

$$\ln S = \ln S(b^*, t) \tag{3.1}$$

In the above, holding the discount rate constant at b^* clearly imposes a severe constraint on S. This, however, may be relaxed by proceeding as follows. Very briefly, in place of writing

$\ln S(b^*, t)$ as done in 3.1, it shall be expressed as

$$\ln S = \ln S(b, t) \tag{3.2}$$

which generalises S to account for a time-variable discount rate, $b = b(t)$, instead.

The rationale behind Equation 3.2 is that the effects of the market, and the economy in general, on S are presumed to enter separately through two fundamental elements, one which is b and the other which comprises everything else that falls outside the reign of b. As the second variable appears as time, t, it renders Equation 3.2 general and, hence, together with $b(t)$, it should capture all the economic and market effects on the price, S. In other words, expressing S in the form of 3.2 effectively removes all the restrictions imposed on it earlier in Equation 3.1.

In view of the above, the *total* time differential of Equation 3.2, subsequently, becomes:

$$\frac{\Delta \ln S(b, t)}{\Delta t} = \left(\frac{\partial \ln S}{\partial t}\right)_b + \left(\frac{\partial \ln S}{\partial b}\right)_t \frac{\Delta b}{\Delta t} \tag{3.3}$$

where Δ denotes time-wise differential—i.e. $\Delta b \equiv b(t+1) - b(t)$. While the first partial differential—i.e. $(\partial \ln S/\partial t)_b$—has been shown to be equal to b (see Equation 2.6), the second, $(\partial \ln S/\partial b)_t$, is simply the stock duration, which is the sensitivity of the price to changes in b at some given point in time.

Being an ‘exact differential’, therefore, the two components in Equation 3.3 are coupled to each other via:

$$\left(\frac{\partial}{\partial b}\left(\frac{\partial \ln S}{\partial t}\right)_b\right)_t = \left(\frac{\partial}{\partial t}\left(\frac{\partial \ln S}{\partial b}\right)_t\right)_b \tag{3.4}$$

Since, by virtue of 2.6, the left-hand side of the above is 1, the above equation simplifies to:

$$\left(\frac{\partial}{\partial t}\left(\frac{\partial \ln S}{\partial b}\right)_t\right)_b = 1 \tag{3.5}$$

which may be integrated twice to yield a general solution of the form:

$$\ln S = bt + \alpha_0 + \alpha_1 b + \tilde{\Psi}(b)$$

where α_0 and α_1 are integration constants and $\tilde{\Psi}(b)$ a yet unknown function of *only* b.

Alternatively, the above may be recast into:

$$\ln S - bt = \Psi(b) \tag{3.6}$$

where $\Psi(b)$ is another function of b. The latter representation conveniently absorbs $\tilde{\Psi}(b)$, α_0 and $\alpha_1 b$ into a single function, $\Psi(b)$.

It thus follows from 3.6 that plotting the quantity $\ln S - bt$ against b should, in theory, produce a *single* curve, depending *only* on b. This transformation, as a result, brings in all the effects of time on $\ln S - bt$ through b. A schematic illustration of this is presented in Figure 1, where a mapping of S versus b into $\ln S - bt$ versus b is shown to introduce some type of regularity to a relatively disordered graph.[9,10]

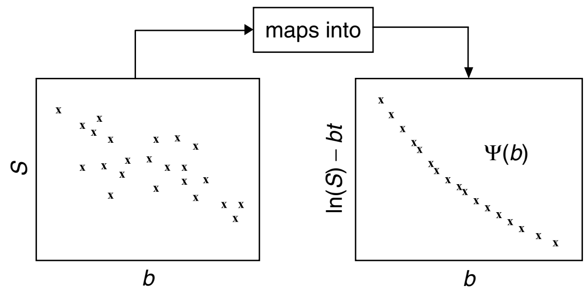


Figure 1: Schematic of the convergence of data points under the proposed coordinate transformation

In light of the derivation so far, it is necessary to mention two points. First, even though Equation 3.6 is extracted from what appears to be too theoretical an approach, it is indeed easy to apply to real situations and, also, as it shall be demonstrated shortly, it does possess other practical uses too. Second, questions relating to what b is—i.e. what interest rate should one use here—have undoubtedly been raised by now. The answer to these, as it will turn out later, happens to be straightforward. Beforehand, however, the same logic is applied next to both the nominal GDP and earnings, as similar transformations are derived.

3.2 Applications to GDP and earnings

It is well accepted that movements in the equity price index are tied closely to corporate earnings and, even more generally, to the economy. Common sense further dictates that a bull market comes typically with a strong economy and a bear market with a weak one. An explanation for this correlation is that the market comprises a subset of the economy—i.e. corporate earnings constitute a (small) fraction of the GDP. This, therefore, should enable one to derive a GDP relationship analogous to the one for equity, as well as for corporate earnings.

Before going into that, however, we need to introduce, with the help of the DCF,[11] a couple of analogies to the equity price index. For this, define V_G and V_E as the 'values' associated with the nominal GDP and corporate earnings, respectively.[12] Therefore, under Propositions 1 and 2, V_G could be represented by

$$V_G \equiv \frac{G_f}{b^*} \tag{3.7a}$$

and V_E by

$$V_E \equiv \frac{E_f}{b^*} \tag{3.7b}$$

where G_f and E_f, respectively, are the time-t expectations of the nominal GDP and corporate earnings one year ahead, at $t + 1$. Hence, with b^* analogous to the discount rate in a 'constant' world, the DCF valuation model is being imposed on the economy as well. It should further be stressed that the one-year-ahead nominal GDP, i.e. $G(t + 1)$, will from now on be implemented instead of the expected for no reason other than convenience, as it shall be assumed that the two converge in an information-efficient economy. For the expected corporate earnings, E_f, on the other hand, Datastream's aggregate I/B/E/S forecasts will be presumed sufficient for the purposes of this work.

Now, with the above analogy in place, it is simple to demonstrate that upon relaxing the constraint on b^* (i.e. replace b^* with b, as it was done in going from Equation 3.1 to 3.2), the same rules that govern the price index should apply as well to V_G and V_E, yielding expressions similar to Equation 3.6, but with V_G and V_E substituted for S. This, consequently, leads to:

$$\ln V_G - bt = \Phi(b) \tag{3.8a}$$

and

$$\ln V_E - bt = \Xi(b) \tag{3.8b}$$

where, as before, $\Phi(b)$ and $\Xi(b)$ are functions of only b.

It should be emphasised that, even though the same transformation that presides over the equity model applies to here as well, the functions $\Phi(b)$ and $\Xi(b)$ may not necessarily be the same as $\Psi(b)$. A comparison of these will be made later; however, certain issues that this raises, namely of the interest rate, 'reversibility' and 'structural or regime shifts', must be addressed beforehand.

3.2.1 Reversibility and structural shifts The representations for the equity price index, GDP and earnings, which are provided in Equations 3.6 to 3.8, lead to the important notions of 'reversibility' and 'structural shifts'. Recognising that structural shifts tend to alter the behaviour of the economy and the markets, an important objective here, as in any economic and financial analysis, would thereby consist of defining ways for detecting and, possibly, classifying them.

To carry this out, observe that $\ln S - bt$, $\ln V_G - bt$ and $\ln V_E - bt$ must depend solely on b via the functions $\Psi(b)$, $\Phi(b)$, and $\Xi(b)$, respectively. The effect of time, as mentioned earlier, enters indirectly through b. Whether or not this functional dependence of Ψ, Φ and Ξ on b is the same in all situations is not of concern now, but, eventually, it shall be dealt with.

An important by-product of such dependence is the concept of 'reversibility', which may be explained via Figure 1 as follows. In reference to this figure, it is noted that, while the unmapped price, S, varies with both b and t and leads to a scattered plot of S versus b, the mapped counterpart changes only with b. This implies that if, for example, the price is S_1 at time t_1, when b equals, let us say, 5%, then at a later time t_2, when b reverts back to 5%, the transformed parameters, $\ln S_1 - bt_1$ and $\ln S_2 - bt_2$, calculated at both times, t_1 and t_2, respectively, must reach the same value again, regardless of the path taken from 1 to 2. This, of course, should apply to V_G and V_E as well, simply by virtue of Equations 3.8a and 3.8b.

Alternatively, a structural or regime shift implies the contrary. If, for instance, a transformed plot produces notably disparate lines, then it is likely that a structural shift has occurred somewhere in between. Schematically, a structural shift is exemplified in Figure 2, where mapping S versus b into $\ln S - bt$ versus b over a given time frame leads to distinctive characteristic patterns. In a similar manner, outliers should, under this type of transformation, appear as shown in Figure 3. Empirical evidence of these phenomena, namely reversibility, regime shifts and outliers, will be provided in Section 4.

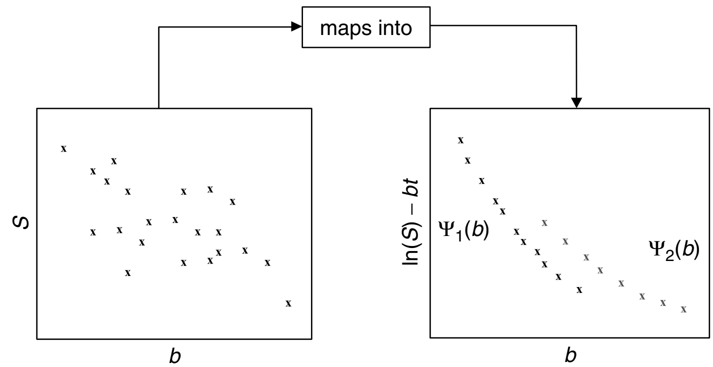


Figure 2: Schematic of how a regime shift manifests itself under the suggested coordinate transformation. A mapping of S versus b into $\ln S - bt$ versus b leads to distinctive characteristic functions, depicted here by $\Psi_1(b)$ and $\Psi_2(b)$, each belonging to a separate regime

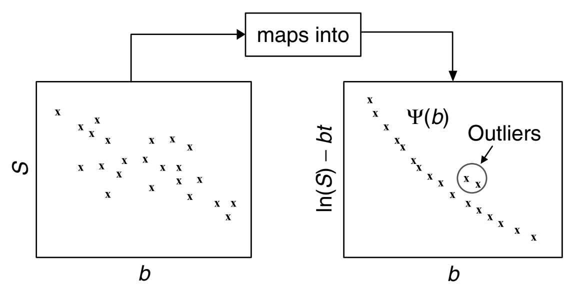


Figure 3: Schematic of how outliers become visible under the suggested coordinate transformation. A mapping of S versus b into $\ln S - bt$ versus b should clearly separate outliers from the function, $\Psi(b)$

3.2.2 The interest rate As mentioned at the end of Section 3.1, the issue of the interest rate is an important one. Putting it more precisely, what should one use for b in Equations 3.6 and 3.8a,b in order to test their validity?

Obviously, several choices exist. These include all the different yields associated with the different, available bond maturities, thus adding to the subjectivity. But, nevertheless, an attempt is made later to settle this point.

Upon following the steps that led to the coordinate transformations in Equations 3.6 and 3.8a,b, it is noted that (bond) maturity or tenor does not enter into the picture. Furthermore, in the context of the reversibility property discussed earlier, it should also not matter which interest rate is used. In other words, using b as the yield of any bond maturity, be it 2 years or 7 years or 30 years, etc., should be acceptable, but only if one moves along a characteristic line, i.e. $\Psi(b)$, $\Phi(b)$, and $\Xi(b)$, which belongs to a certain structural regime. The invariance towards maturity should not be expected to hold across regime shifts and/or to outliers.

4 Evidence of reversibility, outliers and structural shifts

If the hypotheses put forward above were to be proven valid, then upon plotting $\ln X - bt$ against b, where X could signify S, V_G or V_E, one should expect to obtain a single curve, or, more generally, a series of curves, each pertaining to some particular structural regime in the market and/or the economy. Furthermore, it was argued that b could represent the yield associated with any tenor. Examples of each of these, with specific applications to the US, UK and Japan (JP) economies and markets, will be provided in the following sections. Prior to this, however, one must carefully study Table 1, which illustrates how the functions $\Psi(b)$, $\Phi(b)$, and $\Xi(b)$ are calculated.

4.1 Applications to US data

To evaluate the long-run applicability of the model to the US market, refer to Figures 4a,b, where in Figure 4a the S&P price data from 1950 to 2000[13] are plotted both in raw form, as S versus b, and transformed, as $\ln S - bt$ versus b, where b has been chosen to be the 10-year US government bond yield.[14] It is evident here that the raw data, as plotted in Figure 4a, exhibit no regular pattern, whereas the mapped form in Figure 4b definitely displays a convergence that is consistent with theory. A similar conclusion can be derived also from Figures 5a,b, where the aggregated earnings are displayed, both raw and transformed, over the same time period.[15]

Shorter-term, but more detailed, data (quarterly as opposed to annual) for the US, covering from about 1980 to 2004, are presented in Figures 6a–c, where evidence of all the above-mentioned effects, namely convergence, regime shifts and outliers, are clearly depicted. In all instances that follow from now on, the data come from Datastream, using the codes tabulated in Table 2. Also, unless otherwise specified, b will be given by the 10-year government bond yield.

Figures 6a–c present plots of quarterly numbers pertaining to the S&P500 price, I/B/E/S earnings forecast and US GDP, respectively, comparing the raw data against their mapped counterparts. Convergence is noticeable in all cases, although the support is more compelling in the earnings and GDP plots shown in Figures 6b and 6c.

Figure 6a, which pertains to the price index, demonstrates how an outlier, which could otherwise remain hidden in the raw data, stands out in the mapped plane. The outlier highlighted here represents the quarter just before the August 1987 crash, when the overpricing in the S&P500 index, which was then also present in many other national and international indices, led subsequently to the crash.

Figures 6b and 6c, on the other hand, depict structural breaks and regime shifts in the aggregated earnings and GDP. In the interest of objectivity, however, as well as owing to the primary focus of this work, which is to introduce the capabilities of the model rather than guess the causes that could have led to these shifts, there will be no further speculation here. An economist is, perhaps, better suited to undertake this task, by observing the timing of these breaks and connecting them to fundamental (economic and/or market) changes that might have occurred then.

4.2 Applications to UK and JP data

The UK data, concentrating on the FTSE100 price index, aggregated I/B/E/S earnings forecasts and GDP, are presented in Figures 7a–c, respectively. Once again, similar to the US case in

TABLE 1: CALCULATION OF THE FUNCTIONS $\Psi(b)$, $\Phi(b)$ AND $\Xi(b)$ FROM DOWNLOADED US DATA OF THE S&P500 PRICE INDEX, S, 10-YEAR YIELD GOVERNMENT BOND YIELD, b, NOMINAL GDP LEVEL, G, AND THE I/B/E/S AGGREGATE EARNINGS FOR THE INDEX, E_f. DATE OF DOWNLOAD IS FEBRUARY 12, 2004. THE SAMPLE CALCULATIONS IN NOTES 6–10 UNDERNEATH THE TABLE ARE BASED ON Q1 2001

Date	S (1)	b (2)	G (3)	E_f (4)	t (5)	$V_G(b)$ (6)	$V_E(b)$ (7)	$\Psi(b)$ (8)	$\Phi(b)$ (9)	$\Xi(b)$ (10)
Q1 00	1346.09	6.588	9629.4	61.134	19	152167.6	928.0	5.9532	5.2810	5.5813
Q2 00	1437.2	6.556	9822.8	63.983	19.25	153877.4	975.9	6.0084	5.2819	5.6214
Q3 00	1491.72	5.905	9862.1	64.088	19.5	170977.1	1085.3	6.1562	5.4978	5.8382
Q4 00	1367.72	5.696	9953.6	63.668	19.75	178965.9	1117.8	6.0959	5.5700	5.8941
>> **Q1 01**	**1301.53**	**5.223**	10024.8	**60.731**	**20**	**197765.7**	**1162.8**	**6.1267**	**5.7502**	**6.0140**
Q2 01	1291.96	5.398	10088.2	58.173	20.25	193188.2	1077.7	6.0708	5.6783	5.8895
Q3 01	1161.97	4.924	10096.2	56.569	20.5	214094.2	1148.8	6.0485	5.8648	6.0371
Q4 01	1138.65	4.888	10193.9	52.872	20.75	217342.5	1081.7	6.0233	5.8750	5.9720
Q1 02	1104.18	4.919	**10329.3**	54.423	21	218251.7	1106.4	5.9739	5.8604	5.9759
Q2 02	1106.59	5.239	10428.3	56.385	21.25	207037.6	1076.3	5.8958	5.7274	5.8680
Q3 02	928.77	4.29	10542	56.342	21.5	258904.4	1313.3	5.9115	6.1419	6.2580
Q4 02	900.36	4.006	10623.7	55.302	21.75	280736.4	1380.5	5.9315	6.2739	6.3589
Q1 03	851.17	3.93	10735.8	55.078	22	#N/A	1401.5	5.8820	#N/A	6.3807
Q2 03	944.3	3.447	10846.7	56.457	22.25		1637.9	6.0835		6.6342
Q3 03	999.74	4.401	11107	58.581	22.5		1331.1	5.9173		6.2035
Q4 03	1034.15	4.146	11246.3	60.08	22.75		1449.1	5.9981		6.3355
Q1 04	1151.82	4.031	#N/A	62.225	23		1543.7	6.1220		6.4148

(1) S&P500 price index
(2) US 10-year government bond yield.
(3) Nominal level of US GDP.
(4) I/B/E/S earnings forecast.
(5) Reference time, with Q1 00 being taken as 19. The initial value has no impact whatsoever on the final results. Quarterly movements are in steps of 0.25.
(6) Calculation of V_G based on Equation 3.7a, i.e. **19 7765.7** = **10 329.3***100/**5.223**. Note that the GDP forecast at time t, $G_f(t)$, is taken as the value of G a year later, i.e. $G_f(t = 20) = G(t = 21)$.
(7) Calculation of V_E based on Equation 3.7b, i.e. **1162.8** = **60.731***100/**5.223**. Note that the earnings forecast at time t, $E_f(t)$, is taken as the I/B/E/S forecast at time $t = 2$.
(8) Transformed price, based on Equation 3.6. Computed as **6.1267** = ln(**1301.53**) − **5.223***__20__/100.
(9) Transformed GDP, based on Equation 3.8a. This is computed here as **5.7502** = ln(**197 765.7**) − **5.223***__20__/100 − 5.4. The factor of 5.4 has been subtracted at the end to adjust the level of $\Xi(b)$ to about $\Phi(b)$.
(10) Transformed I/B/E/S earnings forecast, based on Equation 3.8b. Computed as **6.0140** = ln(**1162.8**) − **5.223***__20__/100.

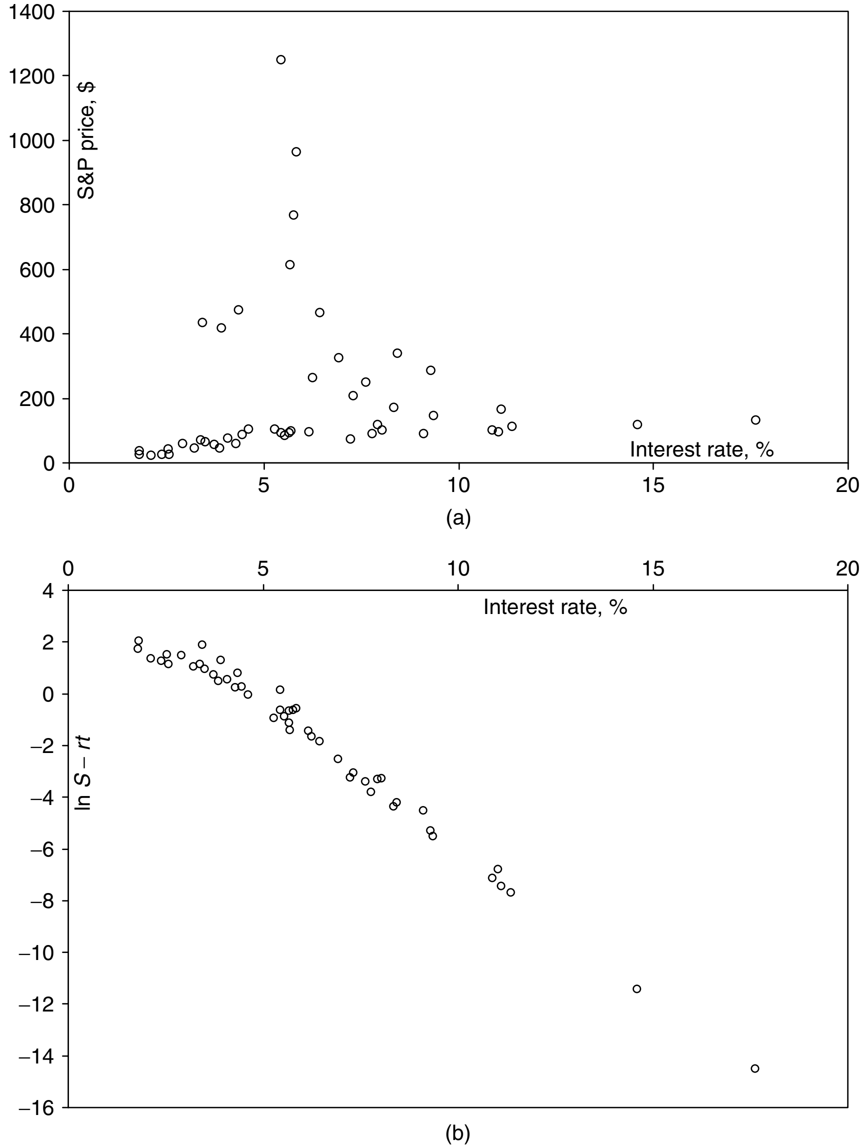


Figure 4: Raw and transformed data, respectively, of the S&P price versus the interest rate from 1950 to about 2000. Transformation of the price data is carried out according to Equation 3.6

Figure 6a, the FTSE100 price index, when mapped, depicts outliers that coincide exactly with time periods immediately prior to the August 1987 crash. In addition, evidence of structural breaks can also be observed in mapped plots of both earnings and GDP.

The JP data, which are included in Figures 8a–c, are substantially different. First, the impact of the transformation on the TOPIX price, as depicted in Figure 6a, is non-existent. Obviously,

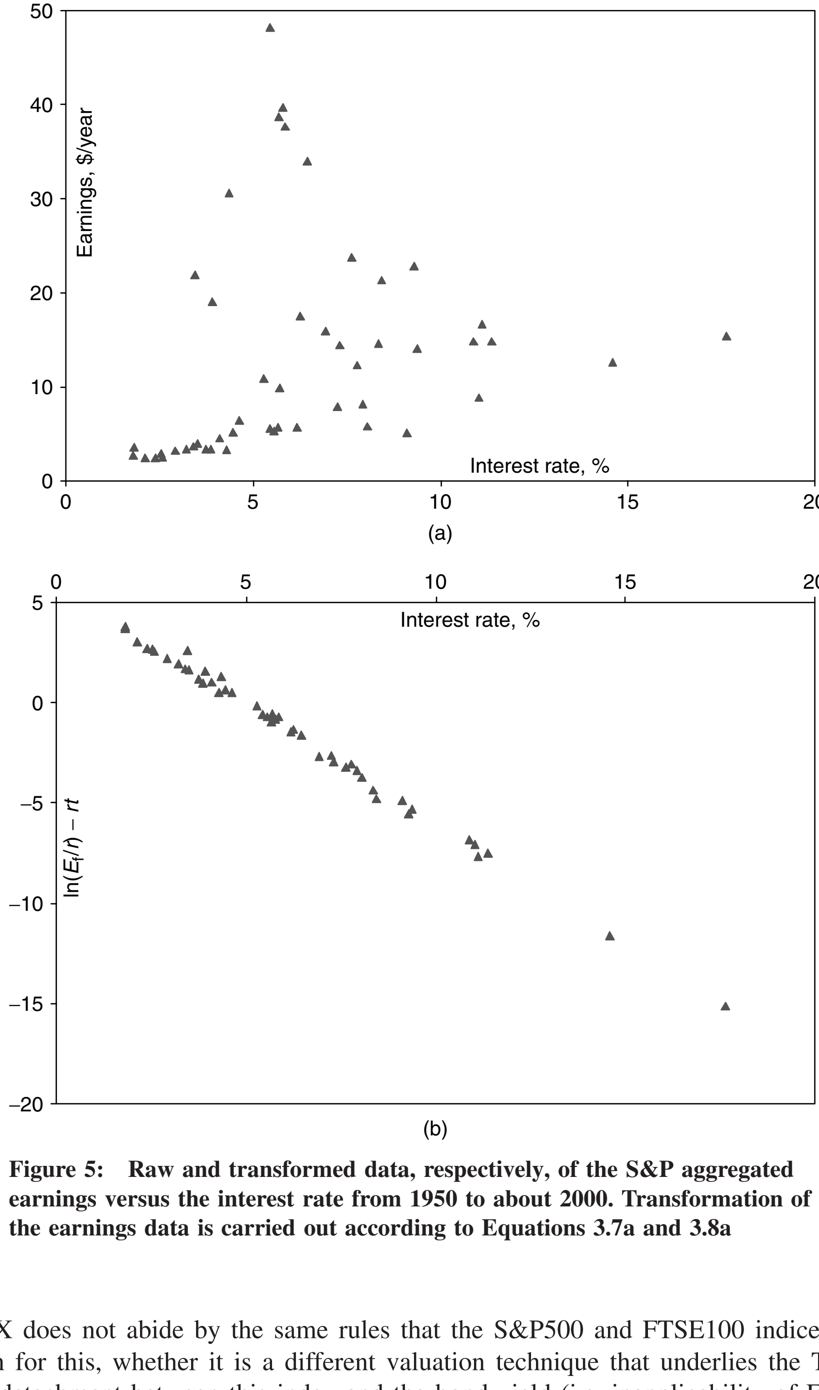


Figure 5: Raw and transformed data, respectively, of the S&P aggregated earnings versus the interest rate from 1950 to about 2000. Transformation of the earnings data is carried out according to Equations 3.7a and 3.8a

the TOPIX does not abide by the same rules that the S&P500 and FTSE100 indices do. As to the reason for this, whether it is a different valuation technique that underlies the TOPIX or a complete detachment between this index and the bond yield (i.e. inapplicability of Equation 3.2 to the TOPIX) is not up for speculation here. What is clear altogether is that this approach does not work for the TOPIX and, hence, cannot be used here.

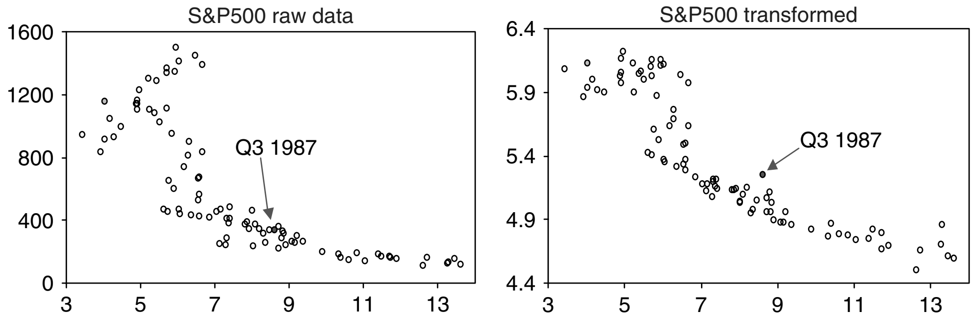


Figure 6a: The S&P500 price index raw data plotted against the 10-year US government bond yield (left) and its transformed counterpart (right). Note the highlighted point representing the quarter prior to the August 1987 crash, where the market was known to be overpriced. Time frame for the plot is Q1 1981 to Q1 2004. The darker point on the top left-hand side is the most current

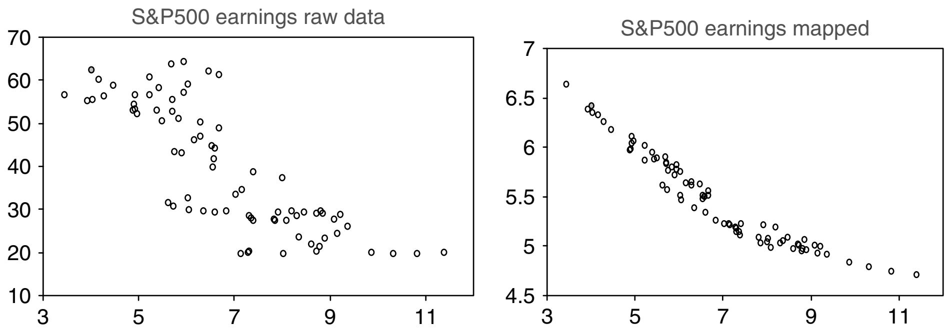


Figure 6b: The S&P500 I/B/E/S earnings forecast raw data plotted against the 10-year US government bond yield (left) and its transformed counterpart (right). Note the existence of a regime shift, similar to that portrayed in Figure 2. Time frame for the plot is Q1 1981 to Q1 2004. The darker point on the top left-hand side corresponds to the latest

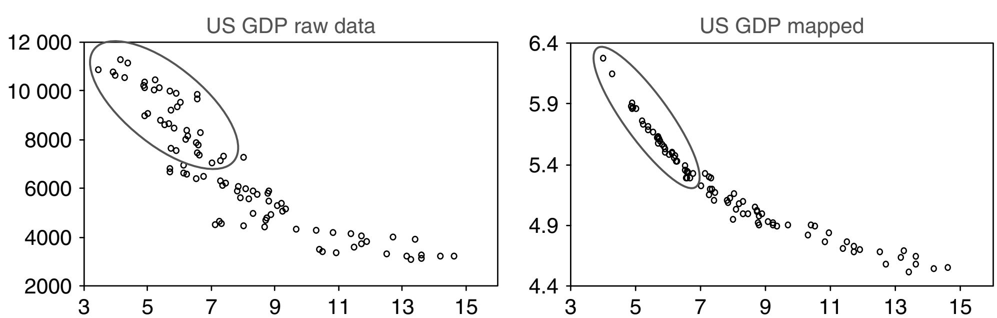


Figure 6c: The US GDP raw data plotted against the 10-year US government bond yield (left) and its transformed counterpart (right). Once again, note the existence of regime shifts. Time frame for the plot is Q1 1981 to Q1 2004. The encircled region covers Q1 2000 to Q4 2003, which, as it appears on the right-hand plot, belongs to a single structural regime and appears to contain no outliers

TABLE 2: DATASTREAM CODES FOR THE QUARTERLY DATA USED IN FIGURES 6 AND THEREAFTER

Country	Parameter	Datastream code
	S&P500	S&PCOMP
	I/B/E/S earnings forecast	@:USSP500(A12FE)
	US GDP	USGDP . . .B
US	30-year US gov. bond yld.	BMUS30Y(RY)
	10-year US gov. bond yld.	BMUS10Y(RY)
	7-year US gov. bond yld.	BMUS07Y(RY)
	5-year US gov. bond yld.	BMUS05Y(RY)
	2-year US gov. bond yld.	BMUS02Y(RY)
	FTSE100	FTSE100
	I/B/E/S earnings forecast	@:UKFT100(A12FE)
	UK GDP	UKGDP . . .B
UK	20-year UK gov. bond yld.	BMUK20Y(RY)
	10-year UK gov. bond yld.	BMUK10Y(RY)
	7-year UK gov. bond yld.	BMUK07Y(RY)
	5-year UK gov. bond yld.	BMUK05Y(RY)
	2-year UK gov. bond yld.	BMUK02Y(RY)
	TOPIX	TOKYOSE
	I/B/E/S earnings forecast	@:JPTOPIX(A12FE)
	JP GDP	JPGDP . . .B
JP	30-year JP gov. bond yld.	BMJP30Y(RY)
	10-year JP gov. bond yld.	BMJP10Y(RY)
	7-year JP gov. bond yld.	BMJP07Y(RY)
	5-year JP gov. bond yld.	BMJP05Y(RY)
	2-year JP gov. bond yld.	BMJP02Y(RY)

In contrast, however, a pattern does emerge when the I/B/E/S earnings forecasts are transformed, as shown in Figure 8b. Here, there is evidence of a structural shift in the earnings, coinciding to around the end of 1994 when the 10-year yield was approximately 4.5%. The JP GDP, on the other hand, which is illustrated in Figure 8c, displays a remarkably tight pattern, showing no signs of any structural change in the economy, at least from Q1 1984 to Q1 2004, the selected range of the data.

In the case of JP, therefore, one could conclude that bond yields (1) are completely detached from the TOPIX price, (2) have an influence on expected earnings and (3) are tightly coupled to the GDP. This, subsequently, could mean that in Japan, the GDP and TOPIX price are not connected to one another, so that any attempt to infer the direction of the TOPIX price, and possibly other Japanese equity indices, from expected movements in either the interest rates and/or the GDP is doomed to fail.

4.3 The impact of bond maturity

Having thus far concentrated only on the 10-year government bond yield, it is time now to question the applicability of the approach to other bond maturities. According to the governing

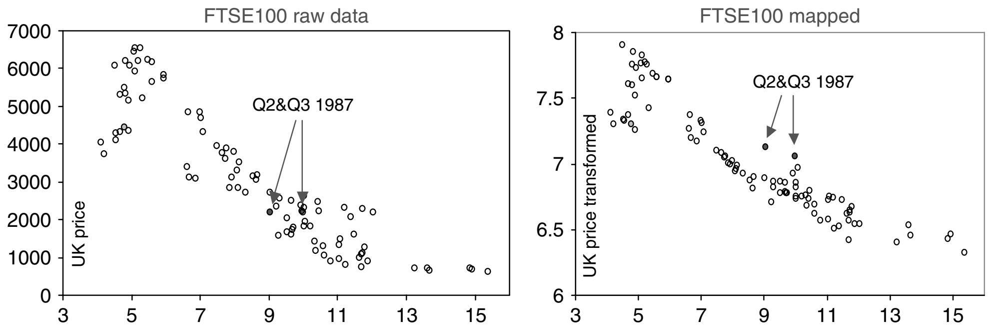


Figure 7a: The FTSE100 price index raw data plotted against the 10-year UK government bond yield (left) and its transformed counterpart (right). Note the highlighted points representing the two quarters prior to the August 1987 crash, where the market was known to be overpriced. Time frame for the plot is Q1 1981 to Q1 2004. The darker point on the top left-hand side corresponds to the latest

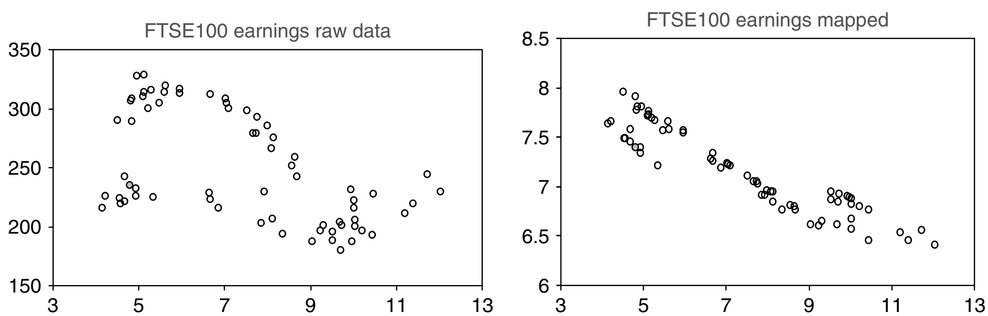


Figure 7b: The FTSE100 I/B/E/S earnings forecast raw data plotted against the 10-year UK government bond yield (left) and its transformed counterpart (right). Note the existence of regime shifts, similar to that portrayed in Figure 2. Time frame for the plot is Q1 1981 to Q1 2004. The darker point on the left-hand side corresponds to the latest

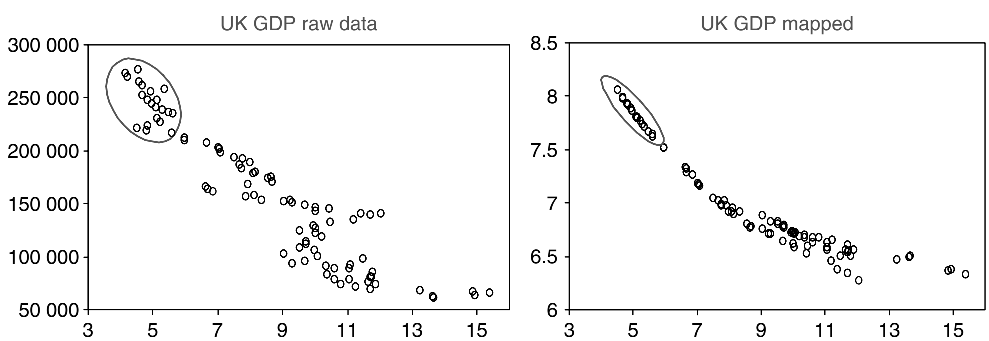


Figure 7c: The UK GDP raw data plotted against the 10-year UK government bond yield (left) and its transformed counterpart (right). Once again, note the existence of regime shifts. Time frame for the plot is Q1 1981 to Q1 2004. The encircled region covers Q1 2000 to Q4 2003, which, as it appears on the right-hand plot, belongs to a single structural regime and contains no outliers

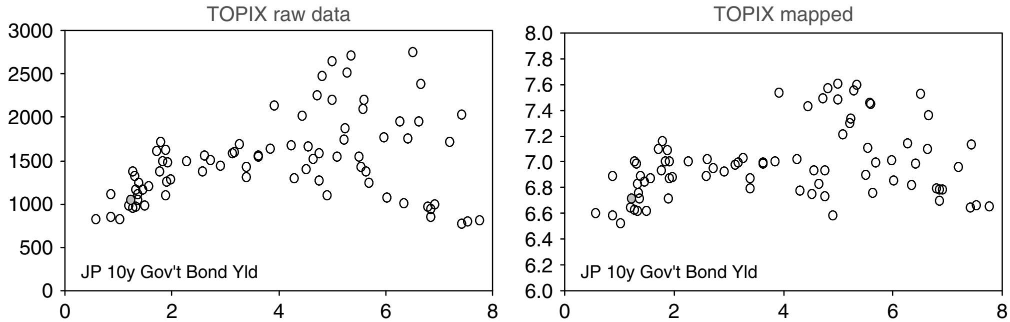


Figure 8a: The TOPIX price index raw data plotted against the 10-year JP government bond yield (left) and its transformed counterpart (right). Note the absence of any convergence in the mapped frame of reference. Time frame for the plot is Q1 1984 to Q1 2004. The darker point on the left-hand side corresponds to the latest

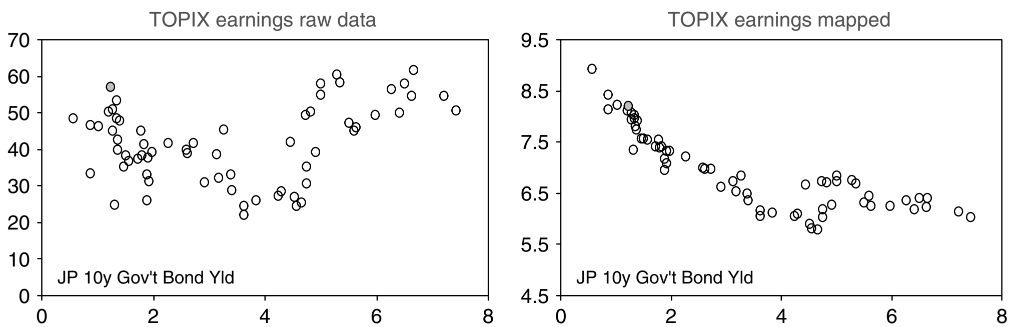


Figure 8b: The TOPIX I/B/E/S earnings forecast raw data plotted against the 10-year JP government bond yield (left) and its transformed counterpart (right). Note the existence of a regime shift, similar to that portrayed in Figure 2. Time frame for the plot is Q1 1984 to Q1 2004. The darker point on the top left-hand side corresponds to the latest

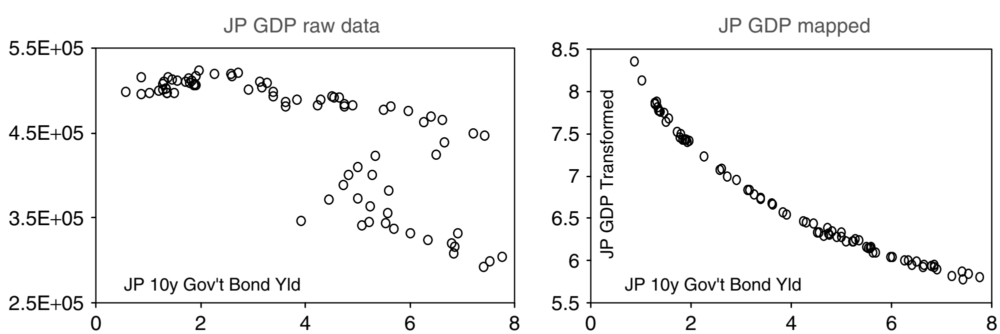


Figure 8c: The JP GDP raw data plotted against the 10-year JP government bond yield (left) and its transformed counterpart (right). Note the absence of regime shifts in the transformed plane, which covers the period Q1 1984 to Q4 2003

equations 3.6–3.8, bond maturity, T, plays *no* role in the model. Therefore, going back to Section 3.2.2, this means that, in the absence of outliers and structural shifts, the characteristic line of convergence in the mapped frame of reference should remain insensitive to the different maturities. More simply stated, all points that result from applying the coordinate transformation using yields from different bond maturities should, under the above conditions, fall exactly on the same line, regardless of maturity.

The validity of the above may now be examined, again visually, by producing plots similar to Figures 6–8. In doing so, care must be taken to select regions where structural shifts and outliers are absent, of which the area encircled in Figure 6c is one. This region contains the time frame Q1 2000 to Q4 2003 for the US GDP. Bearing in mind that the graph was constructed using the 10-year US government bond yield, we now ask what happens if different maturities were also included in the same plot.

The impact of bond maturity on, or rather the absence of its effect in, the present model is clearly demonstrated in Figures 9a–c, which enlarge the areas highlighted in Figures 6c, 7c and 8c, for the US, UK and JP,[16] respectively. In each of these figures, 9a–c, different government bond tenors—namely the 2, 5, 7, 10 and 30 years (20 instead of 30 years in the case of UK)—were plotted together, with the idea that any observable scatter could be attributed to the differences in maturities. Nevertheless, one obtains in all cases a remarkably tight fit, which provides further testimony to the earlier presumption (see Section 3.2.2) that the underlying curve is invariant to different maturities.

5 Potential applications

Prior to going forward with the development of the relative valuation model, two types of applications are brought to mind, both of which could have possible uses in the field of investment.

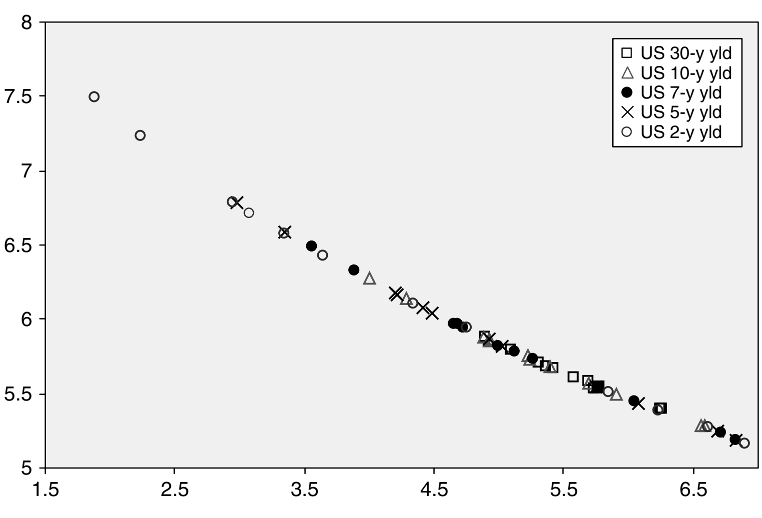


Figure 9a: The transformed US GDP for the area circled in Figure 6c, covering the time frame Q1 2000 to Q4 2003. The plot shows different maturities superimposed on each other. The horizontal and vertical coordinates represent b and $\ln V_G - bt$, respectively

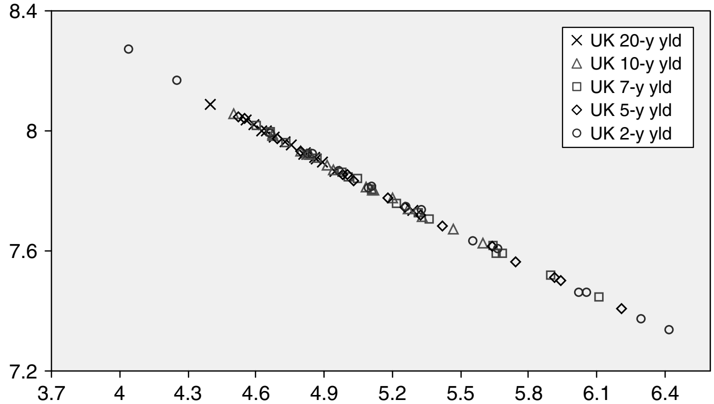


Figure 9b: The transformed UK GDP for the area circled in Figure 7c, covering the time frame Q1 1998 to Q4 2003. The plot shows different maturities superimposed on each other. The horizontal and vertical coordinates represent b and $\ln V_G - bt$, respectively

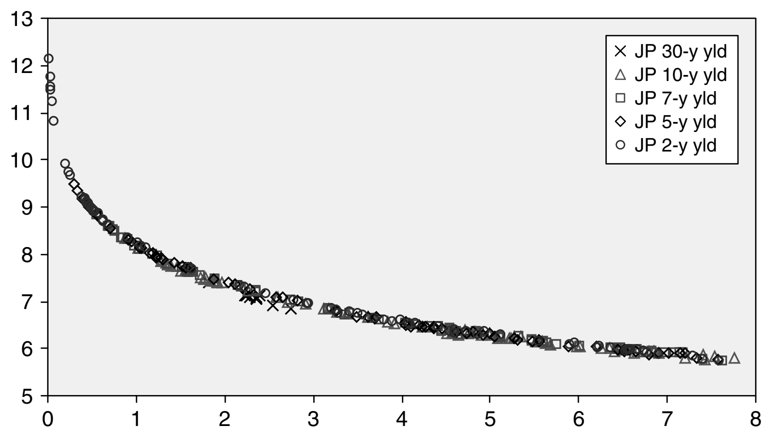


Figure 9c: The transformed JP GDP for all of the area in Figure 8c, covering the time frame Q1 1984 to Q4 2003. The plot shows different maturities superimposed on each other. The horizontal and vertical coordinates represent b and $\ln V_G - bt$, respectively

These applications, which are described next, result from the properties of the curves described in Section 4 and consist of forecasting the GDP and calculating the duration.

5.1 Forecasting the GDP

To illustrate the GDP forecasting capability of the model, one needs to combine Equations 3.7a and 3.8a, replace b^* by b and arrange the result as:

$$G_f(t) = \exp[\Phi(b) + bt + \ln b] \tag{5.1}$$

Recognising that $G_f(t)$ is the GDP expectation, Equation 5.1 then allows one to recover the expected GDP, one year from today, given today's yield, b, as well as the empirically determined function, $\Phi(b)$, which is extractable from plots similar to Figures 9a–c. The assumptions underlying this method are that (1) today's bond yields have the expected GDP priced in them and (2) between now and one year ahead from now, no structural shifts will occur, so that the function $\Phi(b)$ retains its shape over the time period between now and then.

Let us now apply Equation 5.1 to the three cases of interest here, namely the US, UK and JP. Focusing initially on the US, it is observed that a fourth-order polynomial curve runs satisfactorily through all the points in Figure 9a, comprising the yields associated with the different tenors. This curve, therefore, provides an empirical relation for $\Phi(b)$ with an R^2 of 0.99975. The tightness of the fit is noteworthy in Figure 10a, where the polynomial expression is also included.

What follows now is a step-by-step demonstration of how a forecast for the US GDP, let's say of Q1 2005,[17] could be obtained using Equation 5.1. (1) Compute from Table 1 the mapping of GDP to $\Phi(b)$. This, when plotted against b, leads to Figure 10a. (2) A curve fit, similar to the one in that figure, could then be obtained to represent the behaviour of $\Phi(b)$ with respect to b. In this case, a fourth-order polynomial was sufficient to achieve a very tight fit. (3) Return to Equation 5.1 and note that the expected GDP for Q1 2005—i.e. $G_f(t = \text{Q1 } 2004)$—may now be calculated by substituting the values of b, $\Phi(b)$ and the quantity bt, where, in correspondence to Q1 2004, b and t are 4.031% and 23, respectively (see Table 1).

Repeating the above procedure for the different bond maturities leads to Figure 10b, as well as Table 3, where different estimates of the Q1 2005 expected GDP have been obtained. These fall between 11 350 and 11 906 (in appropriate units), which correspond to the stretch of tenors between 2 and 30 years, respectively. A simple average finally provides an overall estimate of 11 619 for the Q1 2005 US GDP. Note that, since this value is based on yields that are market driven and which tend to vary rather gently on a day-to-day basis, the estimate for one-year-ahead GDP should also behave similarly.

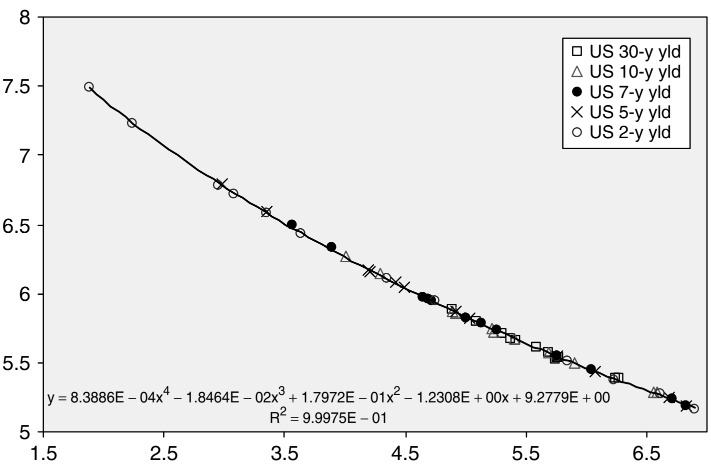


Figure 10a: Same as Figure 9a, but with a fourth-order polynomial curve fit passing through the yields belonging to the different maturities indicated in the legend. The extremely tight fit, as reflected by the high R^2, represents the function $\Phi(b)$

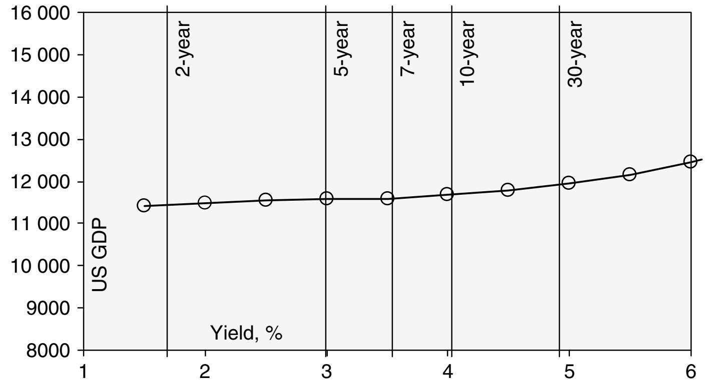


Figure 10b: The expected US GDP for Q1 2005 as a function of the interest rate, as derived from the methodology outlined in Section 5.1. The vertical lines correspond to the different maturity yields as of the time of data download—i.e. February 12, 2004, corresponding to Q1 2004

TABLE 3: EXAMPLE CALCULATION ILLUSTRATING THE GDP FORECASTING PROCEDURE. A VALUE OF 23, CONSISTENT WITH THAT IN TABLE 1, WAS USED HERE FOR t TO SIGNIFY Q1 2004. ALSO, $\Phi(b)$ WAS COMPUTED USING THE POLYNOMIAL FIT IN FIGURE 10a

Maturity, T	Bond yield, b, in %	$bt/100$	$\Phi(b)$	Expected GDP for Q1 05
2 years	1.687	0.38801	7.63116	11 350
5 years	2.998	0.68954	6.77352	11 566
7 years	3.544	0.81512	6.48367	11 601
10 years	4.031	0.92713	6.24891	11 671
30 years	4.911	1.12953	5.86892	11 906
	Average			11 619

To validate these estimates, a back test was performed following the same steps as above. Here, for instance, an estimate for the now historical Q1 2001 GDP level is obtainable from the yields of Q1 2000, as well as upon utilising the same expression for $\Phi(b)$. This back test provides Figures 11a–c, which pertain to the US, UK and JP, respectively. The basis of this is Table 4, which displays the fitted polynomials, as well as the time frames involved, for all three jurisdictions.

5.2 Calculating the duration

In the financial literature, the duration of any parameter, let's say X, is defined as its sensitivity to the interest rate, keeping all else constant. Thus, quantitatively, the duration of X, symbolised here by D_X, is represented by

$$D_X \equiv \left(\frac{\partial \ln X}{\partial b}\right)_t \tag{5.2}$$

Although simplistic in construct, problems abound when trying to calculate D_X in practice. First, since this application involves differentiation, then differentiating any volatile economic or market fundamental, such as the GDP, price, earnings, etc., will lead to even more volatile outcomes. Second, the above definition incorporates a *partial* differentiation with respect to b, which explicitly requires holding the time parameter, t, constant. This is an impossible feat to achieve in practice since expressing Equation 5.2 as the difference, let's say, in GDP level between Q1 2001 and Q1

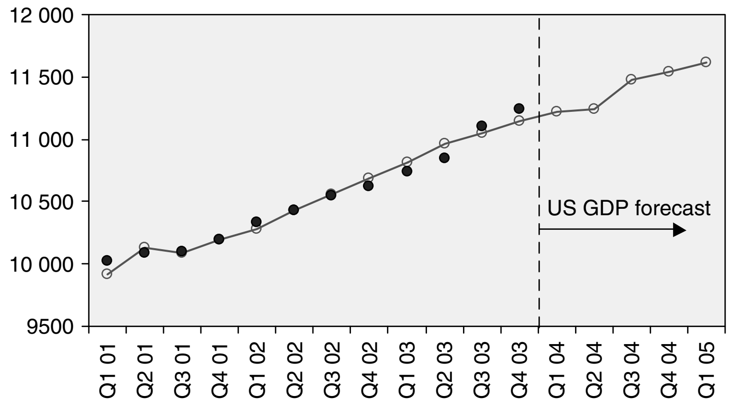


Figure 11a: The US GDP forecast post-Q3 2003 derived by the methodology outlined in Section 5.1. The historical data, which are the solid circles, are also included to demonstrate the close fit between model and data

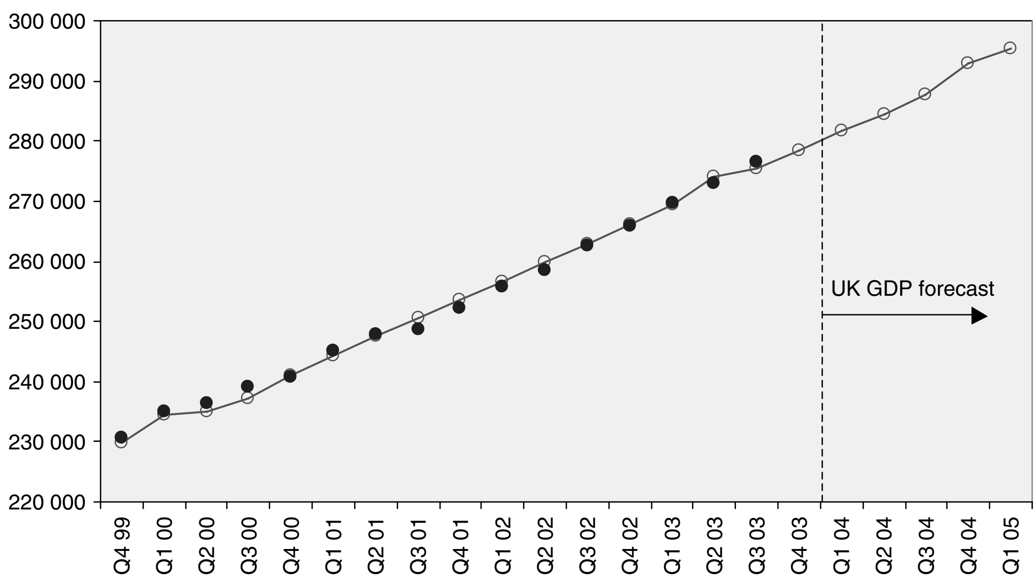


Figure 11b: The UK GDP forecast post-Q3 2003 derived by the methodology outlined in Section 5.1. The historical data, which are the solid circles, are also included to demonstrate the close fit between model and data

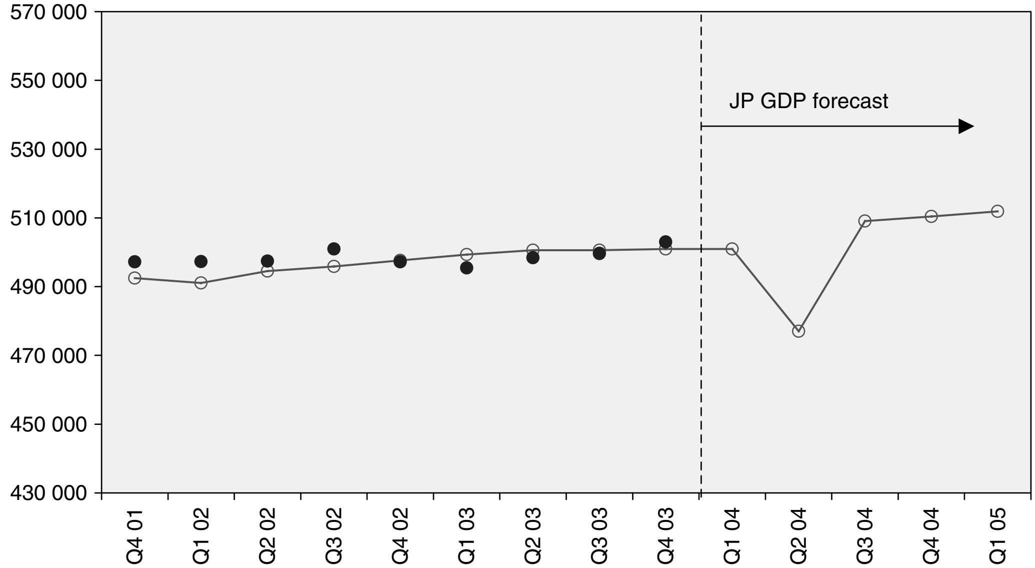


Figure 11c: The JP GDP forecast post-Q3 2003 derived by the methodology outlined in Section 5.1. The historical data, which are the solid circles, are also included to demonstrate the close fit between the model and data

TABLE 4: TIME FRAMES, MATURITIES, POLYNOMIAL FITS AND THE R^2 VALUES UNDERLYING THE CURVES IN FIGURES 9a–c

Market	Time frame of data	Maturities used	4th order polynomial curve fit	R squared
US	Q1 00 - Q1 04	2, 5, 7, 10, 30	$8.3886E - 04{*}b \wedge 4 - 1.8464E - 02{*}b \wedge 3 + 1.7972E - 01{*}b \wedge 2 - 1.2308{*}b + 9.2779$	99.975%
UK	Q4 98 - Q1 04	2, 5, 7, 10, 20	$-3.730118E - 03{*}b \wedge 3 + 8.758313E - 02{*}b \wedge 2 - 0.9.961051{*}b + 11.10780$	99.932%
JP	Q3 00 - Q1 04	5, 7, 10, 30	$0.1246{*}b \wedge 4 - 0.9396{*}b \wedge 3 + 2.7093{*}b \wedge 2 - 4.2651{*}b + 10.509$	99.960%

2000 divided by the yield b, will, implicitly, also involve a change in the time parameter. Thus, there is no way in practice that the above expression could be worked out.

Therefore, how could one get around this? Assuming for the time being that X is the GDP level, then, obviously, with $\Phi(b)$ being independent of time, the duration of the GDP—i.e. its sensitivity with respect to the interest rate while holding all else constant—could be computed by simply applying the partial differentiation to it. This yields the expression

$$D_{GDP} = \left(\frac{\partial \ln G_f(t)}{\partial b}\right)_t = \Phi'(b) + t + \frac{1}{b} \tag{5.3}$$

which greatly simplifies the calculation of the duration of GDP, among other data of interest. In practice, therefore, if one were to calculate the sensitivity of the GDP in Q1 2005 with respect to b, then it could be achieved from the above using $\Phi(b)$ in Table 4, along with the appropriate value for t, which, for example, is 23 for the US, in accordance to Table 1. This approach is mathematically more sound than the existing ones simply because the time parameter, t, is literally being held constant in the process of calculating duration.

6 The relative valuation of an equity price index

Thus far, the model has been developed and applied to forecasting the GDP and computing duration. What remains now is its implementation to relative valuation. This is simple as it only involves superimposing the three empirically determined functions, $\Psi(b)$, $\Phi(b)$ and $\Xi(b)$, directly on top of one another and looking for regions of deviation. It should be noted that this method incorporates no adjustable parameters, except for a basic and necessary one that is discussed in note 9 under Table 1.[18] To illustrate how the model works, we start with a preliminary description, along with a couple of historical examples, and then proceed with some detailed assessments.

6.1 A long-term historical example

For a preliminary demonstration, refer to Figures 4a and 4b, where, respectively, the historical price and earnings are mapped against the US government 10-year bond yield. A direct superposition of the two plots leads to Figure 12a, part of which has been magnified in Figure 12b.

Without dwelling too much on this, it is worth noting that the two data series, when mapped as $\Psi(b)$ and $\Xi(b)$ and superimposed, do fall on top of one another over most of the time covered, thus confirming that, with the exception of the period between 1950 and 1960, price and earnings are reasonably valued relative to each other. This chart, nevertheless, is based on annual data and, hence, does not capture the details that are to follow shortly.[19] Before going into these, however, it is worth alluding to an issue that comes up often in related literature—namely, Irving Fisher's assertion that the stock market was not overvalued just before its crash in 1929. An examination of this is carried out in the next section.

6.2 Irving Fisher and the 1929 stock market crash

Let us now apply the model to provide an answer to a long-debated issue, which is whether Fisher was right in his claim that the stock market was *not* overvalued before its dramatic crash in 1929, around the time when the great depression began. This issue seems to be a popular one, as countless papers have been written on it, each attempting to offer an explanation (see, for example, McGrattan and Prescott (2003) and references therein). We shall also try to provide an answer here, albeit strictly in the context of the present model.

Refer to Figure 13, which portrays a superposition of the three functions, $\Psi(b)$, $\Xi(b)$ and $\Phi(b)$, on each other over the time period 1928–1940. Any deviation observed in this mapped plane should, therefore, reflect the degree of relative valuation between the three fundamentals—being price, earnings and GDP.

First, note that from 1928 to 1931, all three fundamentals lie, more or less, near each other, signifying relative fair valuation. The significant deviation, which can be seen as a drop in the

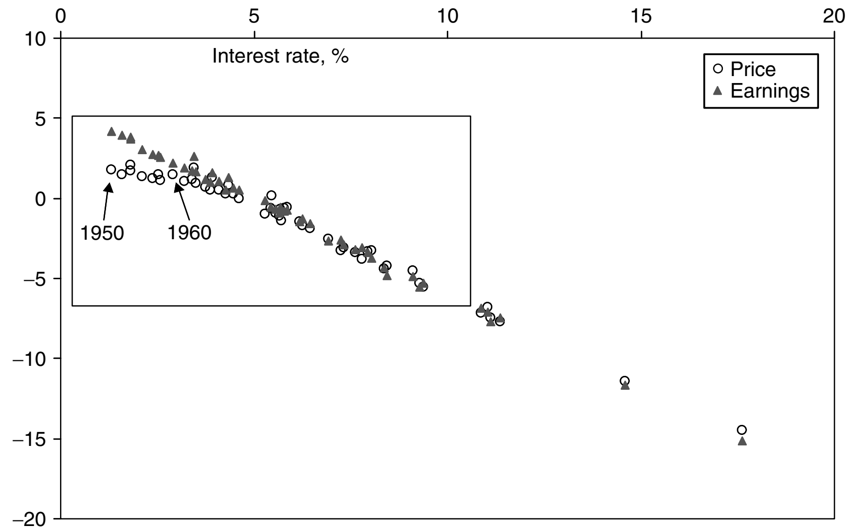


Figure 12a: Figures 4b and 5b superimposed, portraying the notion of relative valuation in the context of this work. Note that the points lie, more or less, on top of one another except for the time frame between 1950 and 1960, during which the convergence was in the process of happening

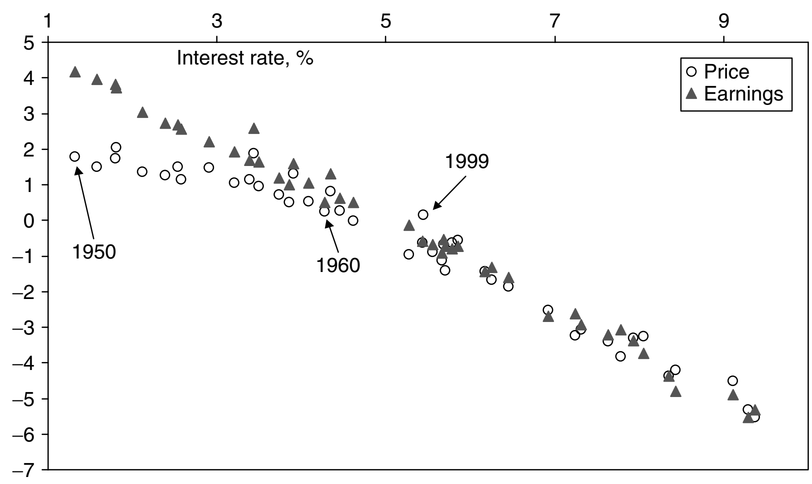


Figure 12b: Magnification of the boxed data in Figure 12a, illustrating the convergence of the two characteristic functions, i.e. $\Psi(b)$ and $\Xi(b)$, at around 1960

mapped price relative to the others, begins at around 1932 and becomes dramatic afterwards. Nevertheless, the mapped earnings and GDP remain reasonably close to one another throughout the whole time period. This, according to the model, means that, just before its plunge, the price was *not* overvalued[20] in relation to the earnings and GDP, but, nevertheless, it did become severely undervalued afterwards. Moreover, the observation that the earnings and GDP remained close to each other during the whole period simply implies that the former reflected the latter fairly well throughout the recession. With this in place, we can go now to the next section and discuss relative valuation in a more current time frame.

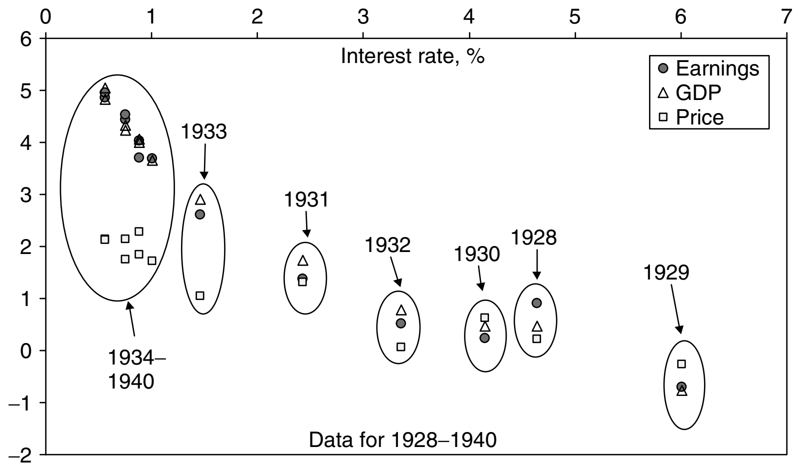


Figure 13a: Superposition of the three functions, $\Psi(b)$, $\Xi(b)$ and $\Phi(b)$, on each other using data covering the period 1928–1940. See Section 6.2 for explanation

6.3 Detailed examination

This section presents a closer look at the more recent time period, whereby the quarterly data displayed in Figures 6–8 are superimposed to exhibit signs of over- and/or undervaluation relative to each other. This is carried out thoroughly for the US and UK, but less so for JP since the TOPIX data, when mapped, lead to inconclusive results (see Figure 8a).

6.3.1 Relative valuation in the US data A relative valuation of the S&P500 price with respect to earnings is illustrated in Figure 14a, revealing the regimes of severe over-undervaluation relative to each other. In this figure, the outlier corresponding to the quarter before the Q3 1987 market crash is highlighted, as well as the time periods of the 1990s tech bubble, the Asian crisis and the post-2001 stock market decline.

Interesting, also, is the close-up view in Figure 14b, focusing on the time frame Q1 1999 to the present, being Q1 2004, and outlining the time-wise progression of the price and yield. This figure essentially displays the dynamics of the price movement, which started initially as overvalued relative to earnings, but eventually crossed the curve at around Q4 2001 to become undervalued, again relative to earnings. In the interest of space, no more will be said here, as the figure is self-explanatory.

Figure 14c displays a superposition of Figures 6a and 6c, relating the behaviours of the S&P500 price and the US GDP. Once again, the 1990s bubble period, as well as the post-2001 market crash, are clearly visible in the shape of deviations of the mapped price, $\Psi(b)$, from the mapped GDP, $\Phi(b)$. Finally, for the US, Figure 14d portrays the mapped S&P500 earnings, $\Xi(b)$, relative to the US GDP. Here, the period coinciding with the 1990s equity bubble is portrayed by a structural regime shift in the shape of a series of earnings data points that fall parallel to, but slightly above, the mapped GDP. Interestingly, however, the post-2001 decline in the market price, which is clearly apparent in Figure 14a, is not reflected at all by the earnings. This supports

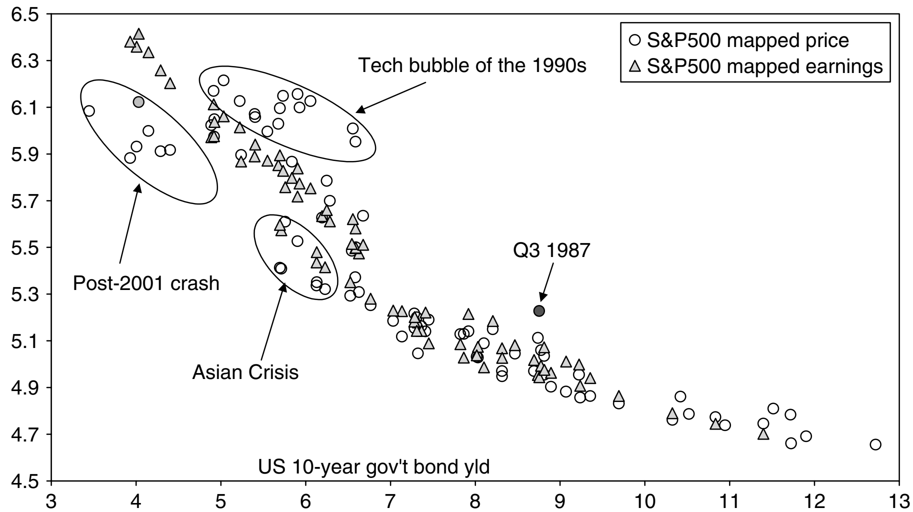
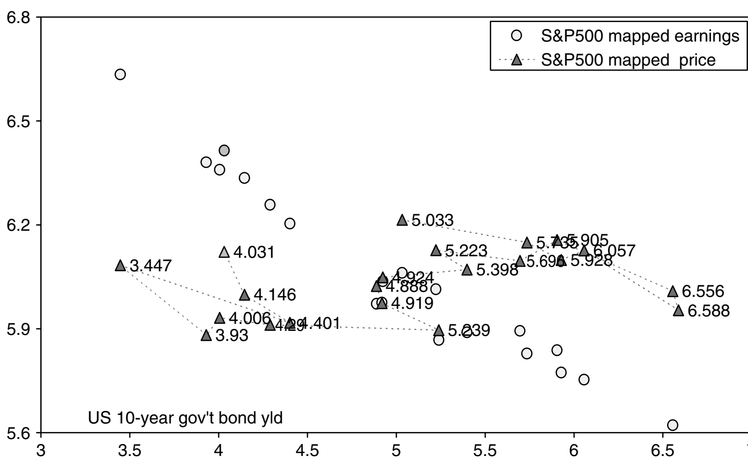


Figure 14a: Relative valuation of the S&P500 price and earnings via superposition of Figures 6a and 6b. Regions of gross deviation are circled. The 1990s tech bubble portrays overvaluation of the stocks relative to earnings and the post-2001 crash shows undervaluation of the former relative to the latter



Q1 99	1237.28	5.033
Q2 99	1333.32	5.735
Q3 99	1332.84	5.928
Q4 99	1424.94	6.057
Q1 00	1346.09	6.588
Q2 00	1437.2	6.556
Q3 00	1491.72	5.905
Q4 00	1367.72	5.696
Q1 01	1301.53	5.223
Q2 01	1291.96	5.398
Q3 01	1161.97	4.924
Q4 01	1138.65	4.888
Q1 02	1104.18	4.919
Q2 02	1106.59	5.239
Q3 02	928.77	4.29
Q4 02	900.36	4.006
Q1 03	851.17	3.93
Q2 03	944.3	3.447
Q3 03	999.74	4.401
Q4 03	1034.15	4.146
Q1 04	1151.82	4.031

Figure 14b: Close-up of Figure 14a, covering the period Q1 1999 to Q1 2004 and depicting the movement of the mapped price relative to mapped earnings. The table on the right-hand side lists the quarter, price and 10-year government bond yield in columns 1, 2 and 3, respectively

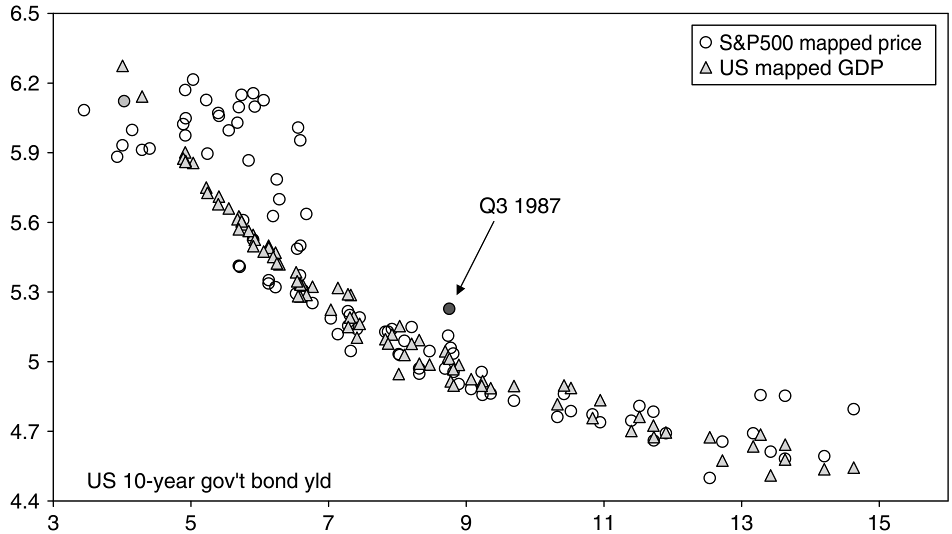


Figure 14c: Relative valuation of the S&P500 price and the US GDP via superposition of Figures 6a and 6c

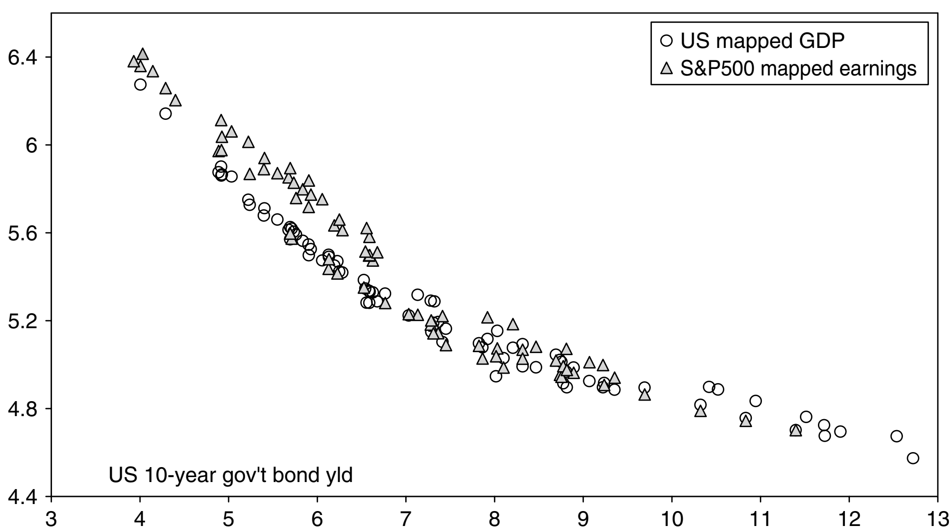


Figure 14d: Relative valuation of the S&P500 earnings and the US GDP via superposition of Figures 6b and 6c

the claim, albeit in retrospect, that the rise in the market's equity price during the 1990s was nothing but a bubble, which ultimately collapsed.

6.3.2 Relative valuation in the UK data Figure 15a, which is a superposition of Figures 7a and 7b, displays the relative behaviour of the FTSE100 price against earnings, both in transformed

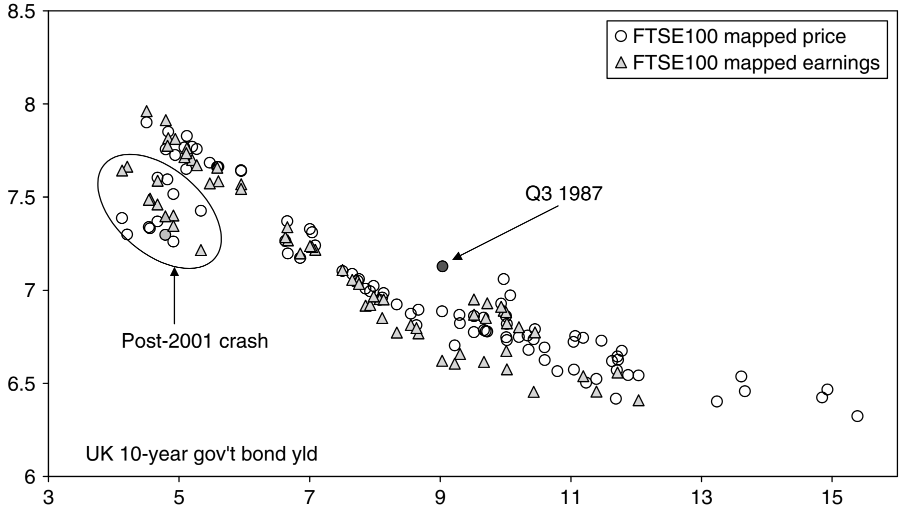
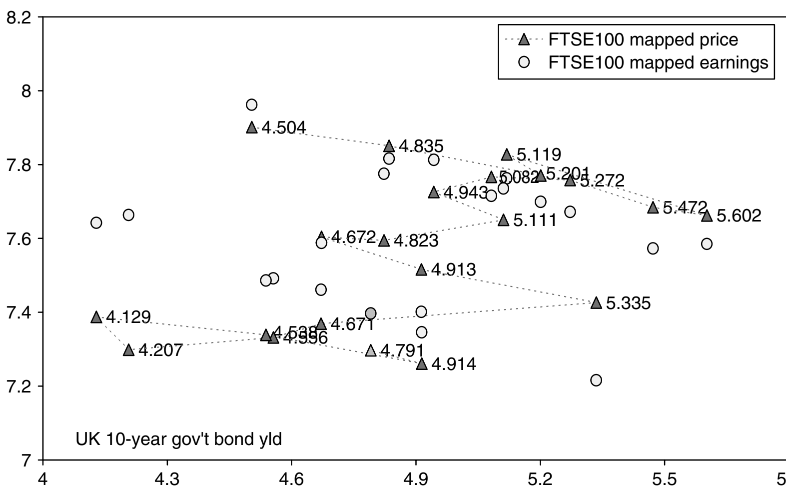


Figure 15a: Relative valuation of the FTSE100 price and earnings via superposition of Figures 7a and 7b. Regions of gross deviation are circled. In contrast to the S&P500 case in Figure 14a, the 1990s tech bubble is completely absent and a structural shift in both price and earnings appears to have occurred post-2001



Q1 99	6074.9	4.504
Q2 99	6206.43	4.835
Q3 99	6201.78	5.201
Q4 99	6550.75	5.119
Q1 00	6164.96	5.602
Q2 00	6232.95	5.472
Q3 00	6543.66	5.272
Q4 00	6440.1	5.082
Q1 01	6088.28	4.943
Q2 01	5914.98	5.111
Q3 01	5342.13	4.823
Q4 01	5290.98	4.672
Q1 02	5154.29	4.913
Q2 02	5217.98	5.335
Q3 02	4329.97	4.671
Q4 02	4115.99	4.556
Q1 03	3729.54	4.207
Q2 03	4049.01	4.129
Q3 03	4272.12	4.538
Q4 03	4354.7	4.914
Q1 04	4442.9	4.791

Figure 15b: Close-up of Figure 15a, covering the period Q1 1999 to Q1 2004 and depicting the movement of the mapped price relative to mapped earnings. The table on the right-hand side lists the quarter, price and 10-year government bond yield in columns 1, 2 and 3, respectively

planes, throughout roughly the last 20 years. The data point pertaining to the quarter prior to the Q3 1987 crash is, once again, highlighted. Here, however, in contrast to the S&P case discussed in Section 6.1.1 and illustrated in Figure 14a, there is no sign, whatsoever, of a price bubble.

In the 1990s, during the peak of the dotcom bubble in the US, the FTSE100 price is observed to follow the earnings consistently. In this case, however, what coincides with the collapse of the

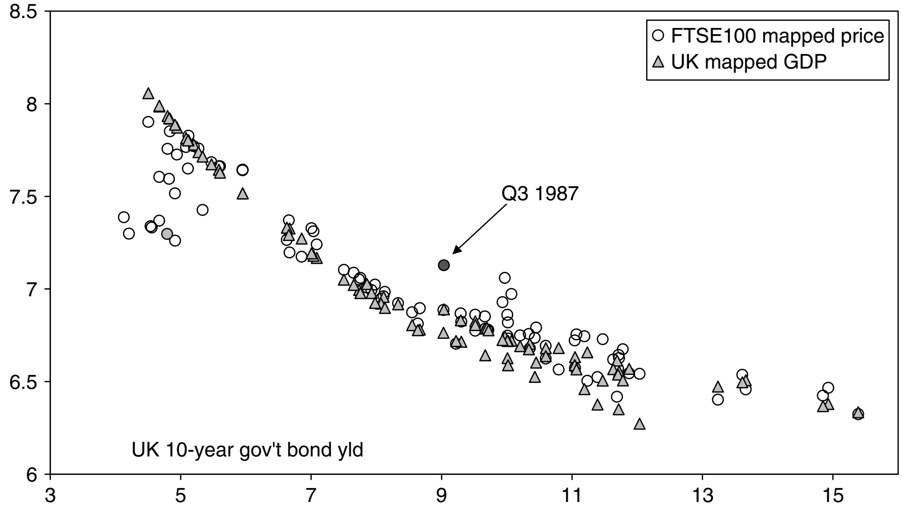


Figure 15c: Relative valuation of the FTSE100 price and the UK GDP via superposition of Figures 7a and 7c

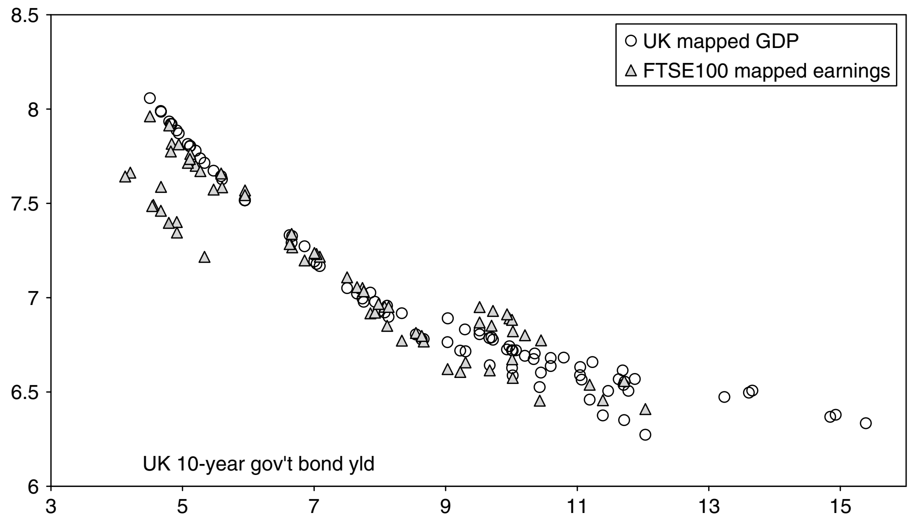


Figure 15d: Relative valuation of the FTSE100 earnings and the UK GDP via superposition of Figures 7b and 7c

price bubble in the S&P is a regime shift in the FTSE100 mapped earnings, which appears also to pull the FTSE100 price with it. This is further confirmed in Figure 15b, where the time-wise movements in earnings and price are depicted in close-up. Again, as in the above and in the interest of remaining objective, we shall not speculate here on the possible reasons for this regime shift (in the behaviour of the earnings and the subsequent fall in the FTSE100 price). Rather, an economist is perhaps better suited to provide an explanation for this.

The lack of a tech bubble, similar to that in the S&P data, in the FTSE price index is again verified in Figure 15c, where the mapped price in Figure 7a is superimposed on the mapped UK GDP in Figure 7b. Moreover, the existence of the regime shift in the FTSE100 earnings, as discussed in the previous paragraph, is found to be quite prominent in Figure 15d, which lays the mapped earnings in Figure 7b directly on top of the mapped UK GDP in Figure 7c.

Altogether, based on the above and without delving into detail, one could deduce that (1) the tech bubble that dominated the S&P500 during the 1990s did not exist in the FTSE100 market and (2) the decline in the FTSE100 price, which coincided with the S&P500 bubble collapse, was initiated by a regime shift in the FTSE earnings. Based on Figure 15d, this regime shift could be ‘corrected’ by either an increase in the interest rate (to shift the post-2001 earnings line in Figure 15d to the right to match the mapped UK GDP), an increase in earnings (to shift the same line in Figure 15d above to match the UK GDP), or a combination of both. Once the mappings coincide, fair valuation will presumably be achieved between earnings, GDP and price, that is if price will follow earnings.

6.3.3 Relative valuation in the JP data The superimposed JP data are displayed in Figures 16a–c. Figure 16a overlays Figures 8a and 8b, representing the mapped TOPIX price and earnings, respectively. Figure 16b, on the other hand, superimposes the mapped price on the mapped JP GDP in Figure 8c. From the perspective of relative valuation not much can be concluded, as there seems to be no pattern established in the mapped price.

Figure 16c lays the mapped I/B/E/S expected earnings of the TOPIX on top of the mapped JP GDP. There is similarity in the patterns here, although the earnings data converge less tightly and, as already discussed in Section 4.2, they do appear to exhibit some sign of a structural shift,

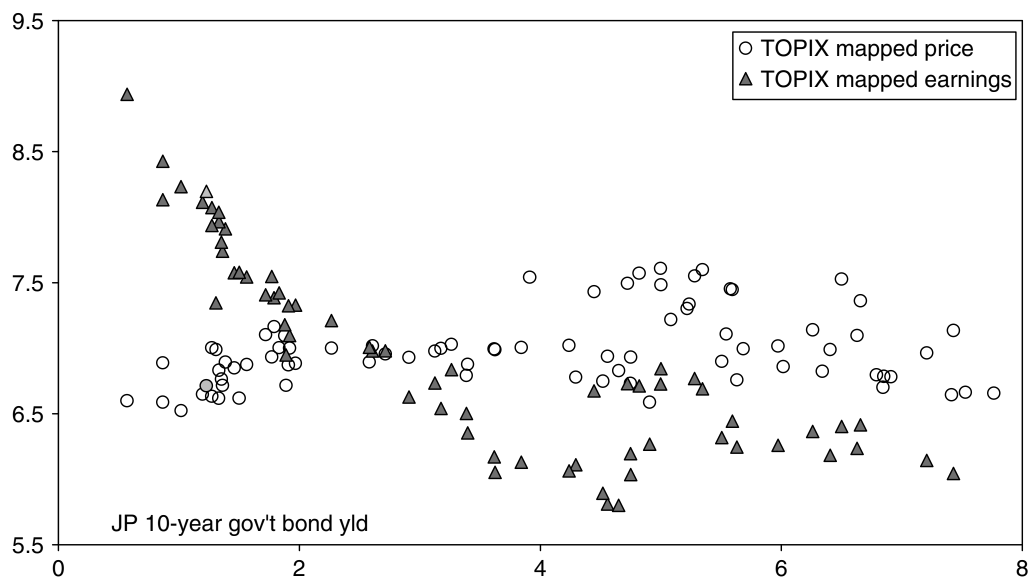


Figure 16a: Superposition of Figures 8a and 8b for the TOPIX mapped price and earnings. The nature of the price prevents any objective assessment of its relative valuation with respect to earnings

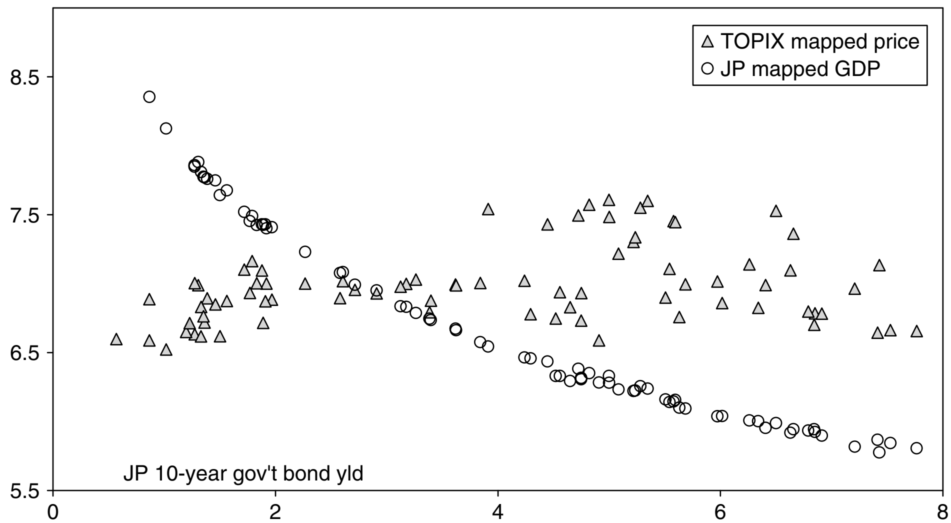


Figure 16b: Superposition of Figures 8a and 8c for the TOPIX mapped price and JP GDP. Once again, as in Figure 16a, the nature of the price prevents any objective assessment of its relative valuation with respect to the GDP

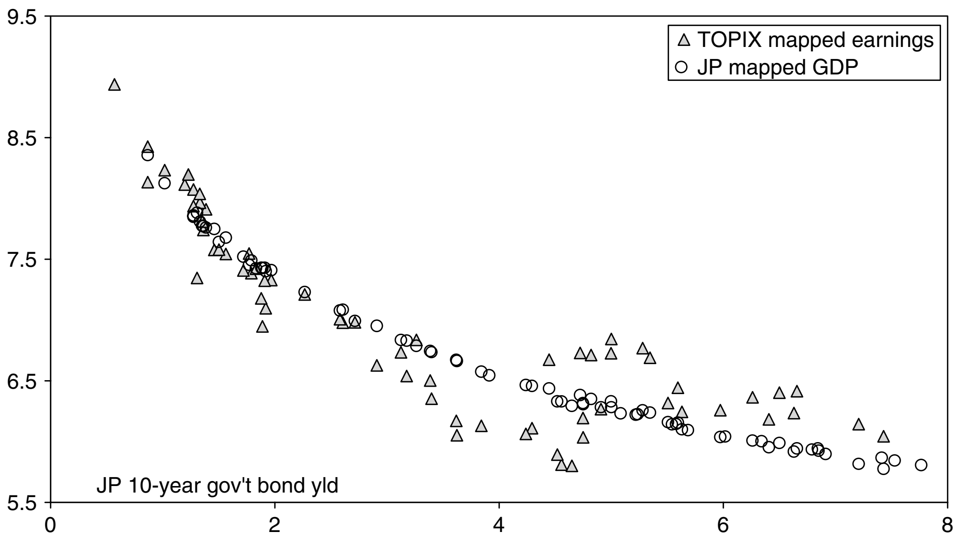


Figure 16c: Superposition of Figures 8b and 8c for the TOPIX mapped earnings and JP GDP

which is absent from the GDP. In terms of relative valuation between the TOPIX earnings and the JP GDP, however, it could be concurred that the two are currently, within the present regime of low interest rates, reasonably close to each other and, hence, the former can be considered to be a fair reflection of the latter.

7 Summary and conclusions

An objective and, hopefully, practical approach to relative valuation of an equity price index has been proposed. The method, which entails a simple mapping, enables one to (1) objectively compare the nominal GDP, corporate earnings and equity index against one another, (2) pinpoint outliers and structural shifts in the data and distinguish between the different regimes, (3) extract an estimate of the GDP forecast for next year, given today's interest rates and (4) obtain a mathematically sound expression for calculating duration. Application of the new method to the US, UK and JP markets and economies led to certain conclusions, some of which are listed below.

1. Fisher's claim that the stock market, just before its dramatic crash in 1929, was not overvalued is supported.
2. A historical, but detailed, assessment of US data, involving the S&P500 price and I/B/E/S earnings forecast, as well as the US GDP, over the last 20 years clearly confirms the existence of the 1990s price bubble in comparison to the earnings and the GDP, and its subsequent collapse in 2001. The collapse brought down the price to fair value relative to both earnings and GDP.
3. An assessment of the UK data, similar to the above, was also undertaken. Here, in contrast to the S&P price data, the results point to the absence of any price bubble in the FTSE100. The subsequent fall in the price, which nonetheless coincided with the collapse of the S&P bubble, occurred as the FTSE100 aggregated earnings underwent a structural shift. A disparate line in Figures 7b, 15a and 15b clearly marks this shift.
4. The situation in JP is markedly different. As depicted in Figures 8a and 16a, b, the mapping transformation has no impact whatsoever on the TOPIX price. In this case, the unmapped price undergoes no change in pattern when subjected to the transformation defined in Equation 3.6. This potentially means that the effect that interest rates or bond yields have on the S&P500 and FTSE100 price indices are totally absent here. As a result, the policy of varying interest rates to manipulate the equity price index does not work in JP under the present circumstances.
   In contrast, the TOPIX earnings and the JP GDP acquire well-defined patterns under the proposed coordinate transformation. A superposition of the two indicates that currently they are both fairly valued relative to each other.

All said, the new model does appear to have some potential as a relative valuation tool and, thereby, might be worth developing further. This could well involve (1) applications to other major equity indices that lie within the same jurisdictions covered here, (2) applications to other jurisdictions and, finally, (3) delving deeper into the other possible uses that were briefly mentioned here—namely, extracting the expected GDP and calculating the duration.

FOOTNOTES & REFERENCES

1. I express these views as an individual, not as representative of companies with which I am connected. E-mail: ruben.cohen@citigroup.com Phone: +44(0)207 986 4645 **Contact address**: Citigroup, London E14 5LB, UK
2. This is also known as the dividend discount model.

3. Note that this is also the return on equity (ROE), which is more an identity rather than a valuation tool.
4. Some might debate here that the DCF or ROE relationship in Equation 2.3 must contain a growth term for the earnings, analogous to the dividend-growth term in Gordon's Growth Model. The argument against including such a term, however, relies on the classical relationship between the plowback ratio and equity growth. The relation, according to literature (see, e.g., Brealey and Myers 1996), as well as intuition, implies that $E_f - \delta_f = \Delta S$, where ΔS is the growth in equity. Dividing both sides of this by the equity, S, leads to an equality between Equations 2.1 and 2.3. This equality first suggests that the total rate of return is the same as the ROE and, second, it reconciles the income statement with the balance sheet. Inclusion of any growth term in Equation 2.3 would, otherwise, produce something inconsistent with the plowback relation provided above.
5. The notion of the risk-free rate is also surrounded by controversy, especially in the empirical literature. Although there is little argument that this number should be based on a government-issued security, questions abound as to what maturity it should take. Another problem, which is more fundamental in nature, addresses the 'riskiness' of the risk-free rate — that is how could government securities be considered risk free when they are, as with any other type of security, volatile and impossible to predict.
6. This, obviously, presents an idealised scenario, but it will be relaxed later as the relative valuation model is developed.
7. Which is especially valid in the absence of volatility.
8. Based on this, therefore, firms pay and/or investors demand dividends because of the uncertainties inherent in the market. Take away these uncertainties — i.e. as per Propositions 1 and 2 — and the dividend yield will disappear altogether from the fundamental relationships, Equations 2.1 and 2.2.
9. Mappings and/or coordinate transformations, whose principal objective is to condense theoretical and empirical data into more manageable formats, have, for nearly a century, played a central role in the field of fluid mechanics. Although a few successful attempts have been made so far to apply this technique to economics (see, for instance, de Jong 1967 and Cohen 1998), as of yet, and as far as we are aware, very few endeavours, if any, have been made to incorporate it into finance.
10. Although materially different in approach from the classical 'dimensional analysis' described in de Jong (1967) and Cohen (1998), among others, the fundamental purpose of the coordinate transformation introduced here remains essentially the same.
11. The DCF model converges with the dividend discount model after 1950 (refer to Cohen 2002).
12. Note the similarity between Equations 3.7b and 2.3, as they are both based on the DCF model.
13. Price and earnings data from Shiller (`http://www.econ.yale.edu/shiller/data/ie_data.htm`). Interest rate data from the Fed website (`http://www.federalreserve.gov/releases/h15/data/m/tcm10y.txt`).
14. As discussed in Section 3.2.2, and as it will also be shown in a later section, the choice of bond maturity does not matter.
15. The earnings data used here are actual, rather than the I/B/E/S forecasts. Therefore, V_E was in this case computed the same way as V_G, where the one-year forward is substituted for

today's forecast of one-year ahead—i.e. $E(t+1)$ used for $E_f(t)$ (refer, for instance, to Table 1 for the method of calculation of $V_G(t)$).

16. Since the JP GDP is all one regime, then Figure 9c contains all the time frame included in Figure 8c.

17. Given that today is Q1 2004.

18. The need for this arises from the scale differential between the GDP and the aggregated index earnings.

19. More detailed, quarterly data will be shown later to clearly capture the 1990s bubble and its collapse.

20. This, therefore, is consistent with Fisher's claim and all the subsequent works that support it.

■ Barth, M. E., Beaver, W. H. and Landsman, W. R. (1998) Relative valuation roles of equity book value and net income as a function of financial health. *Journal of Accounting and Economics*, 25, 1–34.

■ Brealey, R. A. and Myers, S. C. (1996) *Principles of Corporate Finance*, McGraw-Hill, NY.

■ Cohen, R. D. (1998) An analysis of the dynamic behaviour of earnings distributions. *Applied Economics*, 30, 1–17.

■ Cohen, R. D. (2002) The relationship between the equity risk premium, duration and dividend yield. *Wilmott Magazine*, November issue, 84–97. Paper also available in `http://rdcohen.50megs.com/ERPabstract.html`

■ de Jong, F. J. (1967) *Dimensional Analysis for Economists*. North Holland, Amsterdam.

■ D'Mello, J. P., Lahey, K. E. and Mangla, I. U. (1991) An empirical test of the relative valuation of portfolio selection. *Financial Analysts Journal*, 47, 82–86.

■ McGrattan, E. R. and Prescott, E. C. (2003) The 1929 stock market: Irving Fisher was right. Federal Reserve Bank of Minneapolis Research Department Sta. Report 294.

■ Peters, D. J. (1991) Valuing a growth stock. *Journal of Portfolio Management*, 17, 49–51.

10
What the Spreadsheet Said to the Database, Just Before the Regulator Shut Down the Trading Floor...

Brian Sentance

What is the world's most popular trading system? Infinity? No. Front? No. Imagine? Think again. Sophis? Non. Open Bloomberg then? Not even warm!

The most popular trading system in the world, unchallenged in its dominance, is Microsoft Excel.

Billions of dollars are traded and hedged worldwide using the humble spreadsheet. It is a scary thought that the stability of the derivatives markets, not to mention the security of my retirement and yours, might depend upon simple spreadsheet formulas. '$A1 = B1 * C1$' and its variants probably deserve a lot more attention than we currently afford them.

Why is the spreadsheet so popular in derivatives markets? The fundamental answer is that market conditions and trading ideas change on a second-by-second basis. No software system, with the possible exception of the spreadsheet, has yet been designed that can deal with such rapid change in requirements. Years can be spent (and indeed are being spent) designing the perfect trading system, a perfect trading system that will be obsolete from the moment its design, let alone its delivery, is complete.

In a perfect world, new derivative products would be designed, tested and brought to market in hours or minutes, maximising profit margins and market share for the institution that gets them out first. A new product would have its risks understood and be fully integrated with all core risk management systems and processes.

The reality is different. New derivative products are, by their very nature, innovative and as a result less than well understood. A new product may be defined by complex behaviours and require complex data structures to support its pricing and risk management. This complexity often makes new product types difficult to integrate into core trading and risk management systems from both a technical and business process perspective.

As a result, it is relatively common for trading desks to bring a new derivatives product to market using only spreadsheets to price, hedge and manage it. Business users, without the need for extensive systems knowledge, can easily and quickly pull together in a spreadsheet complex instrument data, real-time and historic prices and positional data. And when they change their minds, they can change their spreadsheets to suit. Profit margins, market share and bonuses are up and everyone is happy. Well, not quite everyone.

Risk managers, product control, compliance and IT staff are usually not happy with the extensive use of trader spreadsheets (some traders would say that they are never happy, but that is a separate debate for another day!). Risk managers are understandably concerned that it is the traders themselves who mark the fair market value of the very trades on which their bonuses depend.

Additionally, it may be that only one or two key individuals have been involved in the design and testing of the pricing model for the product. If so, how can the institution really be sure the product has been thoroughly stress-tested? Besides, risk cannot be accurately assessed at an enterprise or even portfolio level since these new instrument types cannot be integrated into core risk management systems quickly enough.

IT staff are concerned that spreadsheets are difficult to support, with undocumented logic and multiple copies of the same spreadsheet to track down. They may be concerned that the trading desk will blame them if the risk numbers cannot be produced due to a (technically) corrupt spreadsheet. IT would also like the traders to use fewer spreadsheets and more of the systems they have spent many man-months developing, often systems that traders themselves requested in the first place.

So, in summary, we are faced with an unhealthy triangle of frustration at some financial institutions. Risk managers are frustrated with 'out-of-system' instrument trades, traders are frustrated with long system integration times for new instruments and IT are trying to make sense of all of this against a background of ever-changing business requirements.

The problems outlined above would be of passing importance for some financial institutions if it wasn't for the part of the regulator. Regulation can translate the above frustrations into hard cash, at which point everyone sits up and takes notice.

Without consistency of data, model risk, control over marks submitted and incomplete risk reports, regulators are not happy. At best, they may slow down CAD approval to address these issues. They may increase regulatory capital as a result of poor derivatives management, increasing both direct interest costs on capital and indirect business opportunity costs. At worst, regulators can raise regulatory capital requirements so high as to undermine the whole viability of the business.

So how do we solve the problem? For the moment I would suggest that what is needed is to maintain all the flexibility of the spreadsheet without incurring the costs and threats for risk management and IT. Traders could continue to implement complex calculation and data logic but with transparency and simple availability for all interested parties.

A number of companies, my own included, are making progress in this direction. Many institutions already make use of Microsoft Excel as a calculation server for complex products. However, this has its limitations, since Excel is not currently designed for scalable and reliable server deployment. Due to client interest, I understand that Microsoft itself is considering the development of a server deployable version of Excel.

Given some of the limitations of the relational data model in supporting array and time-aware data, the introduction of a native XML data type within certain RDBMS will ease the problems associated with ever-changing data requirements for new derivative products.

There are also products aimed specifically at translating existing Excel business logic from spreadsheets into another, more robust form. Of recent note is the approach taken by Savvysoft and its resulting trademark dispute with Microsoft over Savvysoft's TurboExcel product.

At my own company, Xenomorph, we have recently introduced a spreadsheet object that is a native data type within our TimeScape data management software. This spreadsheet-meets-database approach enables traders to carry on designing spreadsheet logic that can be centrally stored, administered and accessed by anyone who is permissioned to do so.

In summary, I believe that spreadsheet data management within the derivatives and wider financial markets is a major issue, and one that deserves everyone's attention due to the opportunities, costs and risks involved.

So what did the spreadsheet say to the database, just before the regulator shut down the trading floor? At such a late stage in the process, my guess would be a dialog box saying 'Microsoft Excel is currently unable to respond'...

11 Emotionomics: Ask Marilyn and Win a Car

Henriette Prast

Every now and then, the Monty Hall problem pops up. Whether in fiction, in job interviews or in academic publications, the problem continues to fascinate those who stumble upon it.

In the 1960s, Monty Hall was a host in the American game show ‘Let’s Make A Deal’. The final contestant in the show could win a car as a prize. The host showed the candidate three doors. Behind one of the doors was a car, behind the other two a goat. The candidate had to pick a door. After the candidate had chosen a door, but before it was opened, Monty Hall used to open one of the other doors, with a goat behind it. He then offered the candidate the opportunity to change his mind and switch doors. Most candidates did not switch.

In 1990, a reader of *Parade* magazine wrote a letter to Marilyn vos Savant, ‘the person with the highest IQ in the world’, who has a column called Ask Marilyn in which she answers mathematics questions sent in by readers. The reader asked Marilyn whether candidates of the Monty Hall show should switch. Marilyn said they should, because that would increase their chances of winning a car, from one-third to two-thirds.

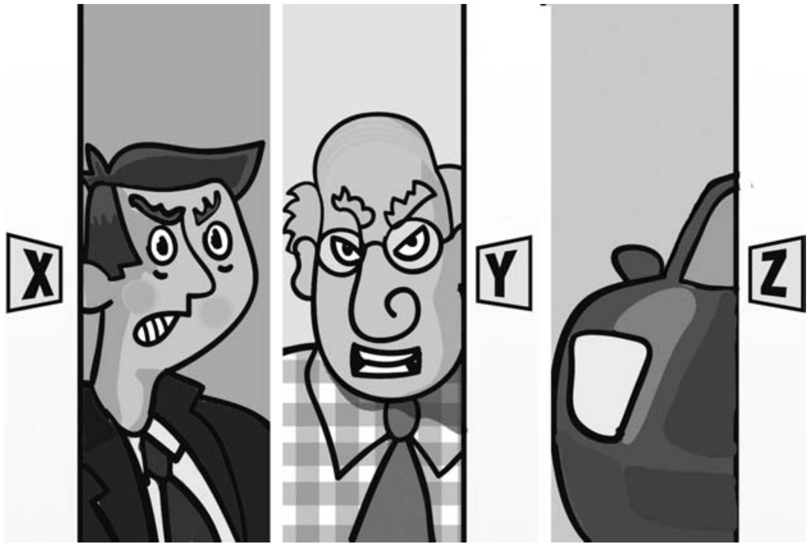
But many readers of *Parade* disagreed with Marilyn, and were of the opinion that it did not make a difference whether the candidate changed doors, as they believed the chances of winning a car were fifty-fifty between switching and not switching. Among those readers were mathematicians with a PhD such as Robert Sachs, PhD from George Mason University, who wrote to Marilyn: ‘I am very concerned with the general public’s lack of mathematical skills. Please help by confessing your error’, and E. Ray Bobo, PhD from Georgetown University, who wrote: ‘How many irate mathematicians are needed to get you to change your mind?’

In 2003, 15-year-old Christopher Boone, the main character of the moving and highly recommended novel *The Curious Incident of the Dog in the Nightime*,[1] explains to the readers of his diary why Marilyn vos Savant is right and all those PhDs were wrong. He does so with the help of maths, and with the following scheme:

Door X = goat

Door Y = goat

Door Z = car



Suppose you choose door X. Monty opens door Y, and if you switch you win a car.
Suppose you choose door Y. Monty opens door X, and if you switch you win a car.
Suppose you choose door Z. Monty opens door X or Y, and if you switch you win a goat.
Hence the chances of winning a car are one-third if you stick to your first choice and two-thirds if you switch.

As Timothy Crack reports in *Heard on the Street: Quantitative Questions for Wall Street Traders*, Wall Street firms often use the Monty Hall problem to assess job candidates.[2] In academic research, the problem is used in experiments to study individual decision making under risk. Some might argue that when in a television show people have only little time to decide and are most likely very nervous. However, from the experimental research it turns out that even when in a tranquil setting, and where there is less at stake than a car, people turn out to behave just like the Monty Hall contestants. For example, Benjamin Friedman (1998), in a series of experiments with individual decision making under risk along the lines of the Monty Hall problem, finds an average switching rate of a poor 30%, implying that a majority of the subjects in his experiment, just like the contestants in 'Let's Make A Deal', did forgo the possibility to double their expected return.[3]

In a recent paper, Brian Kluger and Steve Wyatt use the Monty Hall problem to detect probability judgment errors in an experimental investment context.[4] The main purpose of their study is to see whether the Monty Hall type individual cognitive errors do or do not translate into price and allocation inefficiencies at the aggregate level.

The test was a market experiment, with participants being endowed with three certificates and being able to trade by taking part in a second-price sealed bid auction, followed by an oral double auction. The market experiment consisted of nine steps in each session, and each session consists of 12 trials. It was conducted by 12 different cohorts of six participants, and was structured as follows (see Box 1).

In the market experiment, the expected return of the convertible asset is 67 USE cents and that of the non-convertible asset is 33 USE cents. Hence the efficient price ratio should equal 2. However, this is not what Kluger and Wyatt find.

Box 1: Sequence of events for a trial of the market.

1. Three asset certificates are issued to each participant. The type of asset issued (heart, diamond, or club) will be the same for all participants in the trial but will vary across trials. The assets will be either non-convertible, all-or-none convertible, or partial-convertible.
2. Three cards are randomly placed on a mat, face down. Participants are shown that one card is a heart, one is a diamond, and one is a club. The cards are shuffled and placed on a mat in positions labelled one, two, and three.
3. A ball is chosen randomly from a bingo cage and the corresponding card is set aside. The bingo cage contains three balls, labelled one to three. A ball is chosen and participants are shown the ball. The card corresponding to the ball is set aside.
4. An additional asset certificate is sold in a second-price sealed-bid auction. Participants submit bids to purchase an additional asset certificate from the 'house'. The asset certificate is identical to assets issued at the start of the trial. If the treatment is both-arbitrage, then two auctions are simultaneously conducted, one for non-convertible shares and one for convertibles.
5. Participants can trade shares in an oral double auction. Participants can buy shares from or sell shares to each other via an oral double auction. The auction lasts for two minutes.
6. Participants can trade shares in another oral double auction. Once more, participants can buy shares from or sell shares to each other via an oral double auction. The auction lasts for two minutes.
7. An additional asset certificate is sold in a second-price sealed-bid auction. Participants submit bids to purchase an additional asset certificate from the 'house'. The asset certificate is identical to assets issued at the start of the trial.

8a. Non-convertible shares: no action possible.

8b. All-or-none convertible shares: each participant secretly decides whether or not to convert. The choices are to convert either all or none of his or her shares.

8c. Partial-convertible shares: each participant secretly decides whether or not to convert. The participant may convert all, some, or none of his or her shares.

8d. Both non-convertible and all-or-none convertible shares: each participant secretly decides whether or not to convert all or none of his or her convertible shares.

9. The cards are revealed. The instructor announces the liquidation value of all shares. Participants' profit positions are updated.

Source: Kluger and Wyatt, 2004, p. 976

In those cohorts where all subjects had made judgment errors in the individual experiments, the prices in the subsequent market experiment did reflect this error. Many subjects did not make optimal use of their right to converse assets. Still, there is hope for clever traders. For in cohorts in which at least two rational traders were present—subjects who had not made a single error in the individual experiment assets prices were efficient.

The question whether irrational behavior of individual market participants may lead to inefficiency of the market as a whole is considered as one of the main challenges to behavioral finance.[5] The cognitive bias underlying the Monty Hall problem has its origins purely in limited computational capabilities. As long as some traders are smart enough not to make these mistakes, the bias does not show up in the aggregate, and prices are efficient. This is why the Monty bias is unlikely to appear in large, competitive markets. In this respect the effects of the bias differ from that of other cognitive biases detected in the behavioral finance literature, for example those that originate in preferences and expectations. Sometimes the arbitrage required to compensate for price inefficiencies is simply too costly and risky. For example, in some circumstances the irrationality of the participants in financial markets may increase. If the rational arbitrageur buys undervalued stocks, but market participants grow even more pessimistic, he will incur a loss, no matter how right he may be about fundamentals. This is the well-known 'noise trader' risk and even Marilyn vos Savant cannot get that out of the way.

FOOTNOTES

1. *The Times*: Brilliantly inventive. . .not simply the most original novel in years. . .also one of the best; *The Financial Times*: '. . .extraordinarily moving, often blackly funny. . .'; *TES*: 'A stroke of genius.'

2. Every year sees an update of the questions used. The most recent version is published by the Global-Investor Bookshop, 43 Chapel Street, Petersfield, Hampshire, GU32 3DY, UK; bookshop@global-investor.com

3. Friedman, D. (1998) Monty Hall's three doors: Construction and destruction of a choice anomaly, *American Economic Review*, 88, 933–946.

4. Kluger, B. D. and Wyatt, S. B. (2004) Are Judgment Errors reflected in market prices and allocations? Experimental evidence based on the Monty Hall problem, *The Journal of Finance*, Vol LIX (3), June, pp. 969–997.

5. See Prast, H. M. (2004) *Psychology in Financial Markets: an Introduction to Behavioural Finance*, Series Financial Monetary Studies, NIBE-SVV publishers, Amsterdam, May 2004, ISBN 90 5516 206 X.

12
Risk: The Ugly History

Aaron Brown*

The mathematical study of risk began in 1654 with a famous exchange of letters between Pierre de Fermat and Blaise Pascal. If you like you can push the date back to Isaac Newton in 1610, Gerolamo Cardano in 1525 or Luca Pacioli in 1458, but it is still remarkably late considering that gambling is a universal human activity far older than history. Why didn't some earlier mathematician consider the problem? Why didn't some earlier gambler publish some useful inductions from experience?

The usual explanation is that philosophical and theological obstacles hindered development. But this won't convince anyone trained in finance. The more society discouraged rational approaches to gambling, the greater the rewards to someone who mastered a basic principle or two. We know people were willing to exercise ingenuity to gain a gambling advantage, ancient loaded dice have been excavated. We know people were willing to study the subject, dicing schools and guilds existed in medieval Europe. We know many early mathematicians needed money and used their skill to get it. Two thousand six hundred years ago, Thales of Miletus, the first mathematician known by name, used a complicated analysis to make a fortune by cornering the olive press market. A small fraction of that effort would have provided a lifetime income from gambling.

After Fermat–Pascal

The mystery does not end in 1654. Fermat and Pascal argued over the probability of getting three or more heads in four flips of a fair coin. Compare that to the level of Fermat's number theory or Pascal's projective geometry—anyone with high school algebra can easily solve the probability question in their heads today, while the other work remains challenging to college mathematics majors. After 1654, it took 150 years to derive any results not regarded as trivial today, and 150 years after that to get a reasonably consistent mathematical theory. As late as the beginning of the twentieth century, elementary errors like Bertrand's Paradox were unresolved, and today simple questions like the Necktie Paradox or the definition of a random number do not have fully satisfactory answers. When Ed Thorp figured out how to beat casino blackjack in the 1960s, many

* I would like to thank Paul Wilmott, Deborah Pastor and Dan Tudball for helpful comments and suggestions.
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.

mathematically sophisticated people dismissed the work on the grounds that it was impossible to gain an advantage by varying the bet in a game where the average odds are against you.

It's true that we've had a mathematically rigorous foundation for probability since the 1930s, and not one but four consistent and sophisticated ways to link mathematical probability to risk (by Von Neumann, Arrow-Debreu, Savage and Shannon). But this work does not correspond well with the actual risk faced by humans. In 1921, Frank Knight distinguished between 'risk' and 'uncertainty'. With some oversimplification, he put everything modeled by probability theory in 'risk' and everything people wanted to know about for practical decision-making in 'uncertainty'. In 1972 Daniel Kahneman and Amos Tversky began a field of study that has demonstrated the enormous gap between mathematical and behavioral concepts of risk.

I cannot think of any field of study so basic to human survival that started so late, progressed so slowly or is in such unsatisfactory shape today. The problem is not only theoretical. Simple errors in risk calculations routinely cause large disasters. Individuals clearly mismanage risk according to any reasonable theory. Introduction to statistics is frequently the most unpopular course in the catalog. Elementary statistical principles are commonly ignored in law, which should be almost entirely statistical, and statistical expert witnesses can be found for either side of any case.

Linguistic view

The word 'risk' entered the English language in 1661. Although it comes from French and Italian, its origin and earlier history are unknown. Words that are related today had entirely different meanings. 'Random' meant fast to Shakespeare, only later acquiring a connotation of careless, then haphazard, then unpredictable. The root of 'danger' is the Latin 'dominus' meaning 'master'. The word evolved to mean 'under control of' then later 'liable to a master'. The transfer from the idea of liability or responsibility to a specific person to general possibility of harm came later. 'Peril' meant to try something.

Other risk-related words had specific gambling meanings rather than uncertainty in general. 'Hazard' comes from the Arabic for 'dice'. 'Chance' meant 'falling of the dice'.

Of course, the fact that modern words for risk did not have their contemporary meanings doesn't mean there weren't words for risk in English before the seventeenth century. 'Pleoh' is the Old English word usually translated as 'danger', but it has the sense of 'circumstance' like the modern 'plight'.

Going back further, the most familiar passage referring to risk in the Bible is Ecclesiastes 9:11, '...the race is not to the swift, nor the battle to the strong ... but time and chance happeneth to them all' in the King James translation. However the Hebrew word does not imply randomness but simple circumstance. It's more 'you win some, you lose some' than 'wins and losses are uncaused'. When the Philistines observe the path taken by an oxcart to see if their misfortunes are caused by the Hebrew God or by 'chance' (again King James translation), the Hebrew word only means 'some other cause'. There is no ancient source in the Jewish/Christian/Islamic tradition that clearly refers to the modern sense of risk. However, it's interesting that by 1611 the King James translators used a gambling metaphor to mean 'of unspecified or unknown cause'. This appears to be a concept born in the Reformation. It could not be described in existing languages, instead people applied gambling words beyond the gaming table, and gave gambling connotations to words that did not have them earlier.

To see the distinction, consider the following examples from different games:

- (Chess) Your queen is in danger!

- (Roulette) A martingale strategy of betting $1 on red and doubling the bet after every loss has the danger that black will come up enough times in a row that doubling your bet would exceed the house limit.
- (Poker) Your only danger is that your opponent has a full house.

In chess there is no concealed information, no randomness. ‘Danger’ here means circumstance. It would be pointless to gather historical statistics about how often queens are lost in chess games, or to buy an insurance policy on your queen. An appropriate response to this statement is to consider the ways the queen might be lost, and either prevent them or make sure you have an offsetting gain.

The roulette example is closest to the modern understanding of financial risk. There is no concealed information. For most purposes randomness is a useful model for the sequence of red and black. Historical statistics are certainly relevant here, and an insurance policy against a long sequence of blacks could offset the danger. The appropriate response to this kind of danger is to consider the probability and consequence of the bad event, and weigh that in an overall decision.

In poker after all the cards are dealt, there is no randomness, only concealed information. This is an important intermediate case between the first two examples. For the purposes of one hand, this is the chess situation. Your opponent either has the full house or she doesn’t. But for playing a long series of hands, you have to consider the probabilities. You even have to consider the probabilities of different hands you might have, conditional on the play of the hand up to this point. You know what you have, but your opponent’s actions will be influenced by what you might have. This gets us into the realm of game theory, where probability and strategy are mixed.

Now suppose I say ‘he’s in danger of getting fired’, or ‘having a heart attack’ or ‘being audited’. It’s not clear whether I mean the chess, roulette or poker sense of ‘danger’, or something intermediate. Of course, it’s even harder to interpret historical sources. The only two senses that are clearly more than 400 years old are the chess sense and the roulette sense specifically restricted to gambling games.

Risk and the law

The Code of Hammurabi seems to have no concept of accident or random event. For example, law 120 concerns ‘If any one store corn for safe keeping in another person’s house. . .’ It treats three cases identically: if ‘any harm happen to the corn in storage’, ‘if the owner of the house open the granary and take some of the corn’, or if the owner of the house ‘deny that the corn was stored in his house’. The penalty is the same for accidental loss, theft and fraud. The laws for physicians prescribe penalties if a patient dies without consideration of whether the physician was responsible. In certain cases, an accused is thrown into the river. The details for this are not known, but it is clear that people sometimes drowned and sometimes survived. In a more severe variant, the accused was tied up first, and rarely survived. In either case, the result is not viewed as luck, because if the accused survived, he or she was assumed innocent and the accuser was punished.

Hammurabi *is* careful to excuse liability for acts of God, and the distinction lives on in modern insurance policies. The related concept *force majeure* was also carved in stone in 1800 BCE, Hammurabi excused carriers whose goods were seized by war enemies. So, the Code assumes someone (maybe a god or a rival king) is responsible for everything, nothing is random, there is a legal consequence for every bad result (see the charming Australian comedy *The Man Who Sued God* for the obvious implication). Modern lawyers are not much friendlier to statistics than the ancient King of Babylon.

There are two tantalizing exceptions to the dearth of evidence for modern risk before 1650. The first is from the Hindu holy epic *Mahabharata*, in a section probably composed about 1700 years ago. The poem contains many accounts of people ruined by gambling, invariably because (a) they are obsessed and cannot stop playing despite pleas from friends and (b) the opponent cheats. One such victim, Nala, takes up service with a neighboring king. That king demonstrates his wisdom by counting the leaves and fruits on a large tree through examining a single twig. Nala offers to trade lessons in horsemanship for the secret to this feat. The king agrees, and tells Nala the secret will show him how to win at dice as well. Nala then goes home and wins back his kingdom. This appears not only to show a scientific knowledge of probability useful for dice playing, but connects that skill to reasoning from a sample. However, I know of no other evidence for ancient knowledge of either one.

The second exception also explains the term 'premium' for insurance payments and option prices. 'Bottomry' loans date back at least to the Phoenicians 3000 years ago. They are loans secured by a ship, with the loan forgiven if the ship is lost. Bottomry lenders were granted exemptions from usury laws and allowed to charge a premium to the legal rate of interest. The justification flirts with the idea of expected value, it is acknowledged that the interest on loans for successful voyages must cover the losses on loans for unsuccessful ones. Legal cases are preserved in which the amount of the premium is challenged, but to my knowledge the actual frequency of losses did not enter into the argument. The lender had to prove that he was exposed to significant maritime risk, but did not have to quantify that risk nor relate it to the premium charged.

A closer look at Fermat–Pascal

Fermat's solution to the interrupted game problem was a direct application of mathematical logic to law. Read carefully, it has nothing to do with probability in general. After some analysis the problem comes down to: how to divide a stake in an interrupted game of fair coin flipping, in which A needed m heads to win and B needed n tails first.

Using modern notation, call $A(m,n)$ A's proportion of the stake. Fermat reasoned that $A(k,k) = \frac{1}{2}$ for any k, because both players are in identical positions. Further, $A(m,n) = \frac{1}{2}[A(m-1,n) + A(m,n-1)]$. That allocation makes the next coin flip a simple wager of $\frac{1}{2}[A(m-1,n) - A(m,n-1)]$, which stake should be divided equally according to the first principle. Pascal provided the triangle (which he did not invent) to calculate these values quickly. Note, however, that the solution need not be the expected value of the outcome, in fact the coin probabilities are never used. All you need is a principle for dividing the stake when both parties are in the same situation and *reductio* (one of Fermat's signature techniques in number theory) provides the answer. This is much more similar to the binary version of the Black–Scholes argument than to the binomial distribution in statistics. It applies more generally than the expected value approach.

Pascal then made a fateful error. He confused the equity argument of Fermat with a frequentist argument, essentially that the stake should be divided by Monte Carlo simulation—completing the game many times and dividing the stake in proportion to the number of each player's wins. These are only the same if (a) the outcomes are equally probable (i.e. the coin is fair) and (b) you can rely on the principle that stakes should be divided equally when a tie game is interrupted (this does not work if the stake cannot be divided, and may not be the proper resolution in other cases).

To highlight the difference between these approaches, consider the modern rules for interrupted major league baseball innings (usually as a result of rain). If a game is stopped at the end of an inning, that is, if both teams have had the same number of opportunities, the game is awarded to the team with more runs.[1] The problem arises if the game is stopped mid-inning, when one team (the visiting team) has had more opportunities than the other (the home team).

Pascal's frequentist approach would consider the situation at the interruption and compute the probabilities of different outcomes for the remainder of the inning. Each team would be awarded a share of the win based on its probability of being ahead had the game been completed. The probabilities could be estimated from historical data. For example, if the score is tied in the sixth inning, and the home team has a man on first with one out, then it wins with probability 0.561, based on analysis of completed games. So we could award the home team 0.561 wins and the visiting team 0.439.

The actual rule has Fermatian spirit. There are no splits win, nor any concept of awarding based on expected value at time of interruption. If the home team is ahead, it wins the game. If the visiting team was ahead at the beginning of the inning, the final score stands. In all other cases the game is not awarded to either team (depending on circumstances, it will be ignored, or completed or replayed on another day).

These rules are confusing to baseball fans and are inconsistent with expectation. A team can get credit for a win in a game it had less than an even chance of winning, and not get credit for a win in a game it was almost certain of winning. The home team has a probabilistic advantage. The rule is symmetrical if either team is ahead at the beginning of the inning, but if the game is tied at that point, the home team can get a win but the visiting team cannot. Also, it's easier for the home team to win if the game is stopped in the fifth inning.

But there is a clear argument from justice. We know that it is fair to award the game to the team that is ahead at the end of an inning. We extend the principle to say a team is entitled to a win if it is ahead at the end of an inning and when the game is stopped. The visiting team wins if it is ahead at the end of the last complete inning, and also at the time play is stopped. The home team wins if it is ahead when play is stopped, because that means it would necessarily be ahead if the partial inning were completed.

There is an inconsistency here: the visiting team's argument is based on being ahead at the start of the partial inning, while the home team's case is based on being ahead at the end of the partial inning. Logically, it's possible for one team to be ahead at the beginning and the other at the end, meaning that it's fair to award the game to both teams. The inconsistency is ignored, the visiting team gets the win, even if it's possible the home team would have gone ahead if it had been allowed to complete the partial inning. The home team gets the win even if the visiting team was ahead at the beginning of the inning. This, plus the fact that the rule often fails to award the win to either team, makes it unsatisfactory as a quantitative model. It is statistically unfair. But many legal principles are based on this type of reasoning.

For the next two centuries, probability theory was hobbled by the supposed need to break problems down into equiprobable events for rigor. Practical statisticians used frequentist approaches instead, but without any rigorous foundation. That would require measure theory in the twentieth century, which provided the foundation but at the price of assumptions inapplicable to any real problems outside quantum physics. Bayesian statisticians attempted to avoid the dilemma by positing a subjective prior distribution, which ensures consistency but defines probability only subjectively. Nonparametric statistics also avoid the dilemma, but sacrifices a lot of power. The future of practical statistics probably lies in some combination of frequentism and nonparametric methods, something like the bootstrap, although so far no one has even figured out how to do good bootstrap regression analysis.

As a result of this confusion of theoretical approach, conventional statistics gets the world backward. It starts by introducing randomness, through the abstraction of a random variable generated by a distribution. The problem isn't that the world is predictable and we need a mechanism to introduce randomness; it's that the world is unpredictable and we need mathematics to make predictions. We have the data, and want to know the probability that a hypothesis is true, or that some future event will happen. Instead, statistics tells us to form a 'null hypothesis', generally the opposite of what we want to know, and pretend the data are random (although we have already made the measurements), then compute the probability of getting the data under the null hypothesis. The answer is not the probability or prediction we want, but something called a 'significance' which is hard to define and obeys no consistency rules. The entire process is impossibly abstract, irritates students to no end, takes great skill to perform properly and is subject to well-known paradoxes.

Despite these criticisms, probability theory works extremely well in analyzing gambling games (for which the assumption of introduced randomness is literally true) and for controlled experiments (experiments that have been converted to gambling games for the convenience of statisticians). It works for quantum physics, although a lot of people think it shouldn't. Its record in other fields is mixed. Smart, experienced, honest people, trained in probability theory and statistical practice, can use mathematics to extract truth from an uncertain world. Certain tools developed for specialized problems work reasonably well most of the time. But these qualifications aside, most applications of statistics to uncontrolled situations are problematic.

The breakthrough

For 318 years, there were no statistical methods that combined rigor and practicality. The breakthrough came in 1972, with the publication of the Black-Scholes model. This led to modern derivative pricing which distinguishes between risk–neutral and simulation pricing. The first is firmly in the spirit of Fermat. It requires no assumptions about probabilities, no historical market data and no projections of future market behavior. However, it does require complete market data, which restricts its use to some simple (but important) special cases.

Monte Carlo simulation is Pascal's frequentist approach. It can price anything, but only with a probability model and the assumption that price equals risk-adjusted expected value. Many popular methods combine these two approaches, and in general they serve to verify each other. We only really trust prices for which we can get reasonably tight value ranges by both analytic methods and historical data.

Finance is the first field to recognize the difference explicitly and accept both answers. Unification is still an important goal, but quantitative finance can make rapid progress in both theory and practice without it. This progress holds the best hope for solving the problem and making the history of risk handsome again. Once we've figured out how to price and risk manage everything, then we can fix statistics and figure out the rest of the universe.

FOOTNOTE

1. More than four innings must have been played or it is declared 'no game' and doesn't count for either team.

13 Finformatics: Thirst for Hurst

Kent Osband

Why we have more to thank the Nile for than just an interesting delta.

How does the risk of a trade change with the time T you hold it? When returns for each moment are independent and identically distributed (i.i.d.), the answer is easy. Variance will scale linearly with T, so the volatility will scale with $\sqrt{T}$. With only a bit more effort we can show that skewness of i.i.d. sums shrinks according to $T^{-3/2}$ while the corresponding kurtosis shrinks with T^{-1}.

That suggests a neat way of checking for serial independence. Calculate the volatility, skewness and kurtosis over different holding periods and watch how they scale. Lo and behold we find that the shorter the holding period, the less likely financial market returns are to scale like i.i.d sums.

Hurst exponents

It would be useful to have a single summary measure of how risk scales over time. Statisticians found one in the work of Harold Edwin Hurst, a British civil servant posted to Cairo in 1906. Hurst got interested in Nile flooding as this was crucial to Egyptian grain harvests. Studying 800 years of records, he noticed that high-flood Nile regimes tended to alternate with low-flood regimes, each of them lasting for years at time. (Those of you who've read the Old Testament should know that already.) To summarize his findings, he calculated a coefficient known today as the Hurst exponent.

A Hurst exponent of H basically means that the standard deviation for a holding period T scales with T^H. Hurst exponents higher than 0.5 are said to describe 'persistence': high deviations tend to be followed by low deviations and vice versa. Hurst exponents less than 0.5 are said to describe 'anti-persistence': deviations tend to mean-revert. A Hurst exponent of 0.5 is a good sign of independence.

While positive auto-correlation causes persistence and negative auto-correlation causes anti-persistence, persistence can capture dependencies too subtle or complex to be summarized as correlation. Consider, for example, Brownian motion stuck between two reflecting barriers (the best financial analogue is a currency floating inside a fixed band). The probability of being constrained in the next instant is zero, so the Hurst exponent for short periods must asymptote to 1/2. However, in the long run the standard deviation is capped by the width of the band, so the Hurst exponent must shrink to zero.

Now, Hurst didn't actually calculate multi-period standard deviations. Rather he calculated a related quantity known today as 'rescaled range' or 'the R/S statistic'. To calculate it yourself:

- Divide up the whole period into subperiods of equal length T, preferably non-overlapping to reduce correlation across measures.
- Remeasure each observation as a deviation from the sample mean, so as to mitigate the impact of drift.
- Add the deviations cumulatively.
- Measure the range R as the difference between the cumulative high and the cumulative low.
- Divide by the sample standard deviation S of the observations.
- Average over the various subperiods.

With independent observations, Hurst knew the R/S statistic should converge to $T^{1/2}$ times some constant (I'm not sure Hurst knew the constant. Do you? Hint: Monte Carlo simulations show it's about 1.6). Instead he found the R/S statistic for the Nile was nearly proportional to $T^{0.7}$.

Highfalutin stuff

The R/S statistic has found favor with high-powered finance theorists despite its being relatively easy to calculate. That's because it exists even for non-Gaussian Levy distributions. Actually that's not quite true. Since none of those distributions have well-defined standard deviations you can't divide by S. And half of Levy distributions are too diffuse to have a well-defined mean—the Cauchy distribution marks the divide—so they don't have an expected range either. But the Levy distributions that most theorists are interested in, marked by a tail parameter μ between 1 and 2, do allow an expected range. Moreover, the Hurst exponent for such distributions is constant, with $H = 1/\mu$.

That's not all. Recall that even ardent Levy distribution fans concede they have to truncate them to fit the evidence for finite volatility. But truncation doesn't impair the Hurst exponent at short holding periods. Eventually the Hurst exponent for truncated Levy distributions recedes to 1/2, implying no dependence across long holding periods, but that's what the empirical evidence suggests too. So at first glance it appears as yet another vindication of Levy truncation.

Hurst exponents also are easy to match up with another phenomenon called fractional Brownian motion. Indeed, fractional Brownian motion is defined as the continuous Gaussian process B_H with covariance between values at times t and s of t$\frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H})$ for H, the Hurst exponent. Sadly this is neither a Markov process nor a semimartingale and can't be analyzed using standard Ito stochastic calculus. You have to apply a new fractional stochastic calculus that sprinkles in terms in $\sigma^2 t^{2H}$ everywhere $\sigma^2 t$ appears in standard Ito, ultimately generating a fractional Black–Scholes equation of the form:

$$\frac{\partial V}{\partial t} + H\sigma^2 t^{2H-1} S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rD = 0$$

Clearly this is a fascinating road to travel down, but you'll have to do it without me as I promised to feed my kids' goldfish while they're away. Just remember that to fit the data you'll need to let H recede to 1/2 over time and modify all your equations accordingly.

Simple regime-switching explanations

At first glance, regime-switching under uncertainty can explain persistence just as easily as it explains fat tails. As long as the regime stays the same, the associated drift stays the same, and investors' beliefs will gravitate toward recognizing it. That's persistence. Eventually the regime will change, the drift will change, and investors' beliefs will follow along. Given a sufficiently long time period, regimes and beliefs will switch enough to wash persistence away.

To test this hypothesis, I will perform thousands of Monte Carlo simulations and estimate the Hurst exponents. By running so many simulations I can avoid some of the problems that bedevil Hurst exponent estimation in real life. It won't avoid them all, though. We'll need to make some additional adjustments.

To begin with, we need to be careful about the starting point. For example, if the initial regime is High, should I assume the aggregate market investor knows it ($p_0 = 1$), the investor is completely uncertain ($p_0 = 1/2$), or that the starting belief p_0 is drawn randomly from the stationary distribution f^*? The answer is, none of the above. What we want is the stationary distribution g^* of beliefs given that the initial regime is High. Let us assume High and Low regimes are equally likely in equilibrium; i.e. $\lambda_{HL} = \lambda_{LH} \equiv \lambda$. Applying Bayes' rule and substituting previously derived results,

$$\begin{aligned} g^*(p) &\equiv \text{density}(p|\text{High}) = \frac{\text{density}(p \cap \text{High})}{\text{probability(High)}} = 2pf^*(p) \\ &\propto \frac{1}{p(1-p)^2} \exp\left(-\frac{2Q}{p(1-p)}\right) \text{ for } Q \equiv \lambda/S^2 \end{aligned}$$

Figure 1 provides charts of g^* versus p^* for $Q = 0.25$ and $Q = 0.05$ respectively.

How to lie with rescaling

Here's a more serious problem, discovered in the course of running simulations. The R/S statistic can easily generate spurious evidence of persistence and anti-persistence. Indeed it has to if you detrend every single sample by its own sample mean. That guarantees a detrended return of zero for a one-period holding period; ergo a one-day rescaled range of zero. To generate a positive rescaled range for longer holding periods, the logarithm must jump, not just scale smoothly with the square root of time.

This effect is not just confined to the origin. At any holding period, detrending trims the average range below the levels that would prevail for a genuinely zero drift series, since zero expected drift doesn't ordinarily mean no net price movement. But since the relative impact is more pronounced for shorter series, the R/S statistic scales faster than the square root of time.

Moreover, the ideal R/S statistic should cover the encompass the full intra-day trading range, not just the closing price. For long holding periods this hardly matters. For short periods it matters a lot. This too shows up as spurious evidence for persistence.

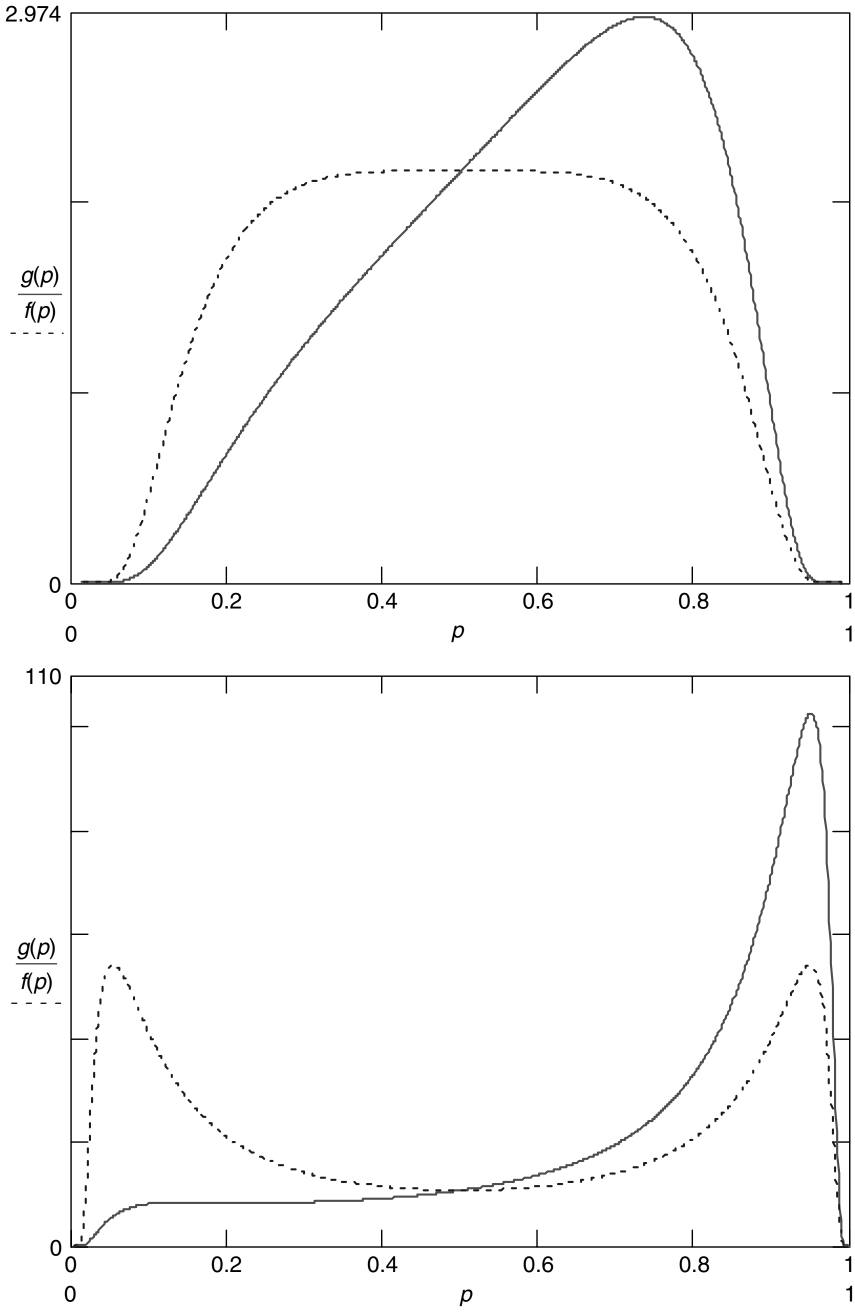


Figure 1: Charts of g^* versus p^*

Figure 2 is a chart in log-log space of the average R/S statistic for a purely random walk, when all samples are individually detrended and the holding period stretches from 2 days to 4096 days. I found 500 samples for each holding period gave fairly robust averages, but calculated 2000 samples each for good measure.

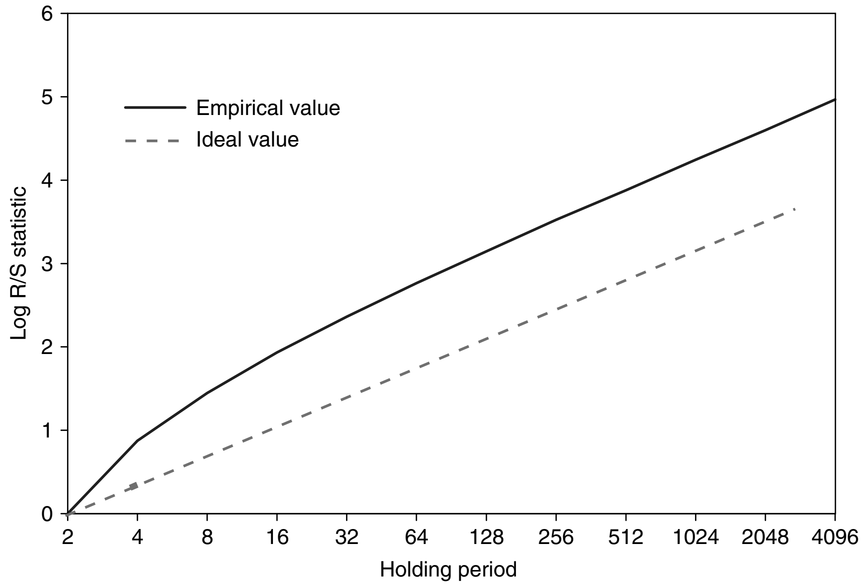


Figure 2: Over-detrended R/S for a random walk

The best linear fit indicates an average persistence of 0.61. A quadratic curve fits the data even better: the R^2 is 0.996. It suggests persistence is 0.9 at two days but shrinks to 0.5 or less at very long horizons. And yet all the observations are independent by construction.

Several writers have noted that the evidence for persistence isn't all it's cracked up to be. The demonstration above is an additional warning. If you're using a range statistic to measure persistence, detrend it only by the true mean or by the mean of a large aggregate of samples. Otherwise your evidence is likely to be too undevious for its own good.

Simulation results

I set up a regime-switching model with annualized values of $\mu_H = -\mu_L = -12\%$, $\sigma = 8\%$, $r = 4\%$, and $\lambda = 0.25$. Seeding beliefs with a random draw from the appropriate conditionally stationary distribution, I ran a Monte Carlo simulation for 4096 days of trading, or just over 16 years. Each day the model updated the regime, the random dividend, beliefs about the regime and price. I then calculated the range for every power-of-two holding period; i.e. 1 day, 2 days, 4 days, and so on up to 4096 days.

Again, that was the setup for one simulation. I ran 1000 such simulations and averaged the ranges. That provided over 16 000 years of simulated daily financial data, a bit more than you are likely to collect in practice.

As it turned out, I had to run the full set of simulations several times, to experiment with various detrending techniques. The first set detrended each sample by its own sample mean, because I didn't initially realize the danger it posed. Not surprisingly I found a lot of persistence at the short end. But after I saw the results for a pure random walk I realized this might be spurious. More convincing was the evidence that persistence declined with holding period. The curve flattened out for long holding periods much more for the regime-switching-under-uncertainty process than for the pure random walk.

I then ran another 1000 simulations without detrending. Figure 3 is the chart in log-log space of range versus holding period. The general shape is very similar to the shape created by excess detrending. The main difference is that excess detrending concentrates most of its force at the short tend, while regime switching has a more even impact.

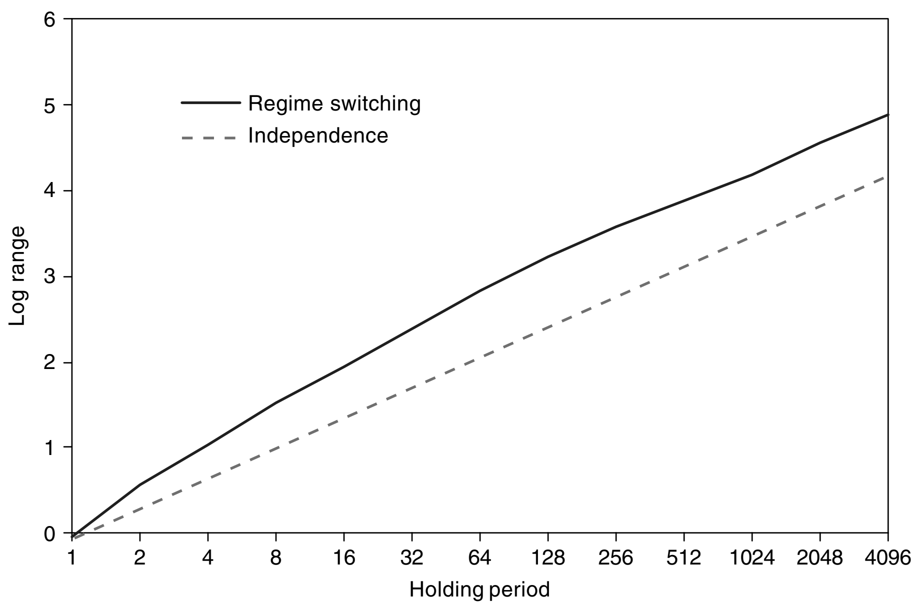


Figure 3: Range without detrending

In this example, the best linear fit has an R^2 of 0.993 and implies a Hurst exponent of 0.58. The best quadratic fit has an R^2 of 0.9997, and implies the Hurst exponent declining from 0.78 to just under 0.5.

In other words, finformatics predicts both persistence and gradual decay of persistence as well as any truncated Levy distribution. It also offers a plausible explanation while the Levy truncators cannot. However, a kind of measurement error can also account for significant persistence and decay of persistence. I am not sufficiently steeped in the evidence to gauge which explanation is more important and by how much.

14 TARNs: Models, Valuation, Risk Sensitivities

Vladimir V. Piterbarg*

We study a new class of interest rate exotics, Targeted Redemption Notes, from the financial modeling perspective. We discuss issues of model selection, develop methods and techniques for valuation, and present approaches for improving numerical properties of risk sensitivities calculations.

1 Introduction

The market for exotic interest rate derivatives has long been dominated by callable Libor exotics. Recently, however, there have been new developments. In addition to new flavors of callable Libor exotics,[1] completely different type of structured interest rate note, called TARN (Targeted Redemption Note), has been introduced.

A comprehensive framework for modeling callable Libor exotics has been developed in a number of papers (Piterbarg 2003, 2004a, 2004b). Unfortunately, many of the methods do not directly apply to TARNs and need to be adapted or replaced. This chapter aims to accomplish just that. We develop and present various methods for analyzing and pricing TARNs and, most importantly, computing their risk sensitivities (Greeks) quickly and efficiently.

A forward Libor model is the workhorse of exotic interest rate modeling. Flexibility of its volatility specification allows calibration to a wide range of market instruments, while controlling forward evolution of the volatility structure. This flexibility comes at a price, as typically only Monte Carlo methods are available for valuation. We start by discussing various variance reduction methods available, and their applications to TARNs. We then proceed to analyze the volatility structure dependence of TARNs in detail. This allows us to adapt a powerful 'local projection'

*I would like to thank Paul Cloke, Leif Andersen and Jesper Andreasen for valuable discussions and feedback.

method to the problem of TARN valuation. The method is based on calibrating a 'local', low-dimensional, PDE-based model to the deal-specific volatility structure components of a 'global' forward Libor model. We show that TARNs are particularly well suited for the method. After reviewing volatility smile dependence of TARNs, we present our 'local' model of choice. Finally, we spend some time demonstrating how TARNs, despite being path-dependent, can be valued by PDE methods.

2 Definitions

The interest rate exotics market, just like all other exotics markets, is driven by investors' interest in structured notes. A structured note works like a bond, where an investor gives a principal to an issuer in return for a promised stream of coupons (that are linked to interest rates at the time when the coupon is set), and a principal repayment at the end of the note.

Investors are primarily interested in receiving a rate of return that is as high as possible, as well as in an opportunity to express a view on future directions of interest rates. A common way to increase the coupon paid to an investor has been to make the note callable (Bermuda-style) by the issuer. While offering an enhanced yield, this feature was not necessarily liked by investors as they typically had no way of knowing when the note would be called.

A recent innovation, an invention of a Targeted Redemption Note, has solved this problem. In a TARN, a structured coupon is paid to an investor. The total return, the sum of all coupons paid to date, is kept track of. When the total return exceeds a pre-agreed target (hence the name of the instrument) the note is terminated. No further coupons are paid, and the principal is returned to the investor.

Issuers do not keep these complicated instruments on their books, and swap them with exotic interest rate trading desks. The principal payment from investors is reinvested at the Libor rate. Thus, from the point of view of the trading desk, a cancelable note looks like a callable Libor exotic. A TARN then looks like an exotic swap that knocks out on the total sum of structured coupons.

Let us define a TARN formally. We sacrifice some generality in the description of the contract for the sake of simplicity, while retaining the features of the contract essential from the modeling prospective.

A TARN is based on a tenor structure, a sequence of times spaced roughly equally apart,

$$0 = T_0 < T_1 < \cdots < T_N,$$
$$\delta_i = T_{i+1} - T_i.$$

We denote zero coupon discount bonds by $P(t, T)$. Forward Libor rates are defined by

$$F(t, T, S) = \frac{P(t, T) - P(t, S)}{(S - T) P(t, S)}.$$

In particular, we define

$$F_n(t) = F(t, T_n, T_{n+1}) = \frac{P(t, T_n) - P(t, T_{n+1})}{\delta_n P(t, T_{n+1})},$$

$n = 0, \ldots, N-1$. The structured coupon is an inverse floating coupon[2] based on the Libor rate. With the strike s, it is defined as

$$C_n(t) = (s - 2F_n(t))^+,$$

observed (fixed) at time T_n and paid at T_{n+1}. This is the coupon promised to an investor. In return, a floating rate payment based on the Libor rate is made. The coupon fixed at time T_n is only paid if the sum of structured coupons up to (and not including) time T_n is below a total return R. Thus, the value of the TARN at time 0 from the investor's viewpoint is given by

$$\begin{aligned}
v &= \mathbf{E}_0\left(\sum_{n=1}^{N-1} B_{T_{n+1}}^{-1} \times X_n(T_n) \times \chi\{Q_n < R\}\right),\\
X_n(t) &= \delta_n \times (C_n(t) - F_n(t)),\\
Q_n &= \sum_{i=1}^{n-1} \delta_i C_i(T_i),\\
Q_1 &= 0, \qquad (2.1)\\
\chi\{A\} &= \begin{cases} 1, & \text{if } A,\\ 0, & \text{if not } A. \end{cases}
\end{aligned}$$

We note that a TARN typically pays some fixed coupons to an investor up front. We do not include them into the contract description as they can be valued off an interest rate curve separately, as they are known in advance.

Let us consider an example. A typical deal at the time of writing has $T_N = 10$ years, $\delta = 1y$, $s = 11.5\%$ and $R = 3\%$. Moreover, 11% (per annum) is paid to the investor up front in the first year. The fixed coupon of 11% in the first year is clearly very high, well above anything available from government bonds. Therein lies the main attraction for the investor.

TARNs are highly leveraged investments. This should be clear from the high up-front coupon that a trading desk is willing to pay. Let us analyze this in a bit more detail. The investor clearly wins if the deal knocks out quickly—he is left with the 14% return (11% up front plus 3% targeted return), and is repaid his principal upon knockout. The deal knocks out on the first possible date (T_2) if $C_1(T_1)$ is 3% or above, or equivalently $F_1(T_1)$ is 4.25% or below. On the contrary, if the rates go up (above 5.75%), and stay up there for 10 years, all coupons C_n become zero, the investor receives nothing for 10 years but has to pay Libor (essentially, he forfeits interest on the principal for 10 years). Figure 1 shows the (risk-neutral) probability of the deal being alive after successive years. The deal stays alive for 10 years (bad for an investor) with about 25% probability, and knocks out after the first two years (good for the investor) with 65% probability. In other words, an investor makes decent money with 65% chance, and loses big with 25% chance (it is awash with 10% chance). This again demonstrates the measure of leverage in TARNs. It is not hard to see that the smaller the target return R, the higher the leverage.

3 Forward Libor models

The first model anyone should apply to a new type of an interest rate exotic, such as a TARN, should be a flexible, fully calibrated (to the full swaption volatility ATM grid and, if possible,

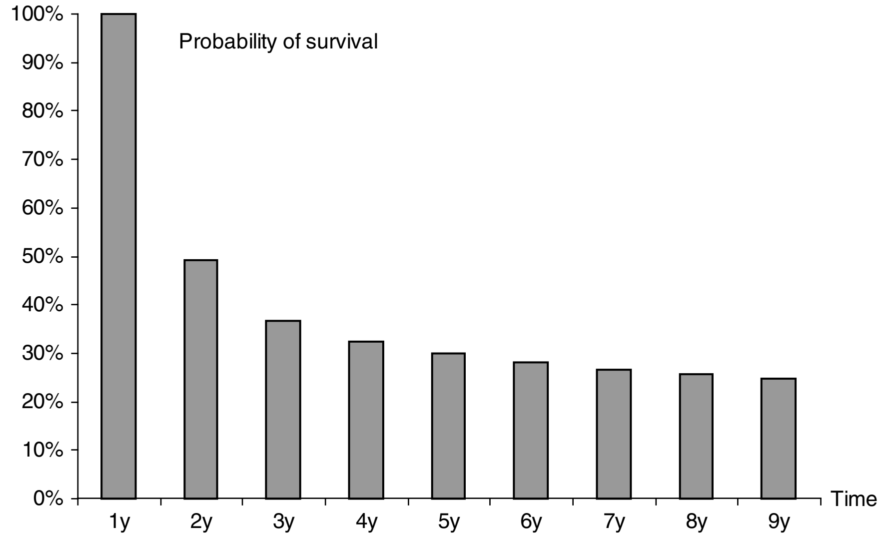


Figure 1: Probability that a TARN is not knocked out on and including given year. Time to maturity 10 years, total promised rate of return 3%, strike 11.5%

volatility smiles as well), 'global' model such as a forward Libor model. Before enough experience with a particular deal type is gained, this approach provides a measure of comfort by the fact that all market information is calibrated to. Using a less flexible model such as the Hull–White model requires one to choose what market volatility information to calibrate the model to. Such judgement is very hard to make and defend. In addition, using a flexible model such as a forward Libor model also allows one to control the dynamics of the volatility structure, something that typically plays an important role in valuation of exotics. Again, 'lesser' models impose that evolution upon the user, and more often than not it cannot be described as reasonable.

We start the definition of the model by specifying a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, together with a sigma-algebra filtration $\{\mathcal{F}(t)\}_{t=0}^{\infty}$.

Different flavors of forward Libor models are available. To avoid burdening this chapter with unnecessary details, and yet to present relevant issues in sufficient generality, we choose to work in the context of a skew-extended forward Libor model (Andersen and Andreasen 2000). The skew-extended forward Libor model introduces a local volatility function $\phi(x)$, independent of time, that is applied to each of the Libor rates. Moreover, we choose to present our analysis for a one-factor model that is based on the same tenor structure as the TARN in the previous section (both restrictions are non-essential for our results). The dynamics (under the appropriate forward measures) for each F_n is given by

$$dF_n(t) = \lambda_n(t)\,\phi(F_n(t))\,dW^{T_{n+1}}(t), \quad n = 1, \ldots, N-1, \quad t \in [0, T_n]. \tag{3.1}$$

A popular choice for $\phi(x)$ is a linear function

$$\phi(x) = ax + b,$$

resulting in a 'displaced-diffusion' type model. Another popular choice, a power function $\phi(x) = x^c$ defines a CEV-type model.

For convenience we define

$$F_n(t) = F_n(T_n), \quad t > T_n.$$

A special numeraire is usually chosen. We define a discrete money-market numeraire B_t by

$$\begin{aligned} B_{T_0} &= 1, \\ B_{T_{n+1}} &= B_{T_n} \times (1 + \delta_n F_n(T_n)), \quad 1 \le n < N, \\ B_t &= P(t, T_{n+1}) B_{T_{n+1}}, \quad t \in \left[T_n, T_{n+1}\right]. \end{aligned} \tag{3.2}$$

The dynamics of all forward Libor rates under the same measure, the measure associated with B_t, is given by

$$dF_n(t) = \lambda_n(t)\phi(F_n(t)) \sum_{j=1}^{n} 1_{\{t<T_j\}} \frac{\delta_j \phi\left(F_j(t)\right)}{1 + \delta_j F_j(t)} \lambda_j(t)\, dt + \lambda_n(t)\phi(F_n(t))\, dW(t), \tag{3.3}$$
$$n = 1, \ldots, N-1,$$

where dW is a Brownian motion under this measure, assumed to be **P**.

We note that the vector-valued process

$$\overline{F}(t) = (F_0(t), F_1(t), \ldots, F_{N-1}(t))$$

is Markov.

Algorithmically, pricing TARNs in a forward Libor model does not present major challenges. As a purely path-dependent contract with no optimal exercise features, a Monte Carlo simulation is straightforward. A TARN, however, has digital-type discontinuities (it knocks out). Simulation error (by which we mean the standard deviation of the Monte Carlo estimate of the true price) is higher for non-smooth payoffs. The problem of noise is especially severe for payoffs with digital-type discontinuities. The noise in the simulated *value* can be controlled relatively successfully by increasing the number of paths. *Risk sensitivities*, however, are a different story. The number of paths required to get a reasonably accurate estimate of risk sensitivities of a payoff with digital discontinuities is very high, and may make the application of forward Libor models impractical. This is particularly a problem for interest rate derivatives as usually a large number of sensitivities are required (the requirement to compute bucketed deltas, gammas and vegas can easily push the required number of valuations for a full set of risk reports into hundreds).

Limitations of Monte Carlo methods for computing risk sensitivities have long been noted, and various methods to alleviate them have been proposed. The book *Monte Carlo Methods in Financial Engineering by* Paul Glasserman (2003) has a wealth of information on the subject. In the next few sections we review some of them with a view towards an application to TARNs. We start by quickly reviewing the methods that actually do not work that well.

4 Pathwise and likelihood ratio differentiation

Methods for computing risk sensitivities in Monte Carlo that do not require separate simulations for the 'base' and 'bumped' value have proven to be extremely successful in certain applications. The methods are very effective for callable Libor exotics, see Piterbarg (2003). There are two types of methods in this category. One is the pathwise differentiation method. In it, the payoff of the underlying is differentiated analytically for each simulated path, and risk sensitivities are computed in the same simulation as the value. The other is the likelihood method, which shifts the differentiation to the density of the process being simulated (both are covered in Glasserman (2003), Chapter 7). The pathwise method is the better (often much better) of the two. Unfortunately, it requires absolute continuity of the payoff, a requirement that precludes its application to TARNs. The likelihood method is applicable, and can be implemented straightforwardly following Glasserman and Zhao (1999). Unfortunately, we found out that the likelihood method is not very effective for TARNs. The main reason is that the standard error of Greeks in the likelihood method is inversely proportional to the time to the first digital. In particular, it blows up as the fixing date is approached. The number of paths required to get decent simulation error is very large.

In view of these results, it appears that there is no alternative but to use the 'bump-and-revalue' method of computing risk sensitivities. So we shift our focus on reducing the simulation noise in Monte Carlo valuation. This is done by smoothing the discontinuities in the payoff.

5 Smoothing by conditioning

Intuitively, it is clear that the biggest contributor to the simulation noise is the first digital, i.e. the feature of the contract that specifies that it knocks out on T_2 if $F_1(T_1)$ is below a certain barrier.[3] The variance of the estimate can be reduced if we could somehow handle this digital explicitly, outside of the Monte Carlo simulation.

To develop the idea formally, let us define

$$S_n = \left(s - \left(R - Q_n\right)/\delta_n\right)/2.$$

In particular,

$$S_1 = \left(s - R/\delta_1\right)/2,$$

and

$$\{Q_2 < R\} \Longleftrightarrow \{F_1(T_1) > S_1\}.$$

Denote by V the value of the coupons that depend on the first knockout (the first coupon $X_1(T_1)$ is paid always and is easy to handle separately)

$$V = \sum_{n=2}^{N-1} B_{T_{n+1}}^{-1} \times X_n(T_n) \times \chi\{Q_n < R\}.$$

Then

$$
\begin{aligned}
v &= \mathbf{E}_0\left(V\right) \\
&= \mathbf{E}_0\left(V \mid F_1\left(T_1\right) > S_1\right) \mathbf{P}_0\left(F_1\left(T_1\right) > S_1\right) + \mathbf{E}_0\left(V \mid F_1\left(T_1\right) \leq S_1\right) \mathbf{P}_0\left(F_1\left(T_1\right) \leq S_1\right).
\end{aligned}
$$

Clearly

$$
\mathbf{E}_0\left(V \mid F_1\left(T_1\right) \leq S_1\right) = 0
$$

so that

$$
v = \mathbf{E}_0\left(V \mid F_1\left(T_1\right) > S_1\right) \mathbf{P}_0\left(F_1\left(T_1\right) > S_1\right).
$$

The first important observation is that the probability of not-knocking-out $\mathbf{P}_0\left(F_1\left(T_1\right) > S_1\right)$ can be computed analytically (or analytically approximated with a high degree of precision). The time T_1 is usually quite short, 1 year or less. Thus, high-quality approximations to the distribution of $F_1\left(T_1\right)$ can be obtained in almost any forward Libor model. For the particular case of a skew-enhanced one, see e.g. Andersen and Andreasen (2000). Since the time to expiry is short, the issue of non-deterministic drift of $F_1\left(T_1\right)$ under the spot Libor measure can be easily dealt with by, for example, freezing the drift along the forward value of the interest rate curve.

The value $\mathbf{E}_0\left(V \mid F_1\left(T_1\right) > S_1\right)$ is interpreted as the value of the TARN under the condition that it does not knock out on the date T_1. This value can be computed in a Monte Carlo simulation by adjusting the drifts of the forward Libor model in such a way as to move the Libor rate F_1 'away' from the knockout region. We do not go into details as we will present a more general scheme shortly.

The idea just presented is related to discontinuity smoothing via conditional expectations, see Glasserman (2003, Section 7.2.3). In addition, it can be viewed as a special form of importance sampling, as will be clear shortly.

The ability to remove the first discontinuity from the payoff being calculated by Monte Carlo reduces the simulation error substantially. We, however, can go even further. Given the information available on the coupon date T_n, we can evaluate the probability of knockout on the next day (quasi) analytically (for the same reasons we can compute $\mathbf{P}_0\left(F_1\left(T_1\right) \leq S_1\right)$—the time to expiry for the digital option in question is short, and excellent approximations to the distribution of the relevant Libor rate are available. Next, we develop a scheme where we integrate *all* discontinuities outside of Monte Carlo.

By an argument detailed in the Appendix, the trade can be valued as follows,

$$
v = \sum_{n=1}^{N-1} \tilde{\mathbf{E}}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \psi_n\right), \tag{5.1}
$$

$$
\psi_n = \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right). \tag{5.2}
$$

Here the measure $\tilde{\mathbf{P}}$ is defined by its Radon–Nikodym derivative with respect to $\mathbf{P}$,

$$
\left.\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\right|_{\mathcal{F}(t)} = \Lambda\left(t\right),
$$

where $\Lambda(t)$ is a non-negative, normalized $\mathbf{P}$-martingale such that

$$\Lambda(t) = \frac{\mathbf{P}_t\left(F_{m+1}\left(T_{m+1}\right) > S_{m+1}\right)}{\mathbf{P}_{T_m}\left(F_{m+1}\left(T_{m+1}\right) > S_{m+1}\right)} \times \prod_{k=1}^{m-1} \frac{\chi\left\{F_{k+1}\left(T_{k+1}\right) > S_{k+1}\right\}}{\mathbf{P}_{T_n}\left(F_{k+1}\left(T_{k+1}\right) > S_{k+1}\right)},$$
$$t \in [T_m, T_{m+1}).$$

This formula (see (5.1)) specifies that the value of a TARN can be computed by Monte Carlo simulation under the measure $\tilde{\mathbf{P}}$ by adding the values of coupons X_n weighted by weights ψ_n. The difference between (5.1) and (2.1) is in weights multiplying the coupons. In the former, these are ψ_n and in the latter, the weights are the knockout indicators $\chi\{Q_n < R\}$. Obviously, the ψ_ns are much smoother functions of a simulated path than the indicators, as in the former the digital discontinuities have been integrated away by computing the probabilities $\mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)$ in (5.2) (quasi) analytically.

The value v is computed by Monte Carlo simulation under the measure $\tilde{\mathbf{P}}$. This amounts to changing the drift of the Brownian motion driving the forward Libor model, see the SDE (A.1) in the Appendix. The measure $\tilde{\mathbf{P}}$ can be seen as the measure under which the TARN never knocks out. The method, which is based on the idea from Glasserman and Staum (2001), can be seen as a flavor of the importance sampling method, where the measure is changed from $\mathbf{P}$ to $\tilde{\mathbf{P}}$, and the likelihood ratio is partially pre-integrated.

Another, quite different approach to smoothing the payoff is detailed in the next section. While it is less effective, it can be simpler to implement.

6 Smoothing by 'sausage' Monte Carlo

Recall the main valuation formula (2.1). If $\omega_j,\ j = 1, \ldots, J$, are simulated paths of interest rates, then the discretized analog of (2.1) is given by

$$\begin{aligned} v &\approx \frac{1}{J} \sum_{j=1}^{J} v_j, \\ v_j &= \sum_{n=1}^{N-1} B_{T_{n+1}}^{-1}\left(\omega_j\right) \times X_n\left(T_n, \omega_j\right) \times \chi\left\{Q_n\left(\omega_j\right) < R\right\}. \end{aligned} \tag{6.1}$$

The main source of simulation noise comes from the non-smooth dependence of the indicator functions $\chi\left\{Q_n\left(\omega_j\right) < R\right\}$ upon the simulated path ω. A small 'bump' to initial conditions will result in a small bump to ω, but that can result in a large change in the indicator $\chi\left\{Q_n\left(\omega_j\right) < R\right\}$. In essence, a whole coupon can be added or lost under a 'bumped' scenario compared to the base one, resulting in significant simulation noise when computing risk sensitivities.

The idea of the 'sausage' Monte Carlo is to replace 'point' estimates of the payoff v_j by their averages over thin 'sausages' centered around simulated paths ω_j. The state of a forward Libor model at time t is defined by $\overline{F}(t, \omega)$. We fix $\varepsilon > 0$, the width of the sausages. For each j, the ε-sausage in the state space is defined by

$$A_j^{\varepsilon} = \left\{\omega : \left\|\overline{F}\left(T_i, \omega\right) - \overline{F}\left(T_i, \omega_j\right)\right\| < \varepsilon \quad \forall i = 1, \ldots, N-1\right\}.$$

The sampling formula (6.1) is approximated with the following expression,

$$\tilde{v} \approx J^{-1} \sum_{j=1}^{J} \tilde{v}_j,$$

$$\tilde{v}_j = \mathbf{E}\left(\sum_{n=1}^{N-1} B_{T_{n+1}}^{-1}\left(\omega_j\right) \times X_n\left(T_n, \omega_j\right) \times \chi\left\{Q_n\left(\omega_j\right) < R\right\} \middle| A_j^{\varepsilon} \right).$$

Since $B_{T_{n+1}}^{-1}(\omega)$, $X_n(T_n, \omega)$ are generally smooth functions of the path ω, we evaluate them just at the sample path (this is accurate to order ε),

$$\tilde{v}_j = \sum_{n=1}^{N-1} B_{T_{n+1}}^{-1}\left(\omega_j\right) \times X_n\left(T_n, \omega_j\right) \times \mathbf{E}\left(\chi\left\{Q_n\left(\omega_j\right) < R\right\} \middle| A_j^{\varepsilon} \right).$$

Using approximate conditional independence and approximate uniformity of the process inside the sausage, the following formula can be obtained:

$$v_j = \sum_{n=1}^{N-1} B_{T_{n+1}}^{-1}\left(\omega_j\right) \times X_n\left(T_n, \omega_j\right) \times p_n\left(\omega_j\right),$$

$$p_n(\omega) = \min\left(\max\left(\frac{R - Q_n\left(\omega_j\right) + \eta_n}{2\eta_n}, 0 \right), 1 \right).$$

Here the exact dependence of η_ns on ε is not important as, in practice, η_ns can be set directly.

The formula replaces a discontinuous payoff $\chi\left\{Q_n\left(\omega_j\right) < R\right\}$ with a continuous one $p_n\left(\omega_j\right)$. Instead of a simple barrier breach/no breach indicator, we introduced a concept of a 'partial barrier breach'. If a barrier breached partially on the date T_n, only a portion of it knocks out. This introduces smooth dependence of the value on the simulation path. The bigger η_ns are, the 'more smooth' this dependence becomes, resulting in smoother risk sensitivities. They, however, cannot be made arbitrarily big as then the approximations used to compute

$$\mathbf{E}\left(\chi\left\{Q_n\left(\omega_j\right) < R\right\} \middle| A_j^{\varepsilon} \right)$$

break down, and the value $\tilde{v}$ becomes biased.

The smoothing methods presented above improve the quality of simulations, particularly for risk sensitivities, dramatically. If, however, speed and accuracy are still an issue, a more advanced approach is required.

7 Local projection method

At the risk of stating the obvious we note that the simpler the model, the better numerical methods are available. Low-dimensional Markovian models afford using PDE methods that have much better numerical properties than Monte Carlo methods available for high-dimensional models.

Even Monte Carlo simulations run faster for simpler models as they typically have fewer variables to evolve forward. As an approach to improve the numerical properties of the valuation algorithm, one can look into building a simpler model.

A powerful general approach for such a task is the 'local projection method'. Conceptually, it specifies finding a simple, 'local' model that is locally calibrated to a 'global' model (such as a forward Libor model) in such a way as to approximate the value of the global model for a particular trade. By local calibration we understand calibrating the simple model to the elements of the global model's volatility structure that have the biggest impact on the value of the instrument being valued (other, non-essential elements are ignored). With the critical elements of the volatility structure matched between the local and the global models, one would expect the values produced by both to closely match.

Identifying the relevant elements of the volatility structure is more of an art than a science. The most successful example of this approach is to Bermuda swaptions (see Andreasen (2001)). It has been determined that a Bermuda swaption price depends essentially only on the following two elements. The first is the collection of volatilities of core, or co-terminal swaptions, and the second is the intertemporal correlations of co-terminal swap rates. A simple 1-factor PDE-based model, such as the Hull–White model, can be constructed and calibrated to these parameters. For calibration, swaption volatilities come directly from the swaption grid, and the correlations are computed from a fully calibrated forward Libor model.

Let us apply this approach to TARNs. To begin, we rewrite the TARN value as follows,

$$v = \mathbf{E}_0 \left(\sum_{n=1}^{N-1} B_{T_{n+1}}^{-1} \delta_n \left((s - 2F_n(T_n))^+ - F_n(T_n) \right) \chi \left\{ \sum_{i=1}^{n-1} \delta_i (s - 2F_i(T_i))^+ < R \right\} \right).$$

Scrutinizing the payoff under the expected value operator, we see that it only depends on the following values

$$\tilde{F} = \{F_1(T_1), F_2(T_2), \ldots, F_{N-1}(T_{N-1})\}$$

(this is trivially true for everything but the numeraire B; for the numeraire we only need to recall the formula (3.2) to realize that this is true for it as well). In particular, *only the values of Libor rates on their fixing dates enter the payoff*. Their values at intermediate times are irrelevant. Thus, only the distributional properties of the $(N-1)$-dimensional vector $\tilde{F}$ are relevant. This is in contrast to a typical contract (such as a Bermuda swaption) that would depend on values of Libor rates at various dates prior to their fixings, with a typical dimension of $(N-1)(N-2)/2$.

From this relatively simple observation, powerful conclusions can be made. Such significant reduction of dimensionality indicates that a simpler model can indeed be used. Focusing on the covariance characteristics only (we will deal with volatility smiles later), and assuming lognormal distributions throughout, it is clear that if two models agree on

1. Term volatilities of Libor rates $\{\text{stdev}(\log F_n(T_n)), n = 1, \ldots, N-1\}$;
2. Intertemporal correlations of Libor rates $\{\text{corr}(\log F_n(T_n), \log F_m(T_m)), n, m = 1, \ldots, N-1\}$,

then the prices of a TARN computed by the two models will closely match.

An application of the local projection method to a TARN proceeds as follows.

- A forward Libor model is calibrated to the full swaption volatility grid and to the modeler's beliefs on the forward evolution of the volatility structure. This is the global model;
- Relevant term volatilities and intertemporal correlations are computed from the global model;
- A simpler model is constructed and calibrated to volatilities/correlations above. This is the local model;
- The local model is used for valuation. When computing risk sensitivities, the calibration info from the global model is recomputed for each bumped scenario, and the local model is recalibrated.

The local model needs enough flexibility in its volatility structure specification to calibrate to the set of volatility information required. Fortunately, the set is not very extensive. Even a simple model such as the Hull–White model has enough degrees of freedom to match the covariance information identified above. We provide more calibration details below, where a more realistic model (an extension of the Hull–White model) is developed.

It is important to emphasize that the local projection method is *not the same* as just using a simple model to value an exotic (something we warned against in the beginning of the chapter). The forward Libor model is an integral part of the method, and is used to extract unobservable but critical volatility information (such as the intertemporal correlations) from the observable market inputs.

A question closely linked to low-dimensional approximations is that of the number of factors required to value various contracts 'properly'. While some instruments may appear to require multi-factor models, this is usually an illusion. A properly calibrated one-factor model is usually more than adequate to price most deal types. This has been convincingly demonstrated for Bermuda swaptions in Andersen and Andreasen (2001). Other types of exotics were discussed in Andreasen (2004). In the latter, it has been noted that as long as the exotic depends on a single rate at each point in time, a one-factor model (properly calibrated to a global forward Libor model) is sufficient. It is only contracts that are linked to different rates observed at the same time (such as CMS spread linked deals) which require two or more factors.

While proper volatility structure modeling is always the main concern for exotics, the effects of the volatility smile should never be ignored.

8 Volatility smile effects

The payoff of the TARN on the date T_1, viewed as a function of the Libor rate fixing $F_1(T_1)$, has a number of important features. There is a digital-type discontinuity at $F_1(T_1) = S_1$, there is a call-option type discontinuity at $s/2$, and the payoff is non-linear for $F_1(T_1) > S_1$. An example is given in Figure 2. This observation serves to demonstrate that a model for a TARN needs to recover the whole distribution of $F_1(T_1)$ (as implied from caplet prices across a range of strikes), and not just some summary statistic such as an implied volatility at a certain strike.

One can argue that the main source of risk at time T_1 in the strike dimension is concentrated at the knockout strike $F_1(T_1) = S_1$. Thus, the argument goes, to value a TARN properly, it is

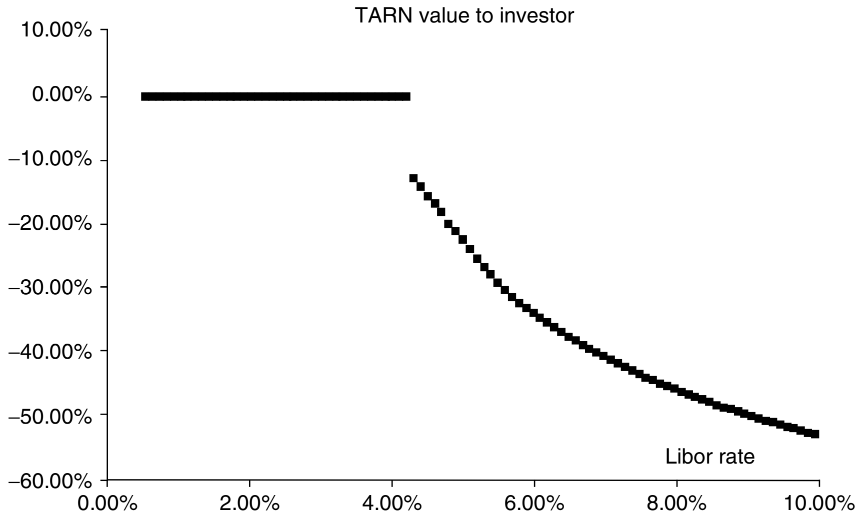


Figure 2: Example of a value of a TARN to investor on the first knockout date as a function of the Libor rate fixing on that date, excluding coupons not contingent on knockout

enough to choose a model that values a digital option with strike S_1 consistently with the market, without trying to match the whole volatility smile. While this argument has some merit for the first date T_1, it breaks down for the subsequent knockout dates. It is quite clear that the value of $F_2(T_2)$ for which the deal knocks out depends on the fixing of $F_1(T_1)$, something that is not known at time $t = 0$. Thus, a model that only matches the implied volatility of $F_2(T_2)$ at a single strike, or matches the slope of the smile at a certain strike, is going to be inadequate. The same holds for subsequent Libor rates.

From this and previous sections, it is clear that a successful candidate for the local model of the local projection method should be a model that has

- A low number of factors;
- Enough flexibility to calibrate to term volatilities and intertemporal correlations of Libor rates;
- The ability to recover term volatility smiles of all Libor rates.

Of all the mechanisms available for generating smiles in interest rate models, stochastic volatility appears to be the most practical choice. We have commented before that the Hull–White model is flexible enough to satisfy the first two requirements. There exists a model that combines the Hull–White model with stochastic volatility. It is called the Stochastic Volatility Cheyette model (SV–Cheyette), see Andersen and Andreasen (2002).

The SV–Cheyette model is defined by the following system of equations.

$$dx(t) = (-\upsilon(t)x(t) + y(t))\,dt + \sqrt{V(t)}\eta(t, x(t), y(t))\,dW(t), \quad x(0) = 0,$$
$$dy(t) = \left(V(t)\eta^2(t, x(t), y(t)) - 2\upsilon(t)y(t)\right)dt, \quad y(0) = 0,$$

$$dV(t) = \varkappa(\theta - V(t))\, dt + \varepsilon\psi(V(t))\, dZ(t), \quad V(0) = 1,$$
$$\langle W, Z \rangle = 0.$$

Here $x(\cdot)$ is the short rate state process, and is the primary driver of the interest rate curve. The short rate is given by $(f(s,t)$ and is the instantaneous forward rate at s for t,

$$r(t) = f(t,t) = f(0,t) + x(t).$$

The other state variable $y(\cdot)$ is an auxiliary variable, and $V(\cdot)$ is the stochastic variance process. The volatility term of the process for dx has a stochastic volatility component $\sqrt{V(t)}$ and a local volatility component $\eta(t,x,y)$, giving it enough flexibility to match a wide variety of volatility smile shapes. Note that the Hull–White model is obtained by setting $\eta(t, x(t), y(t)) \equiv \eta(t)$, $\varepsilon = 0$.

Zero-coupon bonds (and, consequently, all market rates) are functions of x, y. The bond reconstruction formula is

$$P(t,T) = \frac{Pt(0,T)}{P(0,t)} \exp\left(-G(t,T)\,x(t) - \frac{1}{2}G^2(t,T)\,y(t)\right),$$
$$G(t,T) = \int_t^T e^{-\int_t^u v(d)ds} du.$$

The calibration of the SV–Cheyette model for a TARN can be decoupled into three distinct and consequential steps: (a) smile calibration; (b) correlation calibration; and (c) volatility calibration. We review the steps in turn.

Volatility smile generated by the SV–Cheyette model for a particular time horizon is controlled mostly by the volatility of variance parameter ε and the form of the local volatility function η. The relationships between volatility smiles at different expiries are controlled by the speed of mean reversion of variance parameter $\varkappa$. These can be chosen to match volatility smiles of Libor rates $F_n(T_n)$.

The intertemporal correlations of Libor rates $\{\mathrm{corr}t(\log F_n(T_n),\ \log F_m(T_m)), n, m = 1, \ldots, N-1\}$ are controlled by the mean reversion of rates function $v(t)$. In particular, since the tenors of Libor rates are short, the correlations of Libor rates can be approximated by correlations of the short rate state process $x(\cdot)$. The following formulas can be used for calibrating intertemporal correlations,

$$\mathrm{corr}(\log F_n(T_n), \log F_m(T_m)) \approx \mathrm{corr}(x(T_n), x(T_m))$$
$$\approx \left(\frac{\int_0^{\min(T_n,T_m)} e^{2\int_0^u v(s)ds} du}{\int_0^{\max(T_n,T_m)} e^{2\int_0^u v(s)ds} du} \right)^{1/2}.$$

Once the SV parameters and the mean reversion are fixed, the term volatilities of Libor rates (caplet volatilities) in the SV–Cheyette model are determined by the (time dependent) overall level of the function η, i.e. by the function $\eta_0(t) \triangleq \eta(t,0,0)$. In fact, the function $\eta_0(t)$ can be very efficiently bootstrapped from the strip caplet volatilities, as explained in Andersen and Andreasen (2002).

In a sense, the SV–Cheyette model is an 'ideal' choice of a local model for TARNs as it has just enough (and not more) flexibility to calibrate to all relevant covariance/smile information.

Having advocated using a volatility-smile enabled local model, it is only reasonable to have the same level of sophistication in the global model. While originally we presented a skew-enhanced forward Libor model (see (3.3)), in fact a better choice is a proper stochastic volatility forward Libor model. A good choice would be the model described in Andersen and Brotherton-Ratcliffe (2001). If volatility smiles of different Libor rates are substantially different, then a better choice might be the model developed in Piterbarg (2004).

Having explained how to calibrate a low-dimensional local model for a TARN, we now proceed to discuss numerical methods that can be used for such a model.

9 TARNs by PDEs

The SV–Cheyette model admits a PDE-based valuation scheme in 'two-and-a-half' factors, with two 'full' factors x and V and one 'half' factor y (the process $y(\cdot)$ is locally deterministic), see Andersen and Andreasen (2002). However, TARNs are path-dependent contracts. It appears that we are forced to use a Monte Carlo method even with the SV–Cheyette model. While Monte Carlo for the SV–Cheyette model is typically orders of magnitude faster than for a forward Libor model, it would still be of great benefit if we could use a PDE-based valuation.

It turns out that the nature of path dependence in TARNs is such that it can indeed be handled in backward-induction schemes. A general approach is well covered in the book by Wilmott (2000). We quickly review it here, with a focus on TARN valuation.

The path dependency of a TARN is concentrated in the quantity Q_n, the total accumulated return to date. The idea of the method is to introduce an auxiliary state variable to keep track of it. This variable stays constant between fixing dates and is updated on each fixing date by an amount (a coupon paid) known on that date.

For simplicity, we describe the method in the context of a generic one-factor model. Suppose the model is given by the following SDE on the short rate,

$$dr(t) = a(t, r(t))\, dt + b(t, r(t))\, dW(t).$$

Zero-coupon bonds and Libor rates are functions of r, and we use self-evident notations $P(r, t, T)$, $F_n(r, t)$, and so on.

Let $V(t, r, z)$ be the value of the TARN at time t, for the short rate r, assuming that the total accumulated coupon at time t is z. For each particular z, the function V satisfies the following SDE,

$$V_t(t, r, z) + a(t, r)\, V_r(t, r, z) + \frac{b^2(t, r)}{2} V_{rr}(t, r, z) = rV(t, r, z). \tag{9.1}$$

The terminal condition is given at time T_N by

$$V(T_N-, r, z) = 0. \tag{9.2}$$

The following boundary/continuity conditions should be enforced at times T_n,

$$\overline{V}(T_n, r, z) = V\left(T_n, r, z + \delta_n \cdot (s - 2F_n(r, T_n))^+\right), \tag{9.3}$$

$$V(T_n-, r, z) = \overline{V}(T_n, r, z) + \delta_n P(r, T_n, T_{n+1})\left((s - 2F_n(r, T_n))^+ - F_n(r, T_n)\right). \tag{9.4}$$

The scheme works as follows. The function V is initialized with the terminal values (9.2). Then, for each $n = N - 1, \ldots, 0$, the following is done,

1. For each value of z, the PDE (9.1) is solved backwards on $[T_n, T_{n+1})$ with the terminal condition $V(T_{n+1}-, r, z)$;
2. The continuity condition (9.4) is applied 'across z-slices', corresponding to the update of the total return Q_n;
3. The boundary condition (9.3), i.e. the payment of the coupon at time T_n, is applied;
4. The last step gives the new terminal condition, and the steps are repeated with $n = n - 1$.

The final value is computed by

$$v = V\left(0, r^*, 0\right),$$

where r^* is today's value of the short rate.

The variable z (just like t and r) is typically discretized. The scheme amounts to solving PDEs (9.1) independently for each discretized value of z, with the linkage between different 'z-slices' given by (9.3).

10 Conclusions

The local projection method we develop combines a forward Libor model and the Stochastic Volatility Cheyette model. The method provides a robust risk management framework for TARNs. Efficient numerical methods of the SV−Cheyette model are combined with calibration advantages of a forward Libor model. While the SV−Cheyette model is used for routine valuation and risk reporting, periodic benchmarking against the forward Libor model can be performed by using variance reduction techniques presented in the chapter.

Appendix A. Importance sampling for TARNs

We observe that due to non-negativity of $C_i(\cdot)$, the following equalities hold **P**-a.s.

$$\begin{aligned}\{Q_n < R\} &\Longleftrightarrow \left\{Q_{n-1} < R, (s - 2F_{n-1}(T_{n-1}))^+ < (R - Q_{n-1})/\delta_{n-1}\right\}\\ &\Longleftrightarrow \{Q_{n-1} < R, F_{n-1}(T_{n-1}) > S_{n-1}\}\end{aligned}$$

and

$$\begin{aligned}\{Q_n < R\} &\Longleftrightarrow \{Q_1 < R, Q_2 < R, \ldots, Q_n < R\}\\ &\Longleftrightarrow \{F_1(T_1) > S_1, \ldots, F_{n-1}(T_{n-1}) > S_{n-1}\}.\end{aligned}$$

Hence

$$\chi\left\{Q_n < R\right\} = \prod_{k=1}^{n-1} \chi\left\{F_k\left(T_k\right) > S_k\right\}.$$

Define

$$\Lambda_n(t) = \begin{cases} \dfrac{\mathbf{P}_t\left(F_{n+1}\left(T_{n+1}\right) > S_{n+1}\right)}{\mathbf{P}_{T_n}\left(F_{n+1}\left(T_{n+1}\right) > S_{n+1}\right)}, & t \in [T_n, T_{n+1}), \\ \dfrac{\chi\left\{F_{n+1}\left(T_{n+1}\right) > S_{n+1}\right\}}{\mathbf{P}_{T_n}\left(F_{n+1}\left(T_{n+1}\right) > S_{n+1}\right)}, & t \geq T_{n+1}, \\ 1, & t < T_n. \end{cases}$$

We note that $\Lambda_n(t)$ is a non-negative $\mathbf{P}$-martingale. Moreover, $\Lambda_n(t)$ is constant on $[0, T_n]$ and $[T_{n+1}, \infty)$. In addition, $\Lambda_n(t)$ is $\mathcal{F}(T_{n+1})$-measurable for $t \geq T_{n+1}$.

Define $\Lambda(t)$ by

$$\Lambda(t) = \prod_{n=0}^{N-2} \Lambda_n(t).$$

It is not hard to show that

$$\{\Lambda(t), t \in [0, \infty)\}$$

is a martingale as well. Denote the value of the nth coupon, contingent on survival, by

$$x_n = \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \chi\left\{Q_n < R\right\}\right).$$

Then

$$\begin{aligned} x_n &= \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \chi\left\{Q_n < R\right\}\right) \\ &= \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \prod_{k=1}^{n-1} \chi\left\{F_k\left(T_k\right) > S_k\right\}\right) \\ &= \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \prod_{k=1}^{n-1} \frac{\chi\left\{F_k\left(T_k\right) > S_k\right\}}{\mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)} \times \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)\right) \\ &= \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \prod_{k=1}^{n-1} \Lambda_{k-1}\left(T_n\right) \times \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)\right) \\ &= \mathbf{E}_0\left(B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \prod_{k=1}^{N-1} \Lambda_{k-1}\left(T_n\right) \times \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)\right) \\ &= \mathbf{E}_0\left(\Lambda\left(T_n\right) \times B_{T_{n+1}}^{-1} \times X_n\left(T_n\right) \times \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}}\left(F_k\left(T_k\right) > S_k\right)\right). \end{aligned}$$

Define a new measure $\tilde{\mathbf{P}}$ by its Radon–Nikodym derivative with respect to $\mathbf{P}$,

$$\left.\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\right|_{\mathcal{F}(t)} = \Lambda(t).$$

Then

$$x_n = \tilde{\mathbf{E}}_0 \left(B_{T_{n+1}}^{-1} \times X_n(T_n) \times \psi_n \right),$$

$$\psi_n = \prod_{k=1}^{n-1} \mathbf{P}_{T_{k-1}} \left(F_k(T_k) > S_k \right).$$

Strictly speaking, the measure $\tilde{\mathbf{P}}$ is not equivalent to $\mathbf{P}$ because $\Lambda(t)$ can be zero. The important observation, however, is that the value of the TARN is zero for those paths for which $\Lambda(t)$ is zero. Thus, $\tilde{\mathbf{P}}$ and $\mathbf{P}$ are equivalent on the 'relevant' subspace of the sample space Ω. We omit technical details.

To simulate forward Libor rates under the measure $\tilde{\mathbf{P}}$, two approaches can be used. The first one, suggested in Glasserman and Staum (2001), is based on the following idea. If a standard Gaussian random variable ξ is simulated by the formula

$$\xi = \Phi^{-1}(U),$$

where U is a uniform (on [0, 1]) draw and $\Phi(\cdot)$ is the standard Gaussian CDF, then ξ conditioned on the event $\{\xi > b\}$ can be simulated by the formula

$$\tilde{\xi} = \Phi^{-1}\left(\Phi(b) + (1 - \Phi(b))\tilde{U}\right),$$

where $\tilde{U}$ is a uniform draw on $[0, 1]$. To apply this idea in the context of the (multi-dimensional) forward Libor model, we need to rotate the covariance structure of forward Libor rates such that the forward Libor rate relevant for the next knockout is simulated by a single Gaussian draw, and apply the above conditioning to that Gaussian variable.

Another approach is based on treating the change of measure from $\mathbf{P}$ to $\tilde{\mathbf{P}}$ as one induced by a change of drift of the underlying Brownian motion. Let us compute the drift of dW, the Brownian motion that drives the forward Libor model (3.3), under $\tilde{\mathbf{P}}$. The state of the model at time t is completely determined by the vector $\overline{F}(t)$ of all forward Libor rates observed at t. Thus, for every n, the probability $\mathbf{P}_t(F_n(T_n) > S_n)$ can be seen as a deterministic function of $\overline{F}(t)$, and we define a deterministic function $\Psi_n(t, \overline{x})$ by

$$\Psi_n(t, \overline{x}) = \mathbf{E}\left(\chi\{F_n(T_n) > S_n\} \middle| \overline{F}(t) = \overline{x}\right).$$

Clearly, for $t \in [T_n, T_{n+1})$,

$$d\Lambda(t)/\Lambda(t) = d\Psi_{n+1}\left(t, \overline{F}(t)\right)/\Psi_{n+1}\left(t, \overline{F}(t)\right).$$

By Ito's lemma (discarding dt terms because $\Lambda(t)$ is a martingale), under $\mathbf{P}$,

$$d\Psi_{n+1}\left(t,\overline{F}(t)\right)/\Psi_{n+1}\left(t,\overline{F}(t)\right) = \frac{1}{\Psi_{n+1}\left(t,\overline{F}(t)\right)}\sum_{j=1}^{N-1}\frac{\partial\Psi_{n+1}\left(t,\overline{F}(t)\right)}{\partial x_j}dF_j(t)$$
$$= \gamma_{n+1}\left(t,\overline{F}(t)\right)dW(t),$$

where

$$\gamma_{n+1}(t,\overline{x}) = \sum_{j=1}^{N-1}\lambda_j(t)\,\phi\left(x_j\right)\frac{\partial}{\partial x_j}\log\Psi_{n+1}(t,\overline{x}).$$

Define

$$\gamma(t,\overline{x}) = \sum_{n=0}^{N-2}\chi\left\{T_n \le t < T_{n+1}\right\}\times\gamma_{n+1}(t,\overline{x}),$$

then

$$d\Lambda(t)/\Lambda(t) = \gamma\left(t,\overline{F}(t)\right)dW(t), \quad t \ge 0.$$

By Girsanov's theorem,

$$d\tilde{W}(t) = dW(t) - \gamma\left(t,\overline{F}(t)\right)dt$$

is a driftless Brownian motion under $\tilde{\mathbf{P}}$. The forward Libor model can be simulated under the measure $\tilde{\mathbf{P}}$ using the following SDE (compare to (3.3))

$$\begin{aligned}dF_n(t) &= \lambda_n(t)\,\phi(F_n(t))\left(\gamma\left(t,\overline{F}(t)\right)+\sum_{j=1}^{n}1_{\{t<T_j\}}\frac{\delta_j\phi\left(F_j(t)\right)}{1+\delta_jF_j(t)}\lambda_j(t)\right)dt\\&\quad+\lambda_n(t)\,\phi(F_n(t))\,d\tilde{W}(t),\\&\quad n=1,\ldots,N-1.\end{aligned} \tag{A.1}$$

FOOTNOTES & REFERENCES

1. So called 'snowballs' are the most significant recent development in callable Libor exotics. In a snowball, each coupon is a function of interest rates *and the previous coupon*. Such path dependence can be easily handled in a Monte Carlo-based forward Libor model, our model of choice for callable Libor exotics. All the methods developed for 'standard' callable Libor exotics in Piterbarg (2003) can be easily extended to snowballs.

2. A TARN can be based on any type of a structured coupon. Historically the inverse floating coupon was the first one to be used. Our analysis is not specific to a type of coupon, and we use the inverse floating one for concreteness.
3. Sometimes a TARN is structured so that the first digital is virtually worthless, but the second one is important. The discussion that follows should be modified accordingly.

■ Andersen, L. B. G. and Andreasen, J. (2000) Volatility skews and extensions of the Libor Market Model. *Applied Mathematical Finance*, 7, 1–32, March.
■ Andersen, L. B. G. and Andreasen, J. (2001) Factor dependence of Bermudan swaption prices: fact or fiction? *Journal of Financial Economics*, 62, 3–37.
■ Andersen, L. B. G. and Andreasen, J. (2002) Volatile volatilities. *Risk*, 15(12), December.
■ Andersen, L. B. G. and Brotherton-Ratcliffe, R. (2001) Extended Libor market models with stochastic volatility. Working paper.
■ Andreasen, J.(2001) The pricing of Bermuda swaptions. Risk conference, Paris.
■ Andreasen, J. (2004) Markov yield curve models for exotic interest rate products. Lecture notes.
■ Glasserman, P. (2003) *Monte Carlo Methods in Financial Engineering*. Springer.
■ Glasserman, P. and Staum, J. (2001) Conditioning on one-step survival in barrier option simulations. *Operations Research*, 49, 923–937.
■ Glasserman, P. and Zhao, X. (1999) Fast Greeks in forward Libor models. *Journal of Computational Finance*, 3, 5–39.
■ Piterbarg, V. V. (2003) A practitioner's guide to pricing and hedging callable LIBOR exotics in forward LIBOR models. SSRN Working paper.
■ Piterbarg, V. V. (2004a) Computing deltas of callable Libor exotics in forward Libor models. *Journal of Computational Finance*, 7(3), 107–144.
■ Piterbarg, V. V. (2004b) A stochastic volatility forward Libor model with a term structure of volatility smiles. SSRN Working paper.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance*. John Wiley & Sons Ltd.

15
Fast Valuation of a Portfolio of Barrier Options under the Merton's Jump Diffusion Hypothesis

Antony Penaud*

We want to price a large portfolio of barrier options when the underlying follows Merton's jump diffusion process. We do so by solving—for each barrier—the appropriate Fokker Planck equation for the risk neutral probability density function.

1 Introduction

There are nice semi-analytic formulas for the price of European options when the underlying follows Merton's jump diffusion model (see Merton 1976). There are also nice formulas for barrier options under the geometric Brownian motion hypothesis (see Haug 1997). However, for barrier options under Merton's model no simple pricing formula is available. A natural method for pricing a barrier option under Merton's model would be to solve the partial integro differential equation (PIDE) with appropriate final and boundary conditions.

Let's assume that we want to price a portfolio of many thousand barrier options and that the underlying is the same for all deals.

Solving many thousand PIDEs (one for each deal) would be far too long and we need to turn to another approach.

---

*I would like to thank Yanmin Li and James Selfe for helpful suggestions.

We propose a pricing methodology adapted to the problem. Indeed we aim for the method which would best price all the deals together.[1] Our idea is to find—for each barrier—the joint risk neutral probability density function (pdf) as a function of time.[2] Once we know the pdfs we can price knock out options[3] by integrating the payoffs.

The probability density function is found by solving the Fokker Planck equation (a.k.a. forward Kolmogorov equation) that corresponds to the jump diffusion process. This PIDE is solved by an adapted Crank Nicolson scheme (Crank Nicolson for the diffusion and explicit for the jump part).

Our approach would not be the best if we wanted to price a few deals only. Indeed, solving the forward PIDE is not as nice as solving the backward pde (integration of the risk neutral pdf is required and the initial condition is a delta function) and it would not be worth it. But for non Monte Carlo approaches computation time is—as far as we are aware—proportional to the number of deals. Whereas for our pricing methodology computation time is proportional to the number of barriers. So for a large portfolio of options our method becomes better.[4]

First we are going to review Merton's jump diffusion model. Then we mention the natural way for pricing a barrier option under Merton's model. After that we explain our method for pricing the portfolio of barrier options: we write the equations and go through some implementation issues.

Merton (1976) introduced the jump diffusion model and the pricing framework. Andersen and Andreasen (2000) and Lipton (2002) have looked at forward PIDEs. Andersen and Andreasen derive the forward PIDE for the evolution of European call prices as a function of strike and maturity (generalised Dupire equation). They solve it via an alternative direction implicit (ADI) finite different scheme combined with fast Fourier transform (FFT) methods. Lipton develops a new approach for the pricing of path-dependent options on assets driven by jump diffusions with log-exponential Poissonian jumps. His approach is based on fluctuation identities for Levy processes. Metwally and Atiya (2002) use Brownian bridge methods for simulating the jump diffusion process and price barrier options.

2 Merton's jump diffusion model

The underlying S follows the random walk

$$\frac{dS}{S} = (r - \lambda k)\,dt + \sigma dz + dq \tag{1}$$

where r is the risk neutral drift, σ is the volatility, dz is a Gaussian random variable with mean 0 and variance dt and

$$dq = \begin{cases} Y - 1 & \text{if jumps occurs} \\ 0 & \text{otherwise.} \end{cases}$$

During the time interval dt the probability of q jumping is λdt.

If there is a jump, the jump part of the change in the underlying is $dS = (Y - 1)S$, i.e. S goes to YS.

The expectation of $Y - 1$ is called k.

So the expected change in S (from the jump component) is $S\lambda k\,dt$.

If we want r to be the instantaneous total expected rate of return, then we need to subtract λkdt in the diffusion part of the equation, and $E[dS/S] = r\,dt$.

Assuming independence between random variables (dz and dq) the distribution for $S(t)$ is

$$S(t) = e^{(r-\frac{\sigma^2}{2}-\lambda k)t+z(t)}y(n). \tag{2}$$

Above, $z(t)$ is $N(0, \sigma^2 t)$ and $y(0) = 1$, $y(n) = \Pi_1^n Y_i$ with Y_i's iid jumps. Note that n is a random variable too.

2.1 Lognormal jumps

We assume jumps to be lognormal, i.e. we choose Merton's jump diffusion model. In this section we give a summary of European option pricing under Merton's model.

Assume $E[lnY] = \gamma - \frac{\delta^2}{2}$ and var$[lnY] = \delta^2$, then $E[Y] = e^{\gamma} = 1 + k$ and so $\gamma = \ln(1 + k)$.

For n fixed, $y(n)$ is the product of n lognormal random variables so it is lognormal, and therefore $S(t)$ is lognormal too. So any European option can be priced by doing the sum of the Black–Scholes prices weighted by the probability of the corresponding number of jumps occurring.

Quantitatively, the distribution for $S(t)$ (in case of n jumps) is

$$S(t) = S(0)e^{(r_n t-\frac{\sigma^2 t}{2}-\frac{n\delta^2}{2})+\nu_n\phi\sqrt{t}} \tag{3}$$

where $\nu_n^2 = \sigma^2 + n\delta^2/t$ and $r_n = r - \lambda k + n\gamma/t$.

So the European price is

$$\text{European price} = \sum_0^\infty P_n D_n BS(S, T, E, \nu_n, r_n) \tag{4}$$

with $P_n = \frac{e^{-\lambda T}(\lambda T)^n}{n!}$ the probability of exactly n jumps occurring and $D_n = e^{-\lambda kT}(1 + k)^n$ the correcting term for the discounting term (without this term the discounting term would be $e^{-r_n T}$ but it should be e^{-rT}). This can finally simplify to

$$\text{European price} = \sum_0^\infty \frac{e^{-\lambda' T}(\lambda' T)^n}{n!} BS(S, T, E, \nu_n, r_n) \tag{5}$$

with $\lambda' = \lambda(1 + k)$.

2.2 Price as a solution of a PIDE

Using standard arguments, the pricing equation for a European option under Merton's model is the following PIDE:

$$V_t + \frac{\sigma^2 S^2}{2} V_{SS} + \lambda E[V(t, JS) - V(t, S)] + (r - \lambda k)SV_S - rV = 0 \tag{6}$$

together with $V(T, S) = \text{payoff}(S)$ and $V_{SS}(t, S) = 0$ on both sides.

3 Pricing a barrier option

3.1 Geometric Brownian motion

Analytic formulas are available (see Haug 1997). In the PDE framework the barrier option price can be found by solving the following backward equation (see Wilmott 2000):

$$V_t + \frac{\sigma^2 S^2}{2} V_{SS} + rSV_S - rV = 0 \tag{7}$$

together with $V(T, S) = \text{payoff}(S)$, $V(t, B) = 0$ if it is a knockout barrier, and $V_{SS}(t, S) = 0$ on the other side.

3.2 Merton's jump diffusion model

In this case it is the PIDE that needs to be solved backward

$$V_t + \frac{\sigma^2 S^2}{2} V_{SS} + \lambda E[V(t, JS) - V(t, S)] + (r - \lambda k)SV_S - rV = 0 \tag{8}$$

together with $V(T, S) = \text{payoff}(S)$, $V(t, B) = 0$ if it is a knockout barrier, and $V_{SS}(t, S) = 0$ on the other side.

4 Pricing the portfolio of barrier options

4.1 Model

There is no easy way to price a barrier option under the jump diffusion hypothesis.

In this chapter we want to price a portfolio of many thousand barrier options. The volatility[5] and the discount curve are the same for all deals. We obviously need to take this into account when it comes to choosing the methodology. This is why we choose a method that is going to solve a lot of deals at the same time as opposed to a method that takes care of each deal independently.

For each barrier we are going to find the joint risk neutral probability density function[6] for all maturities by solving the appropriate forward Kolmogorov equation. When we know the risk neutral density function it is straightforward to get the price of the knockout barrier option as it is the integral of the payoff times of the pdf. And as there are only a few barriers, only a few forward Kolmogorov equations need to be solved for each underlying. Computation time is not proportional to the number of deals. It is proportional to the number of barriers.[7]

4.2 Mathematical formulation

Let's write down the PIDE satisfied by the risk neutral density function. Let $x = \log(S/S_0)$, then the jump diffusion process is

$$dx = \left(r(t) - \frac{\sigma(t)^2}{2} - \lambda k\right)dt + \sigma(t)\,dX + \log Y dq. \tag{9}$$

For a down option,[8] the corresponding forward Kolmogorov equation for the probability density function is

$$f_t = \frac{\sigma^2(t)}{2} f_{xx} - \left(r(t) - \frac{\sigma^2(t)}{2} - \lambda k\right) f_x + \lambda \left(\int_{\log \frac{B}{S_0}}^{\infty} f(y) \frac{e^{-\frac{(x-y-\gamma+0.5\delta^2)^2}{2\delta^2}}}{\delta\sqrt{2\pi}} dy - f \right) \tag{10}$$

together with initial condition

$$f(0, x) = \delta(x) \tag{11}$$

and boundary condition for a knockout barrier[9]

$$f\left(t, \log \frac{B}{S_0}\right) = 0. \tag{12}$$

Let's try to intuitively explain the meaning of the last term in the Kolmogorov equation. With probability λdt there is a jump. In that case, at x, the variation w.r.t. time of the pdf is equal to the difference between the 'particles' coming from elsewhere to x (the integral term) and the 'particles' leaving x (the term $f(x)$).

Now, the price of the down-and-out barrier ($B < S_0$) maturing at T and with payoff $g(S)$ is

$$\text{Price} = S_0 \text{DF}(T) \int_{\log \frac{B}{S_0}}^{\infty} f(T, x) g(e^x)\, dx. \tag{13}$$

The price of the up-and-out barrier ($B > S_0$) maturing at T and with payoff $g(S)$ is

$$\text{Price} = S_0 \text{DF}(T) \int_{-\infty}^{\log \frac{B}{S_0}} f(T, x) g(e^x)\, dx. \tag{14}$$

In the above equations $\text{DF}(T)$ is the domestic discount factor at time T.

Remark How would we price a down-and-out option if a rebate $b(t)$ is paid when the barrier is knocked out? We would find the probability $p(t)\, dt$ that the barrier is knocked out in the interval $(t, t + dt)$ by

$$p(t) = -\frac{d}{dt} \int_{\log \frac{B}{S_0}}^{\infty} f(t, x)\, dx. \tag{15}$$

And the price of the down-and-out option with rebate would be

$$\text{Price}(b(t)) = \text{Price}(b \equiv 0) + \int_0^T \text{DF}(t) p(t) b(t)\, dt. \tag{16}$$

4.3 Implementation

We solve the PIDE via finite difference. We go for the Crank Nicolson scheme for the partial derivatives (see Wilmott 2000) and we approximate the integral term with a simple explicit approximation.

Integration of the pdf at maturity gives the price of the knockout barrier option. The price of the knockin barrier option is the European option price minus the knockout barrier option price.

Now, we make a few comments on some implementation issues.

The boundary conditions There are two boundary conditions: one on the barrier and one on the other side.

- On the other side of the barrier the node should be far away enough from the current spot price so that it is unlikely that the spot reaches it. The value of the pdf there is zero (second derivative equal to zero does not work well as there should not be loss or gain in the integral on that side).
- Now on the barrier the pdf value should be forced to be zero. This way particles that have already touched the barrier do not contribute anymore to the (jump) diffusion of the pdf. The programmer might not be able to put the barrier on a node; in this case he can use an extra node outside the barrier (so the barrier is between the first and the second node) and assume the pdf is linear in between the two nodes (see Wilmott 2000). The value of the pdf on the first node is then negative.

Initial condition The delta function is approximated by the following hat function:

$$f(t=0,x) = \begin{cases} \dfrac{1}{\delta x} & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

where δx is the space step.

Choosing the space step and the time step We want the time step to be equal to the space step. However, this size does not have to be the same on the whole grid. In fact it is best to start (i.e. for short maturities) with a fine grid and then switch to a grid in which nodes are further apart from each other. Shortly after time 0 (and for a longer time when volatility is small) the pdf looks like a peak. And one gains a lot in accuracy[10] by using a fine mesh. Moreover it does not cost too much computation time as the mesh does not need to be too wide (shortly after time 0 the spot is close to its original value). Eventually the pdf spreads out and one can switch to a grid which is both wider and less fine. The new space step (time step) could be a multiple of the old one so no interpolation is required when we switch from the fine space step grid to the larger one.

5 Conclusion

Standard pricing methodologies focus on the pricing of one exotic deal under one particular model. Here we have looked at the pricing of a large portfolio of exotic options. We have indeed focused on the pricing of many thousand barrier options under Merton's jump diffusion model.

Because of computation time the best method for pricing one deal is not necessarily the best for pricing such a large portfolio. Our method—while not the best for pricing a single deal—is very fast when it comes to pricing many deals. For each barrier we solve the forward Kolmogorov equation that gives us the joint risk neutral probability density function. So simple integration of that function gives the price of a knockout barrier option. Most of the computation time is taken by solving the PIDEs and computation time is proportional to the number of barriers.

FOOTNOTES & REFERENCES

1. We want to know the values of the individual deals. If we were interested in the value of the whole portfolio only we could solve — for each barrier — the backward PIDE and add the relevant payoffs at the relevant times to the portfolio's value.
2. The probability measured by the pdf is the probability for the asset to have a certain value at a certain time *and* that it has not touched the barrier before.
3. The price of a knockin option is the price of the European option minus the knockout option price.
4. We suppose that the number of different barriers is not large.
5. We assume that the volatility is a function of time and is piecewise constant.
6. We mentioned what we mean by joint risk neutral pdf in the introduction.
7. Provided that the longest maturity is about the same for each barrier.
8. If the barrier is above the underlying price then the integral is $\int_{-\infty}^{\log \frac{B}{S_0}}$ as opposed to $\int_{\log \frac{B}{S_0}}^{\infty}$.
9. The price of the knockin barrier is the price of the European option minus the price of the knockout option.
10. Both from the finite difference approximation perspective and the payoff integration perspective.

■ Andersen, L. and Andreasen, J. (2000) Jump-diffusion processes: volatility smile fitting and numerical methods for pricing. *Review of Derivatives Research* 4, 231–262.
■ Gatheral, J. (2003) Case Studies in Financial Modelling Course Notes, Courant Institute of Mathematical Sciences, Fall Term.
■ Haug, E. G. (1997) *The Complete Guide to Option Pricing Formulas*. McGraw Hill.
■ Lipton, A. (2002) Path-dependent options on assets with jumps and a new approach to credit default spreads. 5th Columbia-Jaffe Conference, New York, April 5th.
http://www.math.columbia.edu/lrb/columbia2002.pdf
■ Merton, R. C. (1976) Option pricing when underlying returns are discontinuous. *Journal of Financial Economics*, 3 (January–March): 125–44.
■ Metwally, S. and Atiya, A. (2002) Using Brownian bridge for fast simulation of jump-diffusion processes and barrier options. *The Journal of Derivatives*, Fall, vol. 10, number 1, 43–54.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance* John Wiley & Sons Ltd.

16
An Analysis of Pricing Methods for Basket Options

Martin Krekel, Johan de Kock, Ralf Korn and Tin-Kwai Man

This chapter deals with the task of pricing basket options. Here, the main problem is not path dependency but the multi-dimensionality which makes it impossible to give exact analytical representations of the option price. We review the literature and compare six different methods in a systematic way. Thereby we also look at the influence of various parameters such as strike, correlation, forwards or volatilities on the performance of the different approximations.

1 Introduction

While with many exotic options it is even harder to fully understand the way their final payoff is built up, the construction of the payoff of a (European) basket option is very simple. We define the price of a basket of stocks by

$$B(T) = \sum_{i=1}^{n} w_i S_i(T),$$

i.e. it is the weighted average of the prices of n stocks at maturity T. Here the weights w_i are usually assumed to be positive and to sum up to 1, but also to be quite arbitrary.

Our task is to determine the price of a call ($\theta = 1$) or a put ($\theta = -1$) with strike K and maturity T on the basket, i.e. to value the payoff

$$P_{Basket}(B(T), K, \theta) = [\theta(B(T) - K)]^{+}.$$

Contact address: Fraunhofer ITWM, Department of Financial Mathematics, 67653 Kaiserslautern, Germany
E-mail: krekel@itwm.fhg.de

We price these options with the Black–Scholes model. Note that by the form of the payoff it is not necessary to distinguish between the trading date and the valuation date to calculate the values of these options, since they are not path dependent. Hence without loss of generality we can set $t = 0$ and denote the remaining time to maturity with T. In order to ease the calculations we use the so-called *forward notation*. The T-forward price of stock i is given by

$$F_i^T = S_i(0)\exp\left(\int_0^T \left(r(s) - d_i(s)\right) ds\right)$$

where $r(.)$ and $d_i(.)$ are deterministic interest rates and dividend yields. With its help the stock prices can be represented as

$$S_i(T) = F_i^T \exp\left(-\int_0^T \frac{1}{2}\sigma_i^2 ds + \int_0^T \sigma_i dW_i(s)\right)$$

where the $W_i(.)$ are correlated one-dimensional Brownian motions with correlation of ρ_{ij}. Further, we define the discount factor as

$$\mathrm{DF}(T) = \exp\left(-\int_0^T r(s)ds\right).$$

The forward-oriented notation has two advantages: firstly, contrary to short rates and dividend yields, forward prices and discount factors are market quotes. Secondly, from a computational point of view, it is less costly to work with single numbers, i.e. the forward prices and the discount factor, instead of several term structures, namely the short rates and the dividend yields.

The problem of pricing the above basket options in the Black–Scholes model is the following: the stock prices are modelled by geometric Brownian motions and are therefore log-normally distributed. As the sum of log-normally distributed random variables is not log-normal, it is not possible to derive an (exact) closed-form representation of the basket call and put prices. Due to the fact that we are dealing with a multi-dimensional process, only Monte Carlo or quasi-Monte Carlo (and over multi-dimensional integration) methods are suitable numerical methods to determine the value of these options. As these methods can be very time consuming we will present alternative valuation methods which are based on analytical approximations in different senses.

2 ...and here are the candidates!

(a) Beisser's conditional expectation techniques

Beisser (1993) adapts an idea of Rogers and Shi (1995) introduced for pricing Asian options. By conditioning on the random variable Z and using Jensen's inequality the price of the basket call is estimated by the weighted sum of (artificial) European call prices, more precisely

$$\begin{aligned}
\mathrm{E}\left([B(T) - K]^+\right) &= \mathrm{E}\left(\mathrm{E}\left([B(T) - K]^+ \middle| Z\right)\right) \\
&\geq \mathrm{E}\left(\mathrm{E}\left(\left[B(T) - K\middle|Z\right]^+\right)\right) \\
&= \mathrm{E}\left(\left[\sum_{i=1}^n w_i E[S_i(T)|Z] - K\right]^+\right)
\end{aligned}$$

where

$$Z := \frac{\sigma_z}{\sqrt{T}} W(T) = \sum_{i=1}^{n} w_i S_i(0)\sigma_i W_i(T)$$

with σ_z appropriately chosen. Note that in contradiction to $S_i(T)$, all conditional expectations $E[S_i(T)|Z]$ are log-normally distributed with respect to *one* Brownian motion $W(T)$. Hence, there exists an x^*, such that

$$\sum_{i=1}^{n} w_i E\left[S_i(T) \middle| W(T) = x^* \right] = K.$$

By defining:

$$\widetilde{K}_i := E\left[S_i(T) \middle| W(T) = x^* \right]$$

the event $\sum_{i=1}^{n} w_i E[S_i(T)|Z] \geq K$ is equivalent to $E[S_i(T)|Z] \geq \widetilde{K}_i$ for all $i = 1, \ldots, n$.

Using this argument we conclude that

$$\begin{aligned} \mathrm{E}\left(\left[\sum_{i=1}^{n} w_i E[S_i(T)|Z] - K\right]^+\right) &= \sum_{i=1}^{n} w_i \,\mathrm{E}\left(\left[E[S_i(T)|Z] - \widetilde{K}_i\right]^+\right) \\ &= \sum_{i=1}^{n} w_i \left[\widetilde{F}_i^T N(d_{1i}) - \widetilde{K}_i N(d_{2i})\right] \end{aligned}$$

where $\widetilde{F}_i^T, \widetilde{K}_i$ adjusted forwards and strikes and d_{1i}, d_{2i} are the usual terms with modified parameters.

(b) Gentle's approximation by geometric average

Gentle (1993) approximates the arithmetic average in the basket payoff by a geometric average. The fact that a geometric average of log-normal random variables is again log-normally distributed allows for a Black–Scholes type valuation formula for pricing the approximating payoff. More precisely after rewriting the payoff of the basket option as

$$\begin{aligned} P_{Basket}(B(T), K, \theta) &= \left[\theta\left(\sum_{i=1}^{n} w_i S_i(T) - K\right)\right]^+ \\ &= \left[\theta\left(\left(\sum_{i=1}^{n} w_i F_i^T\right)\sum_{i=1}^{n} a_i S_i^*(T) - K\right)\right]^+, \end{aligned}$$

where

$$a_i = \frac{w_i F_i^T}{\sum_{i=1}^n w_i F_i^T},$$

$$S_i^*(T) = \frac{S_i(T)}{F_i^T} = \exp\left(-\frac{1}{2}\int_0^T \sigma_i^2 ds + \int_0^T \sigma_i dW_i(s)\right)$$

we approximate $\sum_{i=1}^n a_i S_i^*(T)$ by the geometric average, thus

$$\widetilde{B}(T) = \left(\sum_{i=1}^n w_i F_i^T\right)\prod_{i=1}^n \left(S_i^*(T)\right)^{a_i}.$$

To correct for the mean,

$$K^* = K - \left(\mathrm{E}(B(T)) - \mathrm{E}(\tilde{B}(T))\right)$$

is introduced. As approximation for $(B(T) - K)^+$, $(\tilde{B}(T) - K^*)^+$ is used, which—as $\tilde{B}(T)$ is log-normally distributed—can be valued by the Black–Scholes formula resulting in

$$V_{Basket}(T) = \mathrm{DF}(T)\theta\left(e^{\widetilde{m}+\frac{1}{2}\widetilde{v}^2} N(\theta d_1) - K^* N(\theta d_2)\right), \tag{2.1}$$

where $\mathrm{DF}(T)$ is the discount factor, $N(\cdot)$ the distribution function of a standard normal random variable and

$$d_1 = \frac{\widetilde{m} - \log K^* + \widetilde{v}^2}{\widetilde{v}},$$

$$d_2 = d_1 - \widetilde{v},$$

$$\widetilde{m} = \mathrm{E}(\log \tilde{B}(T)) = \log\left(\sum_{i=1}^n w_i F_i^T\right) - \frac{1}{2}\sum_{i=1}^n a_i \sigma_i^2 T \quad \text{and}$$

$$\widetilde{v}^2 = \mathrm{var}(\log \tilde{B}(T)) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \sigma_i \sigma_j \rho_{ij} T.$$

(c) Levy's log-normal moment matching

The basic idea of Levy (1992) is to approximate the distribution of the basket by a log-normal distribution $\exp(X)$ with mean M and variance $V^2 - M^2$, such that the first two moments of this and of the original distribution of the weighted sum of the stock prices coincide, i.e.

$$m = 2\log(M) - 0.5\log(V^2)$$

$$v^2 = \log(V^2) - 2\log(M) \quad \text{and}$$

$$M \equiv \mathrm{E}(B(T)) = \sum_{i=1}^n w_i F_i(T)$$

$$V^2 \equiv \mathrm{E}(B^2(T)) = \sum_{i=1}^{n} w_i w_j F_i^T F_j^T \exp(\sigma_i \sigma_j \rho_{ij} T)$$

result in

$$\mathrm{E}(B(T)) = \mathrm{E}\left(e^X\right) = e^{m+0.5v^2} \quad \text{and} \quad \mathrm{E}(B^2(T)) = \mathrm{E}\left(e^{2X}\right) = e^{2m+2v^2}$$

where X is a normally distributed random variable with mean m and variance v^2.

The basket option price is now approximated by

$$V_{Basket}(T) \approx \mathrm{DF}(T)\,(M N(d_1) - K N(d_2))$$

with

$$d_1 = \frac{m - \ln(K) + v^2}{v},$$
$$d_2 = d_1 - v.$$

Note the subtle difference to Gentle's method. Here, the distribution of $B(T)$ is approximated directly by a log-normal distribution that matches the first two moments, while in Gentle's approximation only the first moment is matched.

(d) Ju's Taylor expansion

Ju (2002) considers a Taylor expansion of the ratio of the characteristic function of the arithmetic average to that of the approximating log-normal random variable around zero volatility. He includes terms up to σ^6 in his closed-form solution.

Let

$$A(z) = \sum_{i=1}^{n} F_i^T \exp\left(-\frac{1}{2}(z\sigma_i)^2 T + z\sigma_i W_i(T)\right)$$

be the arithmetic mean where the volatilities are scaled by a parameter z. Note that for $A(1)$ we recover the original mean. Let $Y(z)$ be a normally distributed random variable with mean $m(z)$ and variance $v(z)$ such that the first two moments of $\exp(Y(z))$ match those of $A(z)$. The appropriate parameters are derived in section (c), only σ_i has to be replaced by $z\sigma_i$. Let $X(z) = \log(A(z))$, then the characteristic function is given as:

$$E[e^{i\phi X(z)}] = E[e^{i\phi Y(z)}]\frac{E[e^{i\phi X(z)}]}{E[e^{i\phi Y(z)}]} = E[e^{i\phi Y(z)}]f(z),$$

where

$$E[e^{i\phi Y(z)}] = e^{i\phi m(z) - \phi^2 v(z)/2}$$
$$f(z) = E[e^{i\phi X(z)}]e^{-i\phi m(z) + \phi^2 v(z)/2}$$

Ju performs a Taylor expansion of the two factors of $f(z)$ up to z^6, leading to

$$f(z) \approx 1 - i\phi d_1(z) - \phi^2 d_2(z) + i\phi^3 d_3(z) + \phi^4 d_4(z),$$

where $d_i(z)$ are polynomials of z and terms of higher order than z^6 are ignored. Finally $E[e^{i\phi X(z)}]$ is approximated by

$$E[e^{i\phi X(z)}] \approx e^{i\phi m(z) - \phi^2 v(z)/2}(1 - i\phi d_1(z) - \phi^2 d_2(z) + i\phi^3 d_3(z) + \phi^4 d_4(z)).$$

For this approximation, an approximation of the density $h(x)$ of $X(1)$ is derived as

$$h(x) = p(x) + \left(\frac{d}{dx}d_1(1) + \frac{d^2}{dx^2}d_2(1) + \frac{d^3}{dx^3}d_3(1) + \frac{d^4}{dx^4}d_4(1)\right)p(x)$$

where $p(x)$ is the normal density with mean $m(1)$ and variance $v(1)$. The approximate price of a basket call is then given by

$$\begin{aligned} V_{Basket}(T) = \mathrm{DF}(T)\Bigg\{&\left[\left(\sum w_i F_i^T\right)N(d_1) - KN(d_2)\right] \\ &+K\left[z_1 p(y) + z_2\frac{dp(y)}{dy} + z_3\frac{d^2p(y)}{dy^2}\right]\Bigg\}, \end{aligned}$$

where

$$y = \log(K), d_1 = \frac{m(1) - y}{\sqrt{v(1)}} + \sqrt{v(1)}, d_2 = d_1 - \sqrt{v(1)}$$

and $z_1 = d_2(1) - d_3(1) + d_4(1)$, $z_2 = d_3(1) - d_4(1)$, $z_3 = d_4(1)$. Note that the first summand is equal to Levy's approximation and the second summand gives the higher order corrections.

(e) The reciprocal gamma approximation by Milevsky and Posner

Milevsky and Posner (1998a) use the reciprocal gamma distribution as an approximation for the distribution of the basket. The motivation is the fact that the distribution of correlated log-normally distributed random variables converges to the reciprocal gamma distribution as $n \to \infty$. Consequently, the first two moments of both distributions are matched to obtain a closed-form solution. Let G_R be the reciprocal gamma distribution and G the gamma distribution with parameters α, β, then per definition:

$$G_R(y, \alpha, \beta) = 1 - G(1/y, \alpha, \beta).$$

If the random variable Y is reciprocally gamma distributed, then

$$E[Y^i] = \frac{1}{\beta^i(\alpha - 1)(\alpha - 2)\ldots(\alpha - i)}$$

and with M and V^2 denoting the first two moments as defined in the previous section, we get:

$$\alpha = \frac{2V^2 - M^2}{V^2 - M^2}$$

$$\beta = \frac{V^2 - M^2}{V^2 M}$$

Basic calculations yield:

$$V_{Basket}(T) \approx \mathrm{DF}(T)\,(MG(1/K, \alpha - 1, \beta) - KG(1/K, \alpha, \beta))$$

Note that we use the parametrisation of the gamma distribution found in Staunton (2002), since this produces more accurate results than that from the original paper by Milevsky and Posner (1998a).

(f) Milevsky and Posner's approximation via higher moments

Milevsky and Posner (1998b) use distributions from the Johnson (1994) family as state price densities to match higher moments of distribution of the arithmetic mean. More precisely, they write the price of a call on a basket as:

$$V_{Basket}(T) = \mathrm{DF}(T)\left[\int_0^{\infty} (x - K)^+ h(x)dx\right]$$

where $h(x)$ is the state price density. Note that we would end up in Levy's approximation if we were using the log-normal density with the first two moments matching those of the mean. Milevsky and Posner, however, use two members of the Johnson family, which is a collection of statistical distributions, that can be represented by a transformation of the normal distribution Z:

$$\text{Type I}: X = c + d\exp\left(\frac{Z - a}{b}\right) \quad \text{or}$$

$$\text{Type II}: X = c + d\sinh\left(\frac{Z - a}{b}\right)$$

The parameters a, b, c and d are chosen, so that the four moments of the arithmetic mean are approximated (since there are no closed-form solutions for them). If the kurtosis of the Type I is close enough to the kurtosis of the mean, they use Type I, otherwise Type II. The closed-form solution for Type I is given by:

$$V_{Basket}(T) \approx \mathrm{DF}(T)\left[M - K + (K - c)N(Q) - d\exp\left(\frac{1 - 2ab}{2b^2}\right)N\left(Q - \frac{1}{b}\right)\right]$$

where

$$M = \sum_i^n a_i F_i^T$$

$$Q = a + b\log\left(\frac{K - c}{d}\right)$$

$$\omega = \frac{1}{2}\sqrt[3]{8 + 4\eta^2 + 4\sqrt{4\eta^2 + \eta^4}} + \frac{2}{\frac{1}{2}\sqrt[3]{8 + 4\eta^2 + 4\sqrt{4\eta^2 + \eta^4}}} - 1$$

$$a = 1/\sqrt{\log(\omega)},\ b = \frac{1}{2}\log(\omega(\omega - 1)/\xi^2),\ d = \ sign(\eta),\ c = dM - e^{(\frac{1}{2b} - a)/b}$$

where ξ is the variance, η the skewness and κ the kurtosis.

3 Test results

As the advantage of analytical methods compared to Monte Carlo or numerical integration is of course speed of computations, we only have to compare the accuracy of the analytical methods presented in the foregoing section.

We will perform a systematic test by looking at the effect of varying correlations, strikes, forward and strikes and volatilities. Our standard test example is a call option on a basket with four stocks and parameters given by

$$\begin{aligned} T &= 5.0, \\ \mathrm{DF}(T) &= 1.0, \\ \rho_{ij} &= 0.5 \qquad (\text{for } i \neq j), \\ K &= 100, \\ F_i^T &= 100, \\ \sigma_i &= 40\% \quad \text{and} \\ w_i &= \frac{1}{4}. \end{aligned}$$

As reference values we compute the prices of all the options below by a Monte Carlo simulation using the antithetic method and geometric mean as control variate for variance reduction. The number of simulations was always chosen large enough to keep the standard deviation below 0.05.

We did not test the method of Huynh (1994), because it is an application of the method of Turnbull and Wakeman (1991) for Asian options (Edgeworth expansion up to the 4th moment) and it is a well-known problem that this approximation gives really bad results for long maturities and high volatilities. See also Ju (2002), who pointed out that the Edgeworth expansion diverges if the approximating random variable is log-normal.

We also tested Curran's (1994) approximation which computes the price by conditioning on the geometric mean. But we do not show the numerical results here, because—if we transformed the forwards to one (simply by multiplying the weights with them)—the prices were *exactly* the same as those of Beisser (1999). If we did not transform the forwards to one, Beisser and Curran gave different prices, but on the other hand Curran's results were mostly worse. For further reading we refer to Deelstray *et al*. (2003) and Beisser (2001) who developed a general framework for the pricing of baskets and Asian options via conditioning.

(a) Varying the correlations

Table 1 shows the effect of simultaneously changing all correlations from $\rho = \rho_{ij} = 0.1$ to $\rho = 0.95$. Note that, except for Milevsky and Posner's reciprocal gamma (MP−RG) and Gentle, all methods perform reasonably well. Especially for $\rho \geq 0.8$, the methods of Beisser, Ju, Levy, the four moments method of Milevsky and Posner (MP−4M) and Monte Carlo give virtually the same price.

TABLE 1: VARYING THE CORRELATIONS SIMULTANEOUSLY

ρ	Beisser	Gentle	Ju	Levy	MP−RG	MP−4M	Monte Carlo CV	StdDev
0.10	20.12	15.36	21.77	22.06	20.25	21.36	21.62	(0.0319)
0.30	24.21	19.62	25.05	25.17	22.54	24.91	24.97	(0.0249)
0.50	27.63	23.78	28.01	28.05	24.50	27.98	27.97	(0.0187)
0.70	30.62	27.98	30.74	30.75	26.18	30.74	30.72	(0.0123)
0.80	31.99	30.13	32.04	32.04	26.93	32.04	32.03	(0.0087)
0.95	33.92	33.41	33.92	33.92	27.97	33.92	33.92	(0.0024)
Dev.[1]	0.700	4.013	0.071	0.203	4.119	0.108		

Dev. $= \frac{1}{n}\sqrt{\sum_{i=1}^{n}(Price - MCPrice)^2}$.

The good performance of Beisser, Ju, Gentle and Levy for high correlations can be explained as follows: all four methods provide exactly the Black−Scholes prices for the special case that the number of stocks is one. For high correlations the distribution of the basket is approximately the sum of the *same* (for $\rho = 1$ exactly the same) log-normal distributions, which is indeed again log-normal. As Levy uses a log-normal distribution with the correct moments, it has to be a good approximation for these cases. The same argumentation applies for Gentle. If we have effectively one stock the geometric and the arithmetic average are the same. The bad performance of MP−RG for high correlations can be explained by the fact that with effectively one stock we are far away from 'infinitely' many stocks, which is the motivation for this method. A test with fixed correlation $\rho_{12} = 0.95$ and varying the remaining correlations symmetrically shows exactly the same result.

In total the prices calculated by Ju's approach (whose method slightly overprices) and MP−4M are overall the closest to the Monte Carlo prices. These approaches are followed by Levy's and Beisser's approximation (whose approach slightly underprices). The other two methods are not recommendable.

(b) Varying the strikes

With all other parameters set to the default values, the strike K is varied from 50 to 150. Table 2 contains the results.

The differences between the prices calculated by Monte Carlo and the approaches of Ju and MP−4M are relatively small. The price curves of the method of Gentle and Milevsky and Posner's

TABLE 2: VARYING THE STRIKE

K	Beisser	Gentle	Ju	Levy	MP−RG	MP−4M	Monte Carlo CV	StdDev
50.00	54.16	51.99	54.31	54.34	51.93	54.35	54.28	(0.0383)
60.00	47.27	44.43	47.48	47.52	44.41	47.50	47.45	(0.0375)
70.00	41.26	37.93	41.52	41.57	38.01	41.53	41.50	(0.0369)
80.00	36.04	32.40	36.36	36.40	32.68	36.34	36.52	(0.0363)
90.00	31.53	27.73	31.88	31.92	28.22	31.86	31.85	(0.0356)
100.00	27.63	23.78	28.01	28.05	24.50	27.98	27.98	(0.0350)
110.00	24.27	20.46	24.67	24.70	21.39	24.63	24.63	(0.0344)
120.00	21.36	17.65	21.77	21.80	18.77	21.73	21.74	(0.0338)
130.00	18.84	15.27	19.26	19.28	16.57	19.22	19.22	(0.0332)
140.00	16.65	13.25	17.07	17.10	14.70	17.04	17.05	(0.0326)
150.00	14.75	11.53	15.17	15.19	13.10	15.14	15.15	(0.0320)
Dev.	0.323	3.746	0.031	0.065	3.038	0.030		

reciprocal gamma approach (MP−RG) run almost parallel to the Monte Carlo curve and represent an underevaluation. The relative and absolute differences of all methods are generally increasing when K is growing, since the approximation of the real distributions in the tails is getting worse and the absolute prices are decreasing.

Again, overall Ju's approximation and MP−4M perform best, while Ju's slightly overprices. Levy is the third and Beisser the fourth best.

(c) Varying the forwards and strikes

The forwards on all stocks are now set to the same value F which is varied between 50 and 150 in this set of tests. Table 3 shows that MP−4M and Ju's method perform excellently, while the second one again typically slightly overprices. Levy and Beisser's method also performs well and Beisser again slightly underprices. The other methods perform worse. These effects can also be seen if some forwards are fixed and the remaining ones are varied.

(d) Varying the volatilities

The next set of tests involves varying the volatilities σ_i. We start with the symmetrical situation at each step, σ_i is set to the same value σ, which is varied between 5% and 100%.

Table 4 shows the results of the test.

The prices calculated by the different methods are more or less equal for 'small' values of the volatility. They start to diverge at $\sigma \approx 20\%$. The Monte Carlo, Beisser, Ju and Levy prices remain close, whereas the prices calculated by the other methods are too low.

The picture obtained so far completely changes if we have asymmetry in the volatilities, precisely if there are groups of stocks with high and with low volatilities entering the basket. This is clearly demonstrated by Figure 1 where we fix $\sigma_1 = 5\%$ and vary the remaining volatilities symmetrically. This time the prices diverge much more. The method of Levy is massively overpricing

TABLE 3: VARYING THE FORWARDS SIM. WITH $K = 100$

F	Beisser	Gentle	Ju	Levy	MP−RG	MP−4M	Monte Carlo CV	StdDev
50.00	4.16	3.00	4.34	4.34	3.93	4.33	4.34	(0.0141)
60.00	7.27	5.53	7.51	7.52	6.56	7.50	7.50	(0.0185)
70.00	11.26	8.91	11.55	11.57	9.95	11.53	11.53	(0.0227)
80.00	16.04	13.13	16.37	16.40	14.10	16.34	16.35	(0.0268)
90.00	21.53	18.11	21.89	21.92	18.97	21.86	21.86	(0.0309)
100.00	27.63	23.78	28.01	28.05	24.50	27.98	27.98	(0.0350)
110.00	34.27	30.08	34.66	34.70	30.63	34.63	34.63	(0.0391)
120.00	41.36	36.91	41.75	41.80	37.32	41.73	41.71	(0.0433)
130.00	48.84	44.21	49.23	49.28	44.49	49.21	49.19	(0.0474)
140.00	56.65	51.92	57.04	57.10	52.08	57.03	57.00	(0.0516)
150.00	64.75	59.98	65.13	65.19	60.05	65.14	65.08	(0.0556)
Dev.	0.316	3.989	0.031	0.072	3.516	0.022		

TABLE 4: VARYING THE VOLATILITIES SIM. WITH $K = 100$

σ	Beisser	Gentle	Ju	Levy	MP−RG	MP−4M	Monte Carlo CV	StdDev
5%	3.53	3.52	3.53	3.53	3.52	3.53	3.53	(0.0014)
10%	7.04	6.98	7.05	7.05	6.99	7.05	7.05	(0.0042)
15%	10.55	10.33	10.57	10.57	10.36	10.57	10.57	(0.0073)
20%	14.03	13.52	14.08	14.08	13.59	14.08	14.08	(0.0115)
30%	20.91	19.22	21.08	21.09	19.49	21.07	21.07	(0.0237)
40%	27.63	23.78	28.01	28.05	24.50	27.98	27.98	(0.0350)
50%	34.15	27.01	34.84	34.96	28.51	34.73	34.80	(0.0448)
60%	40.41	28.84	41.52	41.78	31.56	41.19	41.44	(0.0327)
70%	46.39	29.30	47.97	48.50	33.72	46.23	47.86	(0.0490)
80%	52.05	28.57	54.09	55.05	35.15	48.39	54.01	(0.0685)
100%	62.32	24.41	64.93	67.24	36.45	47.90	65.31	(0.0996)
Dev.	1.22	16.25	0.12	0.69	11.83	5.53		

with all other methods underpricing. We note that Ju's and Beisser's method performs best. Particularly remarkable is the excellent performance of Ju for high volatilities. Since it is a Taylor expansion around zero volatilities, one would not expect the validity of this expansion far away from zero.

The same test but now with $\sigma_1 = 100\%$ results in Table 5 and Figure 2. Note the extremely bad performance for Levy's method for small values of σ which is even outperformed by Gentle's method! Beisser is the only one who can deal with this parameter, while both Milevsky's and Posner's methods are also bad.

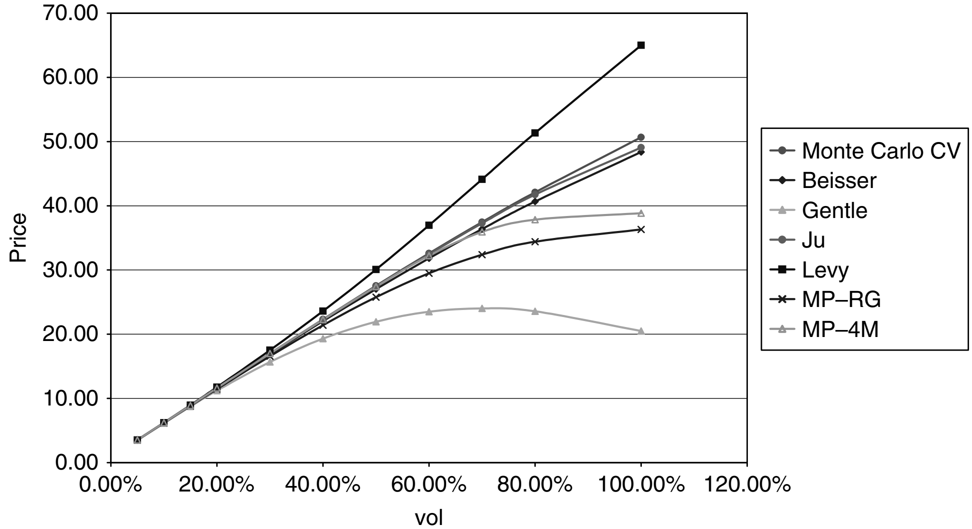


Figure 1: Varying the volatilities sim. with $\sigma_1 = 5\%$, $K = 100$

TABLE 5: VARYING THE VOLATILITIES SIM. WITH $\sigma_1 = 100\%$, $K = 100$ (FIGURE 2)

σ	Beisser	Gentle	Ju	Levy	MP–RG	MP–4M	Monte Carlo CV	StdDev
5.00%	19.45	15.15	35.59	55.46	35.22	18.51	22.65	(0.5594)
10.00%	20.84	16.60	36.19	55.52	35.23	18.64	21.30	(0.3858)
15.00%	22.60	18.08	36.93	55.61	35.24	18.81	22.94	(0.2660)
20.00%	24.69	19.56	37.80	55.71	35.26	19.01	25.24	(0.2124)
30.00%	29.52	22.35	39.97	55.98	35.30	19.42	30.95	(0.1603)
40.00%	34.72	24.73	42.66	56.35	35.36	20.37	36.89	(0.1156)
50.00%	39.96	26.52	45.84	56.89	35.44	20.60	41.72	(0.0894)
60.00%	45.05	27.59	49.39	57.68	35.56	21.72	46.68	(0.0472)
70.00%	49.88	27.87	53.21	58.87	35.72	23.66	51.78	(0.0587)
80.00%	54.39	27.38	57.17	60.70	35.93	27.38	56.61	(0.0742)
100.00%	62.32	24.41	64.93	67.24	36.45	47.90	65.31	(0.0996)
Dev.	1.92	19.18	8.96	22.70	14.48	17.84		

(e) Implicit distributions

In addition we plot the *implicit distribution* of the particular approximations and compare them to the real ones calculated by Monte Carlo simulation. With *implicit distribution* we mean that we derive the underlying distribution of the particular methods by an appropriate portfolio of calls.

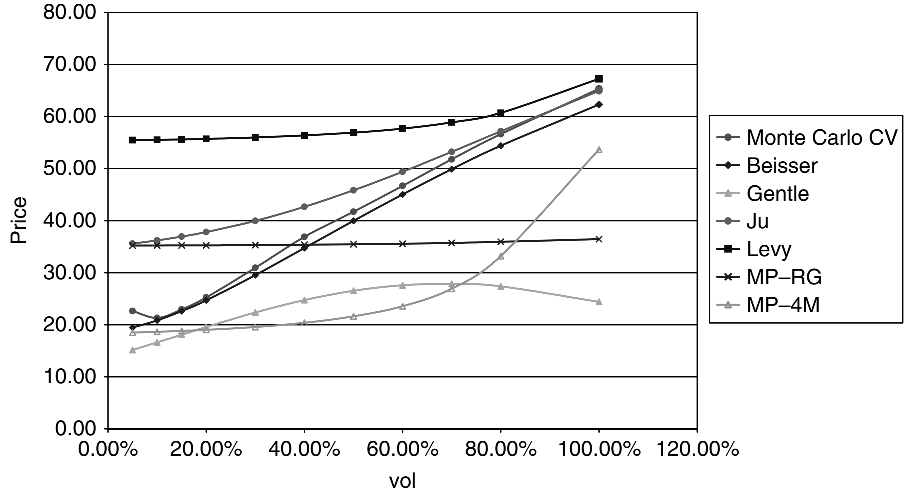


Figure 2: Varying the volatilities sim. with $\sigma_1 = 100\%$, $K = 100$ (Table 5)

Consider the payoff of the following portfolio consisting only of calls:

$$\Pi(B(T)) = \alpha * \left[P_{Basket}\left(B(T), L - \frac{1}{\alpha}, 1\right) - P_{Basket}(B(T), L, 1) - \left(P_{Basket}(B(T), L + \triangle L, 1) - P_{Basket}\left(B(T), L + \triangle L + \frac{1}{\alpha}, 1\right)\right)\right].$$

We notice that the payoff $\Pi(B(T))$ is explicitly given by

$$\Pi(B(T)) = \begin{cases} 0 & : \quad B(T) < L - \frac{1}{\alpha} \\ \alpha\left[B(T) - (L - \frac{1}{\alpha})\right] & : \quad L - \frac{1}{\alpha} \le B(T) \le L \\ 1 & : \quad L \le B(T) \le L + \triangle L \\ 1 - \alpha\left[B(T) - (L + \triangle L)\right] & : \quad L + \triangle L \le B(T) \le L + \triangle L + \frac{1}{\alpha} \\ 0 & : \quad B(T) > L + \triangle L + \frac{1}{\alpha} \end{cases} \tag{3.1}$$

For $\alpha \to \infty$ it is equal to:

$$\Pi(B(T)) = \begin{cases} 0 & : \quad B(T) < L \\ 1 & : \quad L \le B(T) \le L + \triangle L \\ 0 & : \quad B(T) > L + \triangle L \end{cases}$$

So for a sufficiently high α the *value* of our portfolio is approximately the probability that the price of the basket is at maturity in $[L, L + \triangle L]$. To calculate the whole implicit distribution, we shift the boundaries stepwise by $\triangle L$. Instead of applying the underlying distributions, we use this procedure, because we cannot directly determine the distribution for Beisser's approximation. Besides, this procedure seems to be more objective and consistent to compare the approximations.

We examined the distributions for the test cases (a)–(d). The results confirmed our findings from the comparison of the prices. For the cases (a)–(c) the implicit distributions of Ju, Levy and Beisser were consistent with Monte Carlo, and the other ones not. But only Beisser was able to deal with inhomogeneous volatilities in case (d), where Levy showed massive deviations.

We plot an example with $\sigma_1 = 90\%, \sigma_2 = \sigma_3 = 50\%$ and $\sigma_4 = 10\%$ in Figure 3 to test if there is some 'balancing' effect, i.e. observe that $(\sigma_1 + \sigma_4)/2 = \sigma_2$. We see there is one except for Levy's approach.

We did not plot the graph for the state price density method of Milevsky and Posner, because it was running into serious problems for small K. The parameter Q is defined as $a + b\log((K - c)/d)$, hence for all $K < c$ the formula of Milevsky and Posner is not well defined (a similar problem occurs for Type II). But for this parameter set c is around 65, so we simply couldn't calculate all necessary prices.

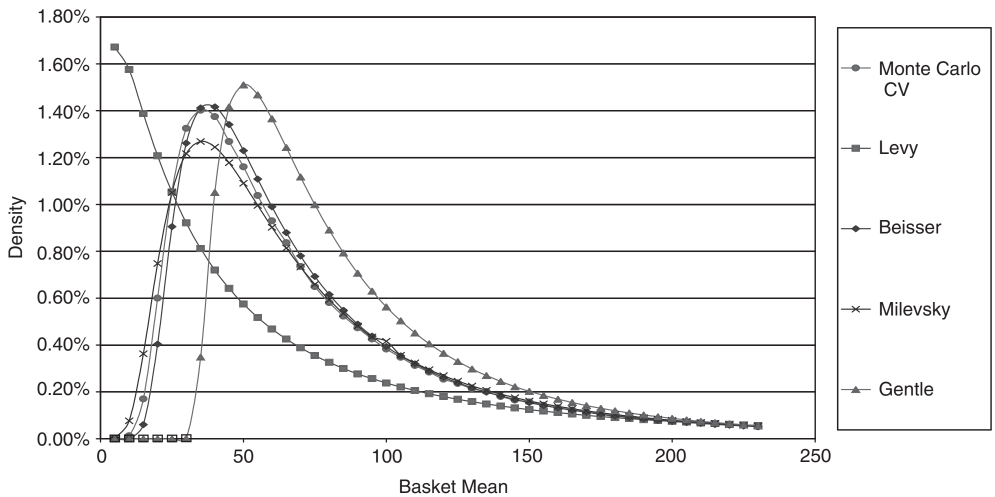


Figure 3: Densities for the standard scenario with $\sigma_1 = 90\%, \sigma_2 = \sigma_3 = 50\%, \sigma_4 = 10\%$

So which method to choose? The tests confirm that the approximation of Ju is overall the best performing method. In addition it has the nice property that it always overprices slightly. Ju's method shows only a little weakness in the case of inhomogeneous volatilities, where Beisser is better. Even though it is based on a Taylor expansion around zero volatilities, it has absolutely no problems with high volatilities, which is quite contrary to both methods of Milevsky and Posner.

Beisser's approximation underprices slightly in all cases. The underpricing of Beisser's approach is not surprising since this method is essentially a lower bound on the true option price. Beisser's approach is the only method which is reliable in the case of inhomogeneous volatilities.

The performances of Milevsky and Posner's reciprocal gamma and Gentle's approach are mostly poor. A reason for the bad performance of MP–RG may be that the sum of log-normally distributed random variables is *only* distributed like the reciprocal gamma distribution as $n \to \infty$. But as in our case where $n = 4$ or even in practice with $n = 30$ we are far away from infinity.

The geometric mean used in Gentle's approach also seems to be an inappropriate approximation for the arithmetic mean. For instance, the geometric mean of the forwards equal to 1, 2, 3 and 4 would be without mean correction 2.21 instead of 2.5. This is corrected, but the variance is still wrong. The MP–4M four moment method is recommended only for low vols.

The Ju method is the best approximation except for the case of inhomogeneous volatilities. The reason for this drawback may be that all stocks are 'thrown' together on one distribution. This is quite contrary to Beisser's approximation, where every single stock keeps a transformed log-normal distribution and the expected value of every stock is individually evaluated. This is probably the reason why this method is able to handle the case of inhomogeneous volatilities.

A rule of thumb for a practitioner would be to use Ju's method for homogeneous volatilities and Beisser's for inhomogeneous ones. But then the question occurs, how to define the switch exactly. So we suggest the following: price the basket with Ju and Beisser; if the relative difference between the two computed values is less than 5% use Ju's price for an upper and Beisser's price for a lower bound. If it is bigger than 5% run a Monte Carlo simulation or if this is not suitable, keep the Beisser result (keep in mind that it is only a lower bound for the prices!).

REFERENCES

■ Beisser, J. (1999) Another way to value basket options. Working paper, Johannes Gutenberg-Universität Mainz.

■ Beisser, J. (2001) Topics in Finance — A conditional expectation approach to value Asian, Basket and Spread Options. Ph.D. Thesis, Johannes Gutenberg-Universität Mainz.

■ Curran, M. (1994) Valuing Asian and portfolio options by conditioning on the geometric mean price. *Management Science*, 40, 1705–1711.

■ Deelstra, G. Liinev, J. and Vanmaele, M. (2003) Pricing of arithmetic basket and Asian basket by conditioning. Working paper, Ghent University, Belgium.

■ Gentle, D. (1993) Basket weaving. *Risk*, 51–52.

■ Huynh,C. B. (1994) Back to baskets. *Risk*, 7(5), 59–61.

■ Johnson, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika*, 36, 149–176.

■ Ju, E. (1992) Pricing Asian and basket options via Taylor expansion. *Journal of Computational Finance*, 5(3), 79–103.

■ Levy, E. (1992) Pricing European average rate currency options. *Journal of International Money and Finance*, 11, 474–491.

■ Levy, E. and Turnbull, S. (1992) Average intelligence. *Risk*, 6, No. 2, February, 5–9.

■ Milevksy, M. A. and Posner, S. E. (1998a) A closed-form approximation for valuing basket options. *Journal of Derivatives*, 54–61.

■ Milevksy, M. A. and Posner, S. E. (1998b) Valuing exotic options by approximating the SPD with higher moments. *Journal of Financial Engineering*, 7(2), 54–61.

■ Rogers, L. C. G. and Shi, Z. (1995) The value of an Asian option. *Journal of Applied Probability*, 32, 1077–1088.

■ Staunton, M. (2002) From nuclear power to basket options. *Wilmott Magazine*, September, 46–48.

■ Turnbull, St. M. and Wakeman, L. M. (1991) A quick algorithm for pricing European average options. *Journal of Financial and Quantitative Analysis*, 26, No. 3, September, 377–389.

17
Pricing CMS Spread Options and Digital CMS Spread Options with Smile

Mourad Berrahoui

1 Introduction

This chapter deals with the smile of spread options in the Black framework. The price of spread options is sensitive to the entire smile of both underlyings. The classical approach uses the Black model without smile. For each underlying, the corresponding at-the-money volatility is taken. This approach ignores the effect of the smile and this is even more of a problem when we deal with digital options, as in this case there is a smile effect caused directly by the slope of the smile.

In general no closed formula exists for pricing a spread option when the strike is different to zero. We don't focus in this chapter on the numerical method. A very detailed survey on the valuation of spread options is given in Carmona and Durrleman (2003).

Dempster and Hong (2001) propose to use the fast Fourier transform (FFT) with stochastic volatility and interest rate environments.

Alexander and Scourse (2003) propose to value spread options with a bivariate normal mixture distribution.

An interesting study has been done, see Cherubini and Luciano (2002), where a non-Gaussian copula has been proposed to associate the marginal distribution. This copula is calibrated using historical data.

The aim of this chapter is to develop a simple approach, easy to implement with exogenous input smile with some application on CMS product.

We start by presenting the current approach used in different banks, then we propose two different methods to take into account the smile. The first method consists of changing the strike where the volatility of each underlying is taken and represents only a partial modeling of the smile. The second method takes into account the full smile of each underlying and involves some numerical integration. These two methods are used to show the errors generated by the old approach. In the last section, we extend the two methods to CMS underlyings, and give some ideas how to generate an artificial smile using the same approach as above.

2 Notations

The following notations are used throughout this document.

Let's consider two assets F_1 and F_2 and an option of maturity T depending on those two assets. We assume that under the T forward probability, each F_i $(i = 1, 2)$ follows a lognormal process according to the stochastic differential equation:

$$\frac{dF_i(t)}{F_i(t)} = \mu(F_i(t), t)\, dt + \sigma_i(t)\, dW_i(t) \tag{1}$$

Correlation between the two assets is represented by the fact that the two standard Brownian processes in equation (1) satisfy:

$$E[dW_1 \cdot dW_2] = \rho dt \tag{2}$$

A spread option also called crack spreads, due to their use in the oil industry, gives the holder the right to exchange F_2 for F_1 at expiry. The payoff is:

$$payoff = \text{Max}(Q_1 F_1 - Q_2 F_2 - K; 0) \tag{3}$$

where Q_1 is the quantity of asset F_1, Q_2 the quantity of asset F_2, and K the strike.

3 Current approach without smile

3.1 Spread option with zero strike

When $K = 0$, a closed-form formula exists (Margrabe 1978). We assume that the drift in equation (1) is deterministic. The price P of this option is:

$$P = Q_1 F_1 B(0, T) e^{\mu_1} N(d_1) - Q_2 F_2 B(0, T) e^{\mu_2} N(d_2) \tag{4}$$

where

$$d_1 = \frac{\ln(Q_1 F_1 / Q_2 F_2) + (\mu_1 - \mu_2 + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

$$\mu_i = \int_0^T \mu_i(t)\, dt; i = 1, 2$$

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$\sigma_i = \sqrt{\frac{1}{T}\int_0^T \sigma_i(t)\,dt};\ i = 1, 2$$

and $B(0, T)$ is the price of a zero coupon of maturity T.

3.2 Spread option with non-zero strike

Theoretical price of a spread option To calculate the spread option price in the case where $K \neq 0$, it is necessary to write equations (1) and (2) differently to use only independent Brownian motions $\tilde{W}_1$ and $\tilde{W}_2$, as follows:

$$\frac{dF_1}{F_1} = \mu_1 dt + \sigma_1 \left(\rho d\tilde{W}_t^1 + \sqrt{1-\rho^2} d\tilde{W}_t^2\right) \tag{5}$$

$$\frac{dF_2}{F_2} = \mu_2 dt + \sigma_2 dW_t^1 \tag{6}$$

The price P is the discounted expectation of payoff (3) under the T forward probability Q_T where T is the maturity of the option we want to price:

$$P = B(0, T) \cdot E_{Q_T}[\mathrm{Max}(Q_1 F_1(T) - Q_2 F_2(T) - K; 0)] \tag{7}$$

There is no closed-form formula but two different numerical methods are available to calculate P: Monte Carlo and semi-analytical.

Monte Carlo approach We simulate the two processes F_1 and F_2. The price P corresponds to the mean of (7) over the set of Monte Carlo paths.

Semi-analytical approach Different approaches exist:

- Apply a conditioning technique to turn the two-dimensional integral into a single one (Ravindran 1993, Shimko 1994)

$$P = B(0, T) \cdot E_{Q_T}\{E_{Q_T}[\mathrm{Max}(Q_1 F_1(T) - Q_2 F_2(T) - K; 0) | F_2(T)]\} \tag{8}$$

- Fast Fourier transform (Carr and Madan 1999, Dempster and Hong 2001).

4 New approach with smile

4.1 A simple way to take into account a partial smile

The problem with the formulas presented in the last section in the presence of smile is what volatility to use for each index. In general, we use the volatility at-the-money for each underlying.

In some special cases it is possible to determine a strike at which to take the volatility of each underlying rather than the money. Let's assume that the asset F_2 is less volatile. So the spread option becomes simply a mono-underlying option and the volatility to use for F_1 corresponds to the strike $F_2(0) + K$. On the other hand, if we suppose that the asset F_1 is less volatile, then the volatility to use for F_2 corresponds to the strike $F_1(0) - K$.

On the basis of this reasoning, we propose to use in general:

$$Vol(F_1) = Vol(Strike = ATM(F_2) + K)$$
$$Vol(F_2) = Vol(Strike = ATM(F_1) - K)$$

We will show later how accurate this approximation is in comparison to the habit of using at-the-money volatility in case of a deeply in/out-of-the-money option.

Just to give an example, imagine that we try to price a spread USD CMS 20Y and USD Libor 3M at 06/17/2003 (Libor 3M = 1.02%, Swap 20Y = 4.299%) with strike equal to 3.279% (4.299% − 1.02%). When the option is at-the-money (as is the case at the beginning of the trade), there is no difference between the two methods. However, when the spread moves, the option becomes deeply out- or in-the-money and the more convex the smile, the greater the difference between the two methods. Even if the option was dealt at zero strike, because the smile for the indexes Libor 3M and CMS 20Y is quite different, the two methods still lead to different prices.

4.2 How to take into account the entire smile

The formula given for the price of a spread option in the previous sections cannot be extended to calculate a price with smile. For this, we need a more general expression for the price which does not assume that F_1 and F_2 follow lognormal distributions. The following formula is true independently of the distribution of the underlying:

$$C = B(0,T)\int_0^{+\infty} \text{Prob}(Q_1F_1(T) > x + K, Q_2F_2(T) \le x)\, dx \tag{9}$$

where Prob(...) is the bivariate cumulative distribution with correlation equal to ρ.

In order to prove (9), we need the following proposition.

Proposition 1 The spread option payoff is a sum of product of digital options:

$$Max(Q_1F_1(T) - Q_2F_2(T) - K; 0) = \int_0^{+\infty} 1_{\{Q_1F_1(T)>x+K\}} \cdot 1_{\{Q_2F_2(T)\le x\}} dx \tag{10}$$

Proof We have just to change the boundary of the integral in (10) by

$$x < Q_1F_1(T) - K$$
$$x \ge Q_2F_2(T)$$

(9) is then obtained by taking the expectation of (10).

The integral in (9) can be calculated numerically using simple methods: trapezoidal rule, Simpson's rule..., or high-order methods: Gauss, Gauss–Kronrod.

All those methods involve approximating (9) in the discrete form:

$$P = B(0,T) \cdot \sum_i w_i \text{Prob}(Q_1F_1(T) > x_i + K, Q_2F_2(T) \le x_i) \tag{11}$$

where w_i is a series of quadrature weights.

We are now faced with the problem of calculating the probability in (11) in the presence of smile. The probability that one asset is above a fixed strike can be retrieved easily from prices of call options. Here we need to calculate a bivariate probability. By no arbitrage, we can find (see Cherubini and Luciano 2002) a lower and upper limit

$$\begin{aligned} P1 - \mathrm{Min}(P1, P2) &\leq \mathrm{Prob}(F_1(T) > x + K, F_2(T) \leq x) \\ &\leq P1 - \mathrm{Max}(P1 + P2 - 1, 0) \end{aligned}$$

with

$$P1 = \mathrm{Prob}(F_1(T) > x + K)$$

$$P2 = \mathrm{Prob}(F_1(T) > x)$$

These limits represent the financial application of the minimal and maximal copulas of the Frechet–Hoeffding inequality.

Copulas help us to calculate the bivariate probability knowing the marginal distribution for each underlying (call spread price), and for that the following assumption is needed:

Gaussian copula assumption

$$\begin{aligned} &\mathrm{Prob}(F_1(T) > x_1, F_2(T) \leq x_2 | \text{Full smile}) \\ &\quad = \mathrm{Prob}(F_1(T) > \tilde{x}_1, F_2(T) \leq \tilde{x}_2 | \sigma_1 = \Sigma_1(T, x_1)); \sigma_2 = \Sigma_2(T, x_2))) \end{aligned} \tag{12}$$

with $\tilde{x}_1$ and $\tilde{x}_2$ such that

$$\mathrm{Prob}(F_1(T) > \tilde{x}_1 | \sigma_1 = \Sigma_1(T, x_1)) = \mathrm{Prob}(F_1(T) > x_1 | \text{Full smile}) \tag{13}$$

$$\mathrm{Prob}(F_2(T) > \tilde{x}_2 | \sigma_2 = \Sigma_2(T, x_2)) = \mathrm{Prob}(F_2(T) > x_2 | \text{Full smile}) \tag{14}$$

$\Sigma_1(T, x_1)$ denotes the implied volatility of $F_1(T)$ at strike x_1 and $\Sigma_2(T, x_2)$ the implied volatility of $F_2(T)$ at strike x_2.

This assumption means that we are using a Gaussian copula to represent the joint distribution of the random variables $F_1(T)$ and $F_2(T)$.

The following algorithm, which relies on the Gaussian copula assumption, can then be used to calculate the price of a spread option with smile as in (11).

Algorithm

- Calculate $\mathrm{Prob}(F_1(T) > x_1 | \text{Full smile})$, $i = 1, 2$, from the price of a call spread.
- Solve equations (13) and (14) for $\tilde{x}_1$ and $\tilde{x}_2$.
- Estimate ρ from historical data for $F_1(t)$ and $F_2(t)$.
- Calculate the joint distribution (normal bivariate) of $F_1(T)$ and $F_2(T)$ using (12).

4.3 Extension to CMS spread options

Introduction If we want to use the model we have proposed above, we need the smile surface for each underlying. This smile is more or less known in the market when the underlying is the short rate (Libor 1M, ..., 12M). But, when the underlying is CMS, the smile is unknown. One idea is to use the swaption smile with the swap maturity equal to the tenor of this CMS. Unfortunately this strategy is not arbitrage free—in theory—particularly when the CMS cap/floor and swap are liquid. In the last part of this section, we propose an idea to build this smile using the prices of CMS caps/floors and swaps. To introduce this idea, first we present the issues involved in pricing CMS products, with a specific section about the timing adjustment necessary for CMS products with fixings in advance. Then we expose a simple approach, widely used in banks, to price CMS swaps and caps/floors using the whole smile of swaptions. This approach is based on a simple idea of replication, which can be used for any complex European payoff.

Issues in pricing CMS products Let us denote SR_t the swap rate at time t. Its value at time t is:

$$SR_t = \frac{B(t, T_0) - B(t, T_N)}{\sum_{i=1}^{N} B(t, T_i)\tau_i}$$

The swap starts at time T_0 and its payments occur at times $T_i (i = 1, \ldots, N)$ with $T = T_0 < T_1 < \ldots < T_N$.

$B(t, T_i)$ is the price at time t of the bond which pays 1 unit at time T_i.

$\tau_i = \frac{T_i - T_{i-1}}{365}$ if SR_t is expressed in basis Act/365.

SR_t is then a martingale (i.e. a driftless process) under the numeraire SM_T defined as:

$$SM_T = \sum_{i=1}^{N} B(T, T_i)\tau_i$$

Prices of FRAs and caplets are given by:

$$\text{FRA}(t) = B(t, T) \cdot E_{Q_T}[SR_T | F_t]$$
$$\text{Caplet}(t) = B(t, T) \cdot E_{Q_T}[\text{Max}(SR_T - K; 0) | F_t]$$

Q_T denotes the T forward measure. Under this measure, SR_t is not driftless and it is difficult to calculate its drift.

The price of a physical swaption is given by:

$$\text{Swaption}(t) = E_{SM_T}[\text{Max}(SR_T - K, 0) | F_t] \cdot \sum_{i=1}^{N} B(t, T_i)\tau_i$$

where E_{SM_T} denotes the expectation with respect to numeraire SM_T.

We can apply the Black formula in this case, because SR_t is driftless.

From this short analysis, we can see that if we can express the payoff of FRAs/caps in terms of the payoff of the swaption, then pricing becomes simple. It is the idea of the replication, which we develop now.

Note that in order to price a cash swaption, which is a more common product than physical swaptions, one has to use instead of the physical swap measure, the cash swap measure where the numeraire is:

$$\text{SCashM}_t = \sum_{i=0}^{N} \frac{1}{(1 + SR_T)^i}$$

Replication of simple products on CMS In this section we develop the idea of replicating the payoff of a CMS swap or cap as a linear combination of swaptions with different strikes. In addition to the mathematical argument of easy derivation given in the last section, another motivation for doing this is that the only simple and liquid way to hedge a product on CMS is to use swaptions.

We want to write a linear payoff (swap/cap/floor) of the form:

$$\text{Max}(SR_T - K; 0)$$

in terms of a non-linear payoff (swaption with cash settlement) of the form:

$$\text{Max}(SR_T - K; 0) \cdot \sum_{i=1}^{N} \frac{1}{(1 + SR_T)^i}$$

So the idea is to find a set of weights w_j and strikes Kj such that:

$$\text{Max}(SR_T - K; 0) = \sum_j w_j \text{Max}(SR_T - K_j; 0) \cdot \sum_{i=1}^{N} \frac{1}{(1 + SR_T)^i} \tag{15}$$

We choose the strikes Kj to be equally spaced, using a discretization step Δ. So we have:

$$K_j = K + j\Delta;\ j = 1, \ldots, M$$

In our experience, $\Delta = 5$ to 10 basis points is a good choice and M can be chosen so that K is about 15%, but it really depends to what limit of strike the trader wants to hedge its CMS products.

The calculus of the weight w_j is straightforward.

Timing adjustment for CMS products with fixings in advance We have seen that the replication technique is based on swaptions with cash settlement, so it can only be used to price CMS products in which the swap rate is observed and paid at the same time. When we deal with CMS products with fixings in advance, e.g. CMS vanilla caps/floors/swaps, the price has to be adjusted.

If the swap rate is observed at time T and paid at $T + \delta$, the forward swap rate SR_0 has to be corrected by a timing adjustment (see Hull 2002):

$$-\frac{SR_0 \delta R_0 \rho \sigma \sigma_R T}{1 + R_0 \delta}$$

where R_0 is the value at time zero of the forward rate between T and $T + \delta$, σ_R is the volatility of this forward rate, σ is the at-the-money volatility of the forward swap rate and ρ is the correlation between the forward swap rate and the forward rate.

Example Let's take the example given by Hull (2002).

$$SR_0 = 5\%$$
$$R_0 = 5\%$$
$$\sigma = 15\%$$
$$\sigma_R = 20\%$$
$$\delta = 0.5$$
$$\rho = 0.7$$

The forward rate has to be adjusted by $-0.0000256T$.

Building CMS smile by arbitrage The process SR_T can be written under the T-forward measure as follows:

$$SR_T = E_{Q_T}[SR_T]\exp\left(-\frac{1}{2}\sigma^2 T + \sigma W_T\right) \tag{16}$$

The application of the replication technique for FRAs gives the expectation value $E_{Q_T}[SR_T]$ of SR_T under the T-forward measure as:

$$E_{Q_T}[SR_T] = \frac{FRA}{B(0, T)}$$

The price of the caplet/floorlet with strike K using the expression (16) of the process SR_T is simply given by Black's formula:

$$\text{Caplet} = \text{Black }(E_{Q^T}[SR_T], \sigma(K), T, K)$$

The unknown variable in this formula is the volatility $\sigma(K)$. At the same time this price can be obtained using the replication technique described above. Hence we can imply the volatility $\sigma(K)$ by:

$$\sigma(K) = \text{Black}^{-1}(\text{Caplet})$$

We can apply this technique for every strike K and thus we build the CMS smile.

We admit that it can be time consuming. At the first approximation we can take the swaption smile.

5 Tests

5.1 Introduction

We first show the difference in price for short rate spread options (Libor 6M–Libor 3M), for given market data: yield curve and smile, with the three methods:

- The approach with taking at-the-money volatility for each index.
- The same approach but with taking as strike for one index, the money for the second index plus/minus the strike of the spread option.
- Pricing with full smile as described in this chapter.

Then we do similar tests on CMS products.

In all our tests, we use the following features:

- Payment frequency: 6M
- Day count: ACT/360
- Yield curve:

	ATM swap rate
1Y	1.14%
2Y	1.55%
5Y	2.63%
7Y	3.12%
10Y	3.61%
15Y	4.12%
20Y	4.36%

- Volatility surface:

	3%	4%	5%	6%	7%	8%	9%
1Y	26.50	21.90	23.30	25.50	26.70	27.80	29.10
2Y	25.80	20.30	19.10	21.30	22.70	23.80	25.20
5Y	23.90	19.50	15.50	15.00	15.60	16.40	17.60
7Y	22.70	18.70	14.80	13.30	13.40	14.00	14.70
10Y	21.50	17.90	14.10	12.10	12.10	12.60	13.10
15Y	20.20	17.00	13.50	11.20	10.90	11.30	11.60
20Y	19.30	16.40	13.10	10.60	10.30	10.70	11.00

5.2 Short rate spread option

We consider a spread option Libor 6M−Libor 3M. First, we consider strike zero and volatility flat. We compare the Margrabe closed-form formula (without smile), the Monte Carlo approach (partial smile), and the full smile method.

$$\text{Libor 6M} = 0.99\%$$

$$\text{Libor 3M} = 0.95\%$$

From Table 1, we check that our model gives the same results as Margrabe's formula in the case where the volatility is flat, for different maturities.

The small difference can be due to the numeric integration method used in our implementation.

Now, we consider the same option but with smiled volatility. We notice that the difference becomes significant when the maturity increases.

TABLE 1: STRIKE = 0, VOLATILITY IS FLAT AT 20%, CORRELATION = 0.7

	Margrabe formula	MC method (10 000 path)	Full smile
1Y	9	9	9
2Y	32	32	32
5Y	153	153	153
7Y	275	275	275
10Y	484	483	483
15Y	854	852	853
20Y	1181	1177	1176

TABLE 2: STRIKE = 0, VOLATILITY WITH SMILE, CORRELATION = 0.7

	Margrabe	Full smile
1Y	12	12
2Y	42	41
5Y	176	170
7Y	278	275
10Y	436	436
15Y	665	690
20Y	860	901

TABLE 3: STRIKE = 0.20%, VOLATILITY WITH SMILE, CORRELATION = 0.7

	Margrabe	Our approach (in basis points)
1Y	4	4
2Y	24	23
5Y	131	126
7Y	217	216
10Y	347	356
15Y	552	583
20Y	725	773

5.3 Building CMS smile surface

We compare the swaption volatility smile (for 10Y fixed swap maturity) with the CMS 10Y, after building the CMS smile as described in this chapter.

In general the CMS smile is less than the swaption smile with the same swap maturity.

TABLE 4: SMILE CMS 10Y (VOLCMS10Y–VOLSWAPTION NX10Y)

	3%	4%	5%	6%	7%	8%	9%
1Y	−0.2	0.1	0.3	0.3	0.4	0.5	0.6
2Y	−0.4	−0.1	0.3	0.5	0.6	0.7	0.8
5Y	−0.8	−0.5	−0.2	0.1	0.4	0.5	0.6
7Y	−0.9	−0.7	−0.4	−0.1	0.1	0.2	0.3
10Y	−0.9	−0.8	−0.5	−0.2	0.1	0.2	0.2
20Y	−0.9	−0.9	−0.7	−0.4	−0.1	−0.1	0.1

5.4 Impact of smile in CMS spread option

We consider the spread option on CMS 20Y and CMS 2Y with strike equal to 2.5% (in but not far from the money).

The first column gives the price of the spread option priced with the volatility at-the-money for each index and the second column with partial smile. The third column shows the price with full smile. In the first two columns, the price is calculated with Monte Carlo.

TABLE 5: SPREAD OPTION ON CMS 20Y AND CMS 2Y WITH STRIKE = 2.50% AND CORRELATION = 0.7

	Vol at-the-money	Partial smile	Full smile
1Y	47	46	47
2Y	79	80	84
5Y	139	155	166
7Y	170	194	210
10Y	211	250	269
15Y	273	334	351
20Y	339	415	428

It is clear from Table 5 that the model with partial smile is closer to the full smile model than the classical approach without smile.

Those differences depend on

- the convexity of the smile
- how far the strike of the spread option is from the money.

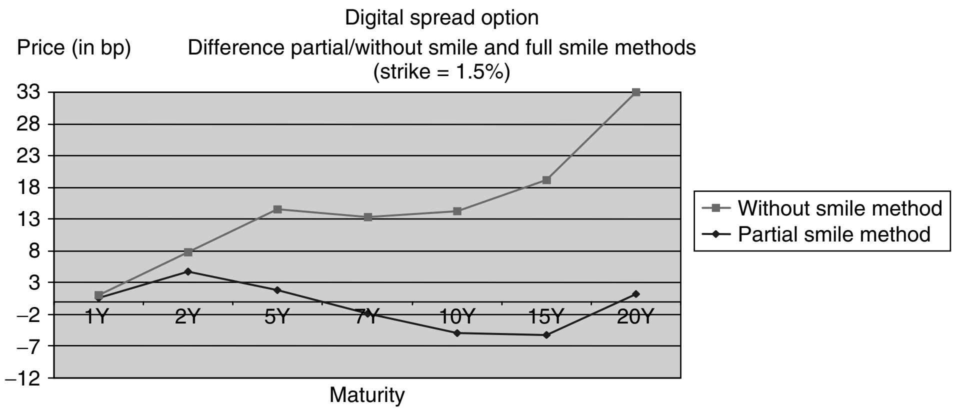
5.5 Impact of smile in digital CMS

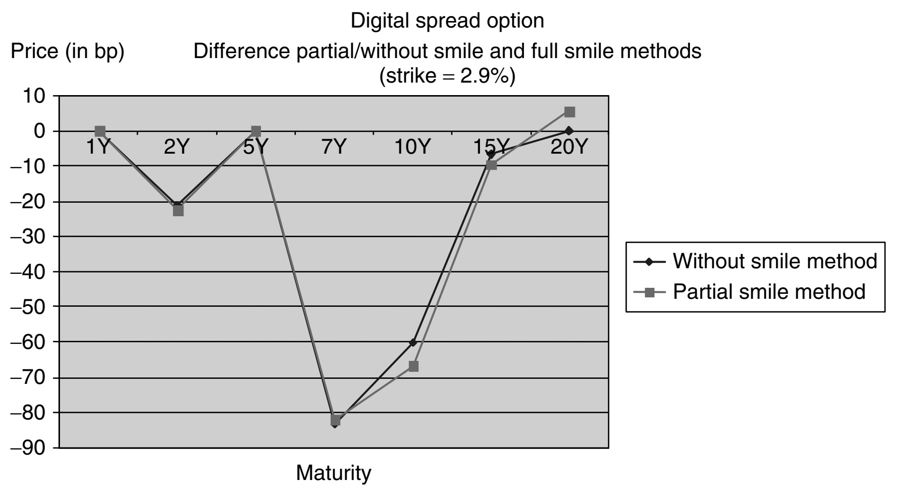
We approximate a digital option as a call spread with a strike shift equal to 10 basis points. We compare the same three models again.

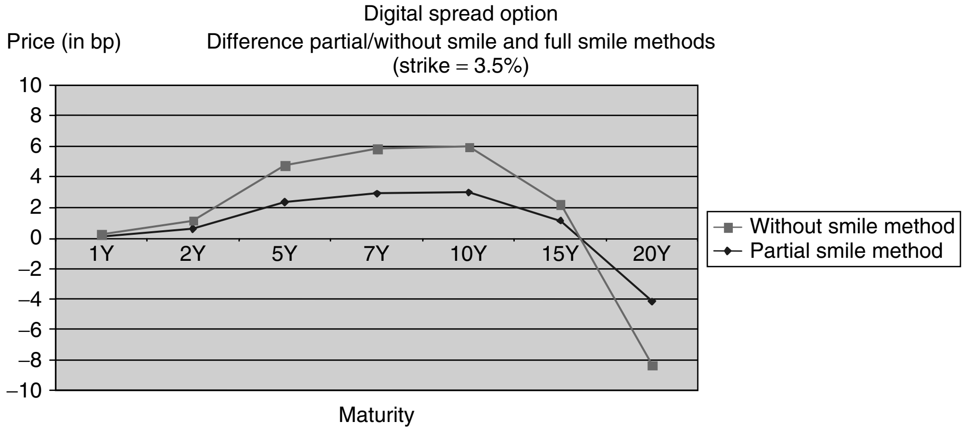
The graphs below show the differences between the prices with the different models as presented in the tables.

TABLE 6: CALL DIGITAL OPTION ON CMS 20Y AND CMS 2Y WITH STRIKE = 1.50% AND CORRELATION = 0.7

	Vol at-the-money	Partial smile	Full smile
1Y	98	98	98
2Y	179	180	176
5Y	320	309	308
7Y	378	361	363
10Y	441	417	421
15Y	508	478	484
20Y	559	528	527







These graphs show that taking the volatility at the right strike (partial smile) gives closer prices to the full smile method especially when the option is deeply in- or-at-the-money.

For digital option at-the-money, the differences between the two models, however, are significant.

6 Conclusions

In this chapter we have exposed two new methods to take into account the smile for spread options and in particular digital spread options.

The most advanced of those two methods is a numerical integration method based on a copula assumption, which uses the entire smile of each underlying.

If the smile is not smooth enough, this method can lead to instabilities. This is why, when this situation occurs, a parameterization of the smile and then using a closed-form formula for $\text{Prob}(F_{i,t} > x_i; \text{smile}(F_{i,t}))$ could be a worthwhile alternative. For a digital option, in this case one needs to consider:

$$\left.\frac{dC}{dK}\right|_{K=K_0} = \left.\frac{\partial C}{\partial K}\right|_{K=K_0} + \left.\frac{\partial C}{\partial \sigma}\right|_{K=K_0} * \left.\frac{d\sigma}{dK}\right|_{K=K_0} \tag{17}$$

where $C(K, \sigma(K))$ is the call option price and $\sigma(K)$ is a parametric volatility function (Example: SABR model).

Another method which we propose is to price spread options taking the volatility at a different strike than the money of each underlying the same time, as follows:

$$Vol(F_1) = Vol(Strike = ATM(F_2) + K)$$
$$Vol(F_2) = Vol(Strike = ATM(F_1) - K)$$

This method is only a partial smile model but we show that it is close to the first, full smile, method.

A separate section of this chapter is dedicated to dealing with CMS underlyings and building the CMS smile.

REFERENCES

■ Alexander, C. and Scourse, A. (2003) Bivariate normal mixture spread option valuation. *ISMA Centre Discussion Papers in Finance*, 2003–15.

■ Carmona, R. and Durrleman, V. (2003) Pricing and hedging spread options in a log-normal model. Working paper.

■ Carmona, R. and Durrleman, V. (2003) Pricing and hedging spread options. *SIAM Review*, 45(4): 627–687.

■ Carr, P. and Madan, D. B. (1999) Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2(4): 61–73

■ Cherubini, U. and Luciano, E. (2002) Multivariate option pricing with copulas. Working paper.

■ Coutant, S., Durrleman, V., Rapuch, G. and Roncalli, T. (2001) Copulas, multivariate risk-neutral distributions and implied dependence functions. Working paper.

■ Dempster, M. and Hong, G. (2001) Pricing spread options with the fast Fourier transform. *Risk*, Europe.

■ Hull, J. C. (2002) *Options, Futures and Other Derivatives*. Prentice Hall.

■ Rapuch, G. and Roncalli, T. (2001) Some remarks on two-asset options pricing and stochastic dependence of asset prices. Working paper.

■ Ravindran, K. (1993) Low fat spread. *Risk Magazine*, 6(10): 66–7.

■ Rosenberg, J. (2003) *Non-parametric Pricing of Multivariate Contingent Claims*. Research and Market Analysis Group, Federal Reserve Bank of New York.

■ Shimko, D. (1994) Option on futures spreads: hedging, speculation and valuation. *Journal of Futures Markets* 14(2): 182–213.

■ Wilmott, P. (1998) *Derivatives: The Theory and Practice of Financial Engineering*. John Wiley & Sons, Ltd.

18
The Case for Time Homogeneity

Philippe Henrotte

Departure from time homogeneity may be the sign of serious modelling deficiency. We show with three important examples that it is possible to calibrate parsimonious time homogeneous models to complex term structures. Our examples include the volatility smile, the credit spread, and the yield curve.

1 Introduction

We explore a simple yet significant modelling issue in finance. In many situations where market prices display a term structure, it seems natural to resort to some time dependent dynamics if one wishes to calibrate a model to the observed market data. We argue that this is almost always a bad idea, a sign that some important underlying stochastic structure has been missed at the modelling stage.

When a simple model fails to capture some economically meaningful pattern, tweaking a few parameters through time is a dangerous way of getting extra mileage out of an exhausted solution, even if this adjustment yields an excellent calibration. For calibration alone should not measure the quality of a model. Adjusting a few parameters through time for the sake of calibration alone almost always implies crazy future scenarios, which, although not theoretically impossible, nevertheless look often extremely awkward. As a result, tweaked models typically lack robustness and time consistency.

Stability can only be achieved once the salient features of the dynamics of the problem are correctly captured, and this implies in turn a careful description of the underlying state variables. Achieving a good calibration with a time homogenous model is a powerful sign that the stochastic structure of the problem has been correctly formulated. The term structure that we wish to calibrate, like the motion of planets in space, is a complex function of time which may be described by many different time inhomogeneous ad hoc theories. A time homogeneous model in finance resembles the law of gravity in physics. It yields a parsimonious explanation where time does not play

Contact address: HEC School of Management, ITO33 SA, 39 rue L'homond, 75005 Paris, France.
E-mail: NewNumberTwo@ito33.com

a direct role. This feat is achieved at the cost of enlarging the state space, by considering for instance speed and acceleration as additional state variables on top of the position in space.

Increasing the dimension of the state space may prove fatal for the numerical tractability of the model. The brute force solution which consists for instance of replacing every time dependent parameter by a general time homogeneous stochastic process is probably doomed to fail. We search instead for a parsimonious solution, the smallest possible state space on which a time homogeneous dynamics can be written with good calibration properties. It would be foolish to push the analogy with physics too far and claim that we would then have discovered some universal law for finance similar to gravitation. Our goal is merely to seek robustness and stability under the constraint of numerical tractability. The objective of this chapter is to point out that this research agenda deserves serious consideration.

We show that in many situations, the increase in the complexity of the state space may be limited to the addition of an abstract regime variable which only assumes a small number of states. We investigate three financial environments where the analysis of a term structure is of the essence: the implied volatility smile, the term structure of credit spread, and the yield curve. In each case we obtain encouraging calibration results, and the added difficulty of working with a larger state space is more than offset by the benefits brought by time homogeneity.

2 The implied volatility smile

The implied volatility schedule of at-the-money calls as a function of maturity is a first important example of term structure in finance. It is well known that a simple tweak to the standard time homogeneous Black–Scholes model will do the job: by allowing the volatility parameter to be a function of time, any term structure can be recovered. If one wishes to fit an entire smile schedule across maturity and strike price, this trick can be extended to a so-called local volatility by letting the volatility be a function of time and spot price. Anyone who ventures down this avenue knows that the journey ends in a bitter numerical fiasco. The seemingly natural extension is in fact all but natural. It lacks robustness, yields chaotic predictions for future smile patterns, and generates hedges and prices for exotic instruments way out of line with market practices. One could hardly paint a gloomier picture.

The good and somewhat surprising news is that one need not introduce a very sophisticated state space in order to recover time homogeneity. Tables 1, 2 and 3 show that a few regimes with a simple time homogeneous Markov structure are enough to capture the jumps and the stochastic volatility needed to calibrate not only to an entire vanilla option smile schedule, but also to some key liquid exotic instruments such as digital or forward start options.[1] Whereas the vanilla option prices are used for the implied volatility smile calibration, a few liquid exotic instruments help capture the dynamics of the smile. The simple tweak to the Black–Scholes volatility fails so miserably because it cannot capture the smile dynamics, as reflected in the prices of the exotic instruments.

The bad news is that by extending, even a little, the state space, the markets are no longer complete. This means that the perfect delta hedge, the cornerstone of the Black and Scholes analysis, is lost and the heavy machinery of incomplete markets must be brought to bear if one is to derive meaningful dynamic hedging strategies.

TABLE 1: CALIBRATED PARAMETERS OF THE REGIME-SWITCHING MODEL (3 REGIMES)

	Brownian diffusion	Total volatility
Regime 1	9.57%	11.67%
Regime 2	6.24%	32.23%
Regime 3	2.25%	11.88%

	Jump size	Jump intensity
Regime 1 → Regime 2	−9.07%	0.2370
Regime 2 → Regime 1	62.67%	0.0855
Regime 1 → Regime 3	2.72%	3.3951
Regime 3 → Regime 1	−3.17%	2.9777
Regime 2 → Regime 3	24.63%	1.0944
Regime 3 → Regime 2	−22.66%	0.2040

3 The term structure of credit spread

A second example of term structure is the schedule of credit spread of an issuer as a function of maturity. This topic is attracting a lot of attention today with the development of the equity to credit paradigm. Insurance instruments such as credit default swaps are becoming liquid for maturities up to five or ten years. In reduced form models, the term structure of credit spreads is often captured by a default intensity parameter which is assumed to be a function of time and spot. One immediately sees the parallel with the local volatility. Tweaking the default intensity does the job and yields simple numerical procedures. But this is achieved at the cost of hiding the stochastic structure of the default process. The term structure contains some key information about this structure which is revealed in a time homogeneous framework with a few constant parameters.

Calibrating a slightly more complex model with constant parameters reveals far more of the underlying stochastic nature of the problem than resorting to a seemingly simpler model with fewer parameters which must be tweaked every period. Tables 4, 5 and Figure 1 show that a simple model with two or three regimes and a time homogeneous Markov structure captures quite nicely most credit spread patterns, even for relatively long maturities.

4 The yield curve

A third obvious example of term structure in finance is the yield curve. Two major modelling schools have emerged, which differ in the way they describe the state variable. One school lets the state variable be the short-term interest rate while the other one uses the entire yield curve.

TABLE 2: QUALITY OF FIT OF A FULL IMPLIED VOLATILITY SURFACE WITH THE REGIME-SWITCHING MODEL

	Strike										
Maturity (years)	80	85	90	95	100	105	110	115	120	130	140
	Market	19.00%	16.80%	13.30%	11.30%	10.20%	9.70%				
0.18	**Model**	19.22%	16.38%	13.35%	11.69%	10.38%	10.29%				
	Market	17.70%	15.50%	13.80%	12.50%	10.90%	10.30%	10.00%	11.40%		
0.43	**Model**	17.56%	15.85%	13.97%	12.43%	11.14%	10.08%	10.07%	11.53%		
	Market	17.20%	15.70%	14.40%	13.30%	11.80%	10.40%	10.00%	10.10%		
0.70	**Model**	17.34%	15.90%	14.37%	13.00%	11.85%	10.87%	10.11%	10.20%		
	Market	17.10%	15.90%	14.90%	13.70%	12.70%	11.30%	10.60%	10.30%	10.00%	
0.94	**Model**	17.22%	15.93%	14.60%	13.39%	12.36%	11.47%	10.69%	10.23%	11.04%	
	Market	17.10%	15.90%	15.00%	13.80%	12.80%	11.50%	10.70%	10.30%	9.90%	
1.00	**Model**	17.19%	15.93%	14.65%	13.48%	12.46%	11.60%	10.83%	10.32%	10.86%	
	Market	16.90%	16.00%	15.10%	14.20%	13.30%	12.40%	11.90%	11.30%	10.70%	10.20%
1.50	**Model**	16.99%	15.98%	14.97%	14.03%	13.19%	12.46%	11.80%	11.24%	10.56%	10.89%
	Market	16.90%	16.10%	15.30%	14.50%	13.70%	13.00%	12.60%	11.90%	11.50%	11.10%
2.00	**Model**	16.87%	16.03%	15.20%	14.42%	13.71%	13.07%	12.48%	11.98%	11.17%	10.76%
	Market	16.80%	16.10%	15.50%	14.90%	14.30%	13.70%	13.30%	12.80%	12.40%	12.30%
3.00	**Model**	16.74%	16.12%	15.52%	14.94%	14.40%	13.89%	13.42%	12.99%	12.26%	11.67%
	Market	16.80%	16.20%	15.70%	15.20%	14.80%	14.30%	13.90%	13.50%	13.00%	12.80%
4.00	**Model**	16.68%	16.19%	15.72%	15.26%	14.83%	14.42%	14.03%	13.67%	13.03%	12.48%
	Market	16.80%	16.40%	15.90%	15.40%	15.10%	14.80%	14.40%	14.00%	13.60%	13.20%
5.00	**Model**	16.63%	16.24%	15.85%	15.48%	15.12%	14.78%	14.45%	14.14%	13.58%	13.09%

Source: S&P500 index in October 1995

TABLE 3: QUALITY OF FIT OF THE ONE-TOUCH PRICE STRUCTURE

Maturity (years)		One-touches									
		−5%	−10%	−20%	−30%	−50%	50%	30%	20%	10%	5%
0.175	**Market**	0.58%	−1.16%	−3.70%	−5.27%	−6.38%	−6.14%	−7.91%	−8,35%	−6.30%	−3.70%
	Model	0.60%	−1.16%	−3.70%	−5.27%	−6.38%	−6.17%	−7.90%	−8.36%	−6.28%	−3.67%
1.5	**Market**	7.17%	6.26%	2.51%	−1.67%	−8.25%	−3.13%	−6.92%	−8.22%	−6.91%	−4.09%
	Model	7.17%	6.26%	2.51%	−1.66%	−8.24%	−3.09%	−6.91%	−8.24%	−6.92%	−4.08%
5	**Market**	8.10%	8.70%	7.47%	5.06%	−0.95%	−0.09%	−2.78%	−4.25%	−4.53%	−3.30%
	Model	8.09%	8.69%	7.45%	5.04%	−0.97%	−0.08%	−2.76%	−4.23%	−4.52%	−3.29%

TABLE 4: CALIBRATED PARAMETERS OF A TIME-HOMOGENEOUS REGIME-SWITCHING MODEL (TWO REGIMES ONLY)

	Hazard rate
Regime 1	0.15%
Regime 2	7.15%

Jump intensity	
Regime 1 → Regime 2	0.7400
Regime 2 → Regime 1	0.1270

TABLE 5: QUALITY OF FIT OF THE TERM STRUCTURE OF SPREADS OF CREDIT DEFAULT SWAPS WITH TWO REGIMES IN A REGIME-SWITCHING MODEL

Maturity (years)	Recovery rate	Market	Model
1	0.45	1.08%	1.16%
2	0.45	1.72%	1.78%
3	0.45	2.10%	2.14%
5	0.45	2.65%	2.53%
7	0.45	2.73%	2.72%
10	0.45	2.79%	2.86%
15	0.45	3.00%	2.96%

Source: General Motors 30/09/2003

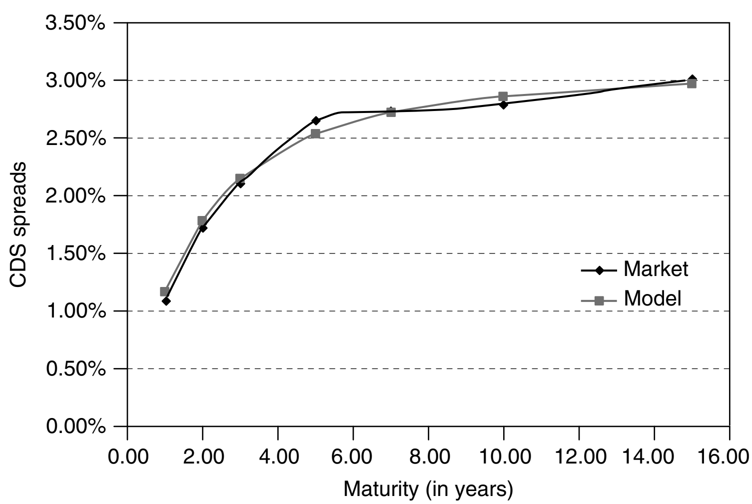


Figure 1: Quality of fit of the term structure of spreads of credit default swaps

The ability to fit a given initial yield curve is a major modelling requirement. For the short-term interest rate school, this is achieved by arm twisting the parameters of the short-term rate process through time so as to generate the desired yield curve. The second school avoids such painful contortion since the yield curve is viewed as an input, a parameter of the model which need not be calibrated. The main drawback here is that any information on the stochastic structure of the problem which may be contained in the shape of the yield curve is lost.

For both schools, producing a simple time homogeneous model of the yield curve seems a remote and lost cause. This is a very unfortunate outcome, probably dictated by a more sombre agenda: the need to produce quasi closed form pricing solutions, or at least elementary numerical procedures such as one-dimensional trees.

It is instructing to realize that a very simple time homogeneous process with no more than three abstract regimes can fit reasonably well almost any yield curve together with the prices of a few interest rate derivatives (see Tables 6, 7, 8, 9 and Figures 2 and 3). Such a model must be solved numerically, but the state variable is so parsimonious that calibration need not be a nightmare.

TABLE 6: CALIBRATED PARAMETERS OF A TIME HOMOGENEOUS REGIME-SWITCHING MODEL (3 REGIMES) NOVEMBER 1995

	Short rate
Regime 1	5.417%
Regime 2	10.930%
Regime 3	2.626%

	Jump intensity
Regime 1 → Regime 2	0.0402
Regime 2 → Regime 1	0.0783
Regime 1 → Regime 3	0.1903
Regime 3 → Regime 1	0.1005
Regime 2 → Regime 3	0.1574
Regime 3 → Regime 2	0.2615

TABLE 7: QUALITY OF FIT OF THE YIELD CURVE USING THREE REGIMES IN A REGIME-SWITCHING MODEL

Maturity (years)	Market	Model
0.25	5.410%	5.383%
0.5	5.333%	5.357%
1	5.311%	5.324%
2	5.322%	5.316%
5	5.495%	5.486%
10	5.798%	5.802%

Source: US Government zero coupon yield curves, November 1995

TABLE 8: CALIBRATED PARAMETERS OF A TIME HOMOGENEOUS REGIME-SWITCHING MODEL (3 REGIMES) OCTOBER 1978

	Short rate
Regime 1	7.388%
Regime 2	0.400%
Regime 3	22.753%

	Jump intensity
Regime 1 → Regime 2	0.6996
Regime 2 → Regime 1	0.5556
Regime 1 → Regime 3	1.5346
Regime 3 → Regime 1	0.4503
Regime 2 → Regime 3	0.7516
Regime 3 → Regime 2	1.6144

TABLE 9: QUALITY OF FIT OF THE YIELD CURVE USING THREE REGIMES IN A REGIME-SWITCHING MODEL

Maturity (years)	Market	Model
0.25	8.937%	8.933%
0.5	9.503%	9.513%
1	9.657%	9.640%
2	9.246%	9.261%
5	8.826%	8.819%
10	8.662%	8.664%

Source: US Government zero coupon yield curves, October 1978

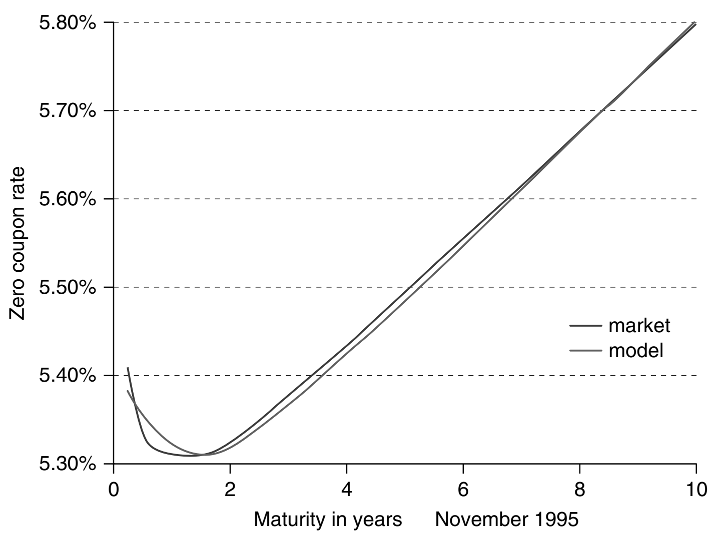


Figure 2: Quality of fit of the yield curve using three regimes in a regime-switching model. *Source*: US Government zero coupon yield curves, November 1995

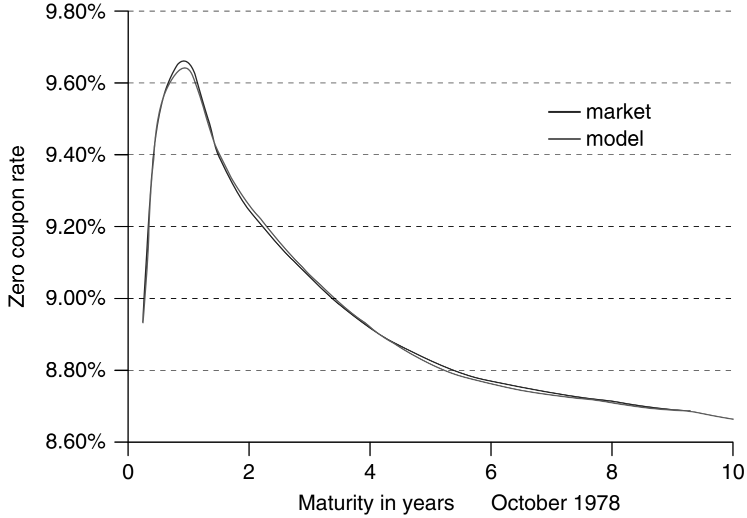


Figure 3: Quality of fit of the yield curve using three regimes in a regime-switching model. ***Source*****: US Government zero coupon yield curves, October 1978**

5 Conclusion

We have made the case for parsimonious time homogeneous models as a powerful way to decipher the stochastic structure underlying a complex collection of market data. In some instances, an event announced for a specific date will destroy the time homogeneity and there are situations where time should indeed be considered as a state variable after all. These cases should be treated as exceptions and not as the rule. We conclude with a simple sanity check for a financial model: any departure from time homogeneity should be the cause of great concern and should therefore be strongly motivated, lest it is the sign of some serious modelling deficiency.

FOOTNOTE

1. See E. Ayache, P. Henrotte, S. Nassar, and X. Wang. Can anyone olve the smile problem? *Wilmott*, January 2004.

19
Hybrid Stochastic Volatility Calibration

Domingo Tavella,[*] Alexander Giese[**] and Didier Vermeiren[**]

We present a hybrid stochastic volatility model which improves the calibration to spot implied volatilities over a wide range of maturities and strikes, while at the same time preserving the desirable properties of the purely stochastic volatility model. This hybrid stochastic volatility model is obtained by superposing a (small) local volatility component to a (dominant) stochastic volatility component. We illustrate this approach by combining the constant parameter Heston model with a parametric local volatility model. Results based on realistic market data indicate that this combination effectively extends the ability of the Heston model to calibrate to a larger range of maturities and strikes.

1 Introduction

A hybrid stochastic volatility model consists of a combination of a stochastic volatility model with a local volatility component. We can construct a hybrid model by choosing an appropriate stochastic volatility process, such as in the Heston model, and creating the instantaneous hybrid volatility as a weighted sum of the spot-independent stochastic volatility and of a spot-dependent local volatility. Depending on the proportions of the stochastic and local volatility components, the properties of a hybrid model will be a compromise between the properties of a purely stochastic volatility model and those of a purely local volatility model.

For an asset process given by:

$$\frac{dS}{S} = r\,dt + \sigma(S,t)\,dW_1 \tag{1}$$

Contact address: [*]Octanti Associates Inc., San Francisco, USA.
[**]HypoVereinsbank, Quantitative Research & Structuring, Arabellastr. 12, D-81925 Munich.

we define the instantaneous hybrid volatility $\sigma(S, t)$ as the superposition of stochastic and local components, σ_{SV} and σ_{LV}, respectively:

$$\sigma(S, t) = \sigma_{SV}(t)\eta + \sigma_{LV}(S, t)(1 - \eta) \tag{2}$$

where η denotes the fraction of the stochastic volatility component in the instantaneous hybrid volatility.

The stochastic volatility component follows a process given by:

$$d\sigma_{SV} = a(\sigma_{SV}, t)\, dt + b(\sigma_{SV}, t)\, dW_2 \tag{3}$$

where W_1 and W_2 are correlated standard Wiener processes.

The local volatility component is a suitably parameterized function of the asset spot and time:

$$\sigma_{LV} = f(S, t; \lambda_1(t), \ldots, \lambda_n(t)) \tag{4}$$

where the λ_i are time-dependent parameters.

Why would we want to consider such a hybrid volatility model? To appreciate this, we first summarize the main features and limitations of both stochastic and local volatility models.

Local volatility models are able to fit perfectly (at least in theory, Dupire 1994) a given implied volatility surface over its full range of strikes and maturities. However, its main limitations arise from the fact that, as maturity increases, forward starting smiles flatten out as a function of maturity (Andersen and Andreasen 1999). This behavior does not correspond to the market rule of thumb that the forward smile is more or less stationary in time.

This effect of maturity on the forward smiles is improved using stochastic volatility models. There are, however, practical limitations about the type of process that stochastic volatility models can follow. Practicality of computation dictates that acceptable candidates for stochastic volatility models should be Markovian. If this is not the case both analytical tools and effective computational techniques, such as finite differences, are no longer applicable to the calibration process.

Stochastic volatility models with constant parameters lack the time scales necessary to accommodate the change in the time dimension observed in market implied volatility surfaces. This means that stochastic volatility models with constant parameters can in general be calibrated to medium- and long-term maturities, but encounter problems fitting the short-term implied volatility smile.

One way to address this issue is to construct a hybrid stochastic volatility model by combining a standard stochastic volatility model with a local volatility component. This has been done in a multiplicative manner by Blacher (2001) and by Lipton (2002). The instantaneous hybrid volatility in this case takes the following form:

$$\sigma(S, t) = \sigma_{SV}(t)\sigma_{LV}(S, t) \tag{5}$$

However, with this particular form, it is not possible to separate the influence of the stochastic component from the local component in an intuitive manner. This is the reason why we prefer to define the instantaneous hybrid volatility as a weighted sum of a stochastic component and a local component:

$$\sigma(S,t) = \sigma_{SV}(t)\eta + \sigma_{LV}(S,t)(1-\eta) \tag{6}$$

Such a hybrid model is particularly interesting when a small local volatility component is sufficient to insure a proper fit of the market implied volatility surface for all maturities and strikes while the desirable properties of the dominant stochastic component are preserved.

In what follows we take the Heston model, one of the most popular stochastic volatility models, as the basis for constructing a hybrid model (Heston 1993).

2 Model framework

The hybrid calibration requires a very efficient solution for pricing vanilla call options. We achieved this through the numerical solution via finite differences of the two-dimensional Fokker–Plank equation (FPE) for the joint probability density of the asset and volatility states. Using the numerically computed solutions of the FPE, it is straightforward to compute the value of vanilla calls and, through the use of an optimizer, adjust the parameters of the model until a sufficiently good fit is obtained.

In the case of the Heston model combined with local volatility, it is much easier to solve the FPE numerically when the Heston process is expressed in terms of log variance. Our hybrid model is given by:

$$\frac{dS}{S} = r\,dt + [(1-\eta)\sigma_{LV} + \eta\sigma_{SV}]\,dW_1 \tag{7}$$

$$d\log\sigma_{SV}^2 = \frac{1}{\sigma_{SV}^2}\left(k(\theta - \sigma_{SV}^2) - \frac{1}{2}\phi^2\right)dt + \frac{\phi}{\sigma_{SV}}dW_2 \tag{8}$$

The second equation is the logarithmic transformation of the standard Heston model:

$$d\sigma_{SV}^2 = k(\theta - \sigma_{SV}^2)\,dt + \phi\sigma_{SV}\,dW_2 \tag{9}$$

The local volatility function, $\sigma_{LV}(S,t)$, is defined as follows:

$$\sigma_{LV}(S,t) = \lambda_0(t) + \lambda_1(t)\log\left(\frac{S}{S_0}\right) + \lambda_2(t)\log^2\left(\frac{S}{S_0}\right) \tag{10}$$

where S_0 is the current spot asset price, and $\lambda_0(t)$, $\lambda_1(t)$, and $\lambda_2(t)$ are piecewise linear continuous functions.

The joint probability density $p(S, \log\sigma_{SV}^2, t)$ of these processes is given by the two-dimensional forward FPE:

$$\frac{\partial p}{\partial t} = -\frac{\partial\mu_1 p}{\partial S} - \frac{\partial\mu_2 p}{\partial\log\sigma_{SV}^2} + \frac{1}{2}\frac{\partial^2 p}{\partial\log{\sigma_{SV}^2}^2} + \frac{\partial^2\rho\sigma_1\sigma_2 p}{\partial S\partial\log\sigma_{SV}^2} + \frac{1}{2}\frac{\partial^2\sigma_2^2 p}{\partial S^2} \tag{11}$$

where ρ is the correlation coefficient between W_1 and W_2, $\mu_1 = r$, $\mu_2 = \frac{1}{\sigma_{SV}^2}\left(k(\theta - \sigma_{SV}^2) - \frac{1}{2}\phi^2\right)$, $\sigma_1 = (1-\eta)\sigma_{LV} + \eta\sigma_{SV}$, and $\sigma_2 = \frac{\phi}{\sigma_{SV}}$.

We solve this equation numerically subject to the initial condition:

$$p(S, \log\sigma_{SV}^2, 0) = \delta(S - S_0, \log\sigma_{SV}^2 - \log\sigma_{SV_0}^2) \tag{12}$$

where σ_{SV_0} is the current spot stochastic volatility component.

3 Calibration considerations

The calibration strategy for the hybrid model consists of two stages. In the first stage, you select appropriate parameters for the basic stochastic volatility model. If you select these parameters such that the medium- and long-term maturity market smiles are captured as tightly as possible, the local volatility will be a thin layer superimposed to the basic stochastic volatility. The purpose of the local volatility layer is then simply to enable tighter calibration over the entire range of maturities, especially for short-term maturities.

In the second stage, the full FPE numerical solution is used to calibrate the local volatility component with respect to the functions $\lambda_0(t)$, $\lambda_1(t)$, and $\lambda_2(t)$, for all maturities.

The finite difference solution of the two-dimensional FPE is accomplished with an ADI scheme and requires careful attention to boundary conditions, aliasing, and oscillation issues.

Very accurate calibrations can be obtained with maturities of up to five years by using a high-resolution finite difference grid and carefully selecting resolution and computational parameters to avoid aliasing and oscillatory behavior.

In what follows, we selected the Heston parameters to be $k = 0.86, \theta = 0.03, \phi = 0.2$, and the correlation between spot and stochastic volatility returns equal to -0.5. We chose the fraction of the stochastic volatility component to be 90%.

Figure 1 shows the market implied volatility surface used in this case. It is fundamentally impossible to conduct a satisfactory Heston calibration to this entire surface. The hybrid model, however, allows for very tight calibrations over the full range of strikes and maturities included in the data.

Figure 2 compares the 0 to 0.5-year market spot smile with the implied volatility smiles generated by the hybrid and the purely local volatility model. Notice that there is a very close agreement between all three.

Figure 3 shows the 0.5 to 1-year forward smiles. Both models produce very similar results. This is consistent with the assumption that for short-term maturities, it is the time scales of the market data that determine the shape of the forward smiles, not the dynamics of the calibrated model. In other words, for short time horizons, if you were able to fully calibrate a stochastic volatility model you should expect the resulting forward smiles to be very close to the ones you would derive from a purely local volatility model.

Figure 4 shows the 1.5 to 2-year forward smile. We can see an incipient flattening of the smile as captured by the purely local volatility model compared with the hybrid model.

Figures 5 and 6 show the 4.0 to 4.5- and the 4.5 to 5-year smiles. For these maturities, we observe even flatter forward smiles generated by the purely local volatility model, and more

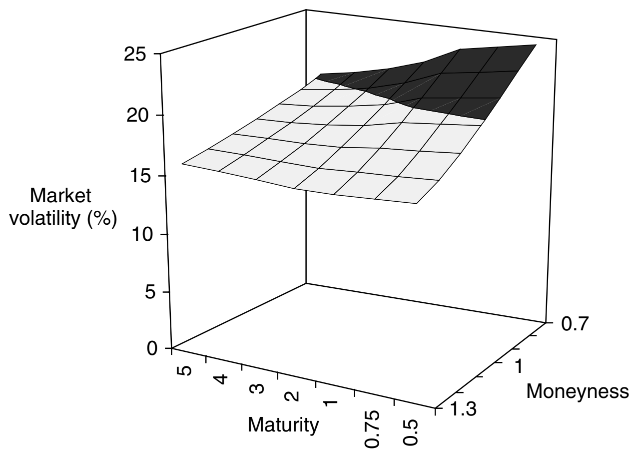


Figure 1: Market implied volatility, in percentage terms

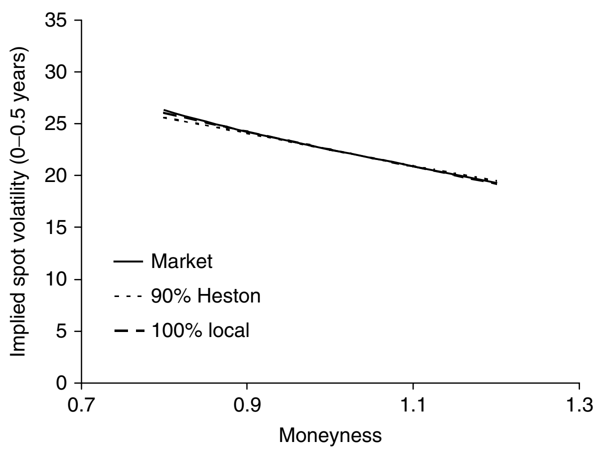


Figure 2: Spot market smile versus calibrated local and hybrid smiles

convex forward smiles generated by the hybrid model, in comparison with today's smile. More convex forward smiles are typical of purely stochastic volatility models (Bergomi 2004).

4 Conclusions

Calibration of constant parameter stochastic volatility models to a narrow window of a given market implied volatility surface is straightforward, but it is usually not possible to extend the calibration to the entire strike and maturity range in a satisfactory manner.

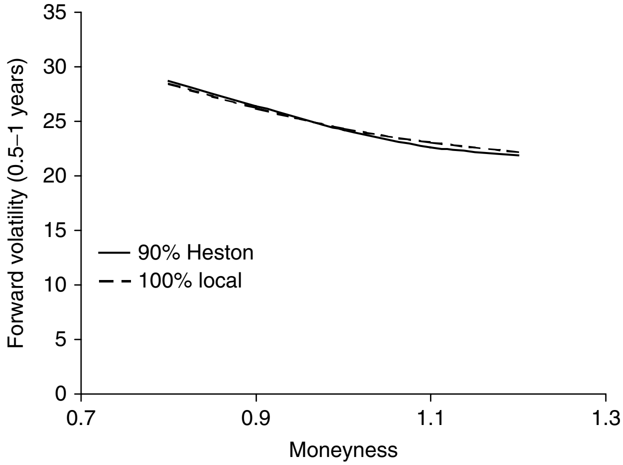


Figure 3: Forward 0.5 to 1-year smiles from hybrid and local volatility models

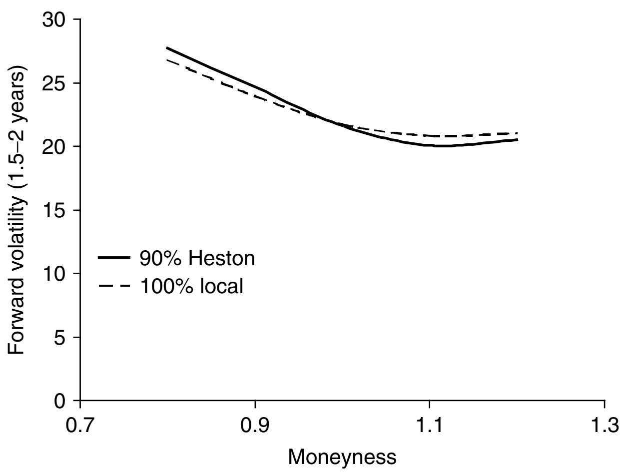


Figure 4: Forward 1.5 to 2-year smiles from hybrid and local volatility models

To address this issue, we proposed a hybrid model consisting of a superposition of a stochastic and a local volatility model. As an example of such a hybrid model, we selected a combination of the Heston model and a parametric local volatility model. The calibration errors obtained with the constructed hybrid model were significantly smaller than those obtained with a pure Heston model when calibrated to realistic market data. The analysis of forward smiles revealed that the hybrid model with a dominant stochastic component preserves the forward smile dynamics of a purely stochastic volatility model, in particular it does not induce a flattening of the forward smile as in a purely local volatility model.

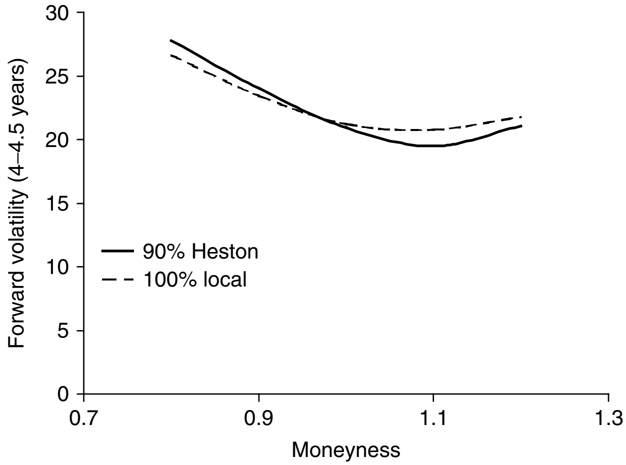


Figure 5: Forward 4 to 4.5-year smiles from hybrid and local volatility models

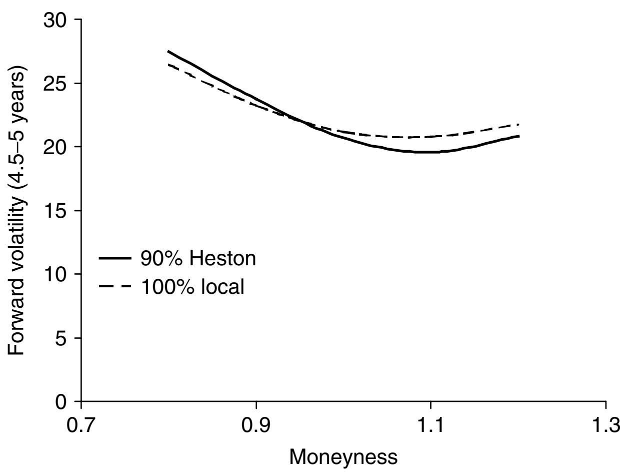


Figure 6: Forward 4.5 to 5-year smiles from hybrid and local volatility models

REFERENCES

- Andersen, L. and Andreasen, J. (1999) Jumping smiles. *Risk*, Nov., 12, 65–68.
- Bergomi, L. (2004) Smile dynamics. *Risk*, Sep., 17, 117–123.
- Blacher, G. (2001) A new approach for designing and calibrating stochastic volatility models for optimal delta-vega hedging of exotic options. Conference presentation at Global Derivatives, Juan-les-Pins.

■ Dupire, B. (1994) Pricing with a smile. *Risk*, Jan., 7, 18–20.
■ Heston, S. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6, 327–343.
■ Lipton, A. (2002) The vol smile problem. *Risk*, Feb., 15, 61–65.

20
Can Anyone Solve the Smile Problem?

Elie Ayache,* Philippe Henrotte, Sonia Nassar† and Xuewen Wang‡**

One of the most debated problems in the option smile literature today is the so-called 'smile dynamics'. It is the key both to the consistent pricing of exotic options and to the consistent hedging of all options, including the vanillas. Smiles models (e.g. local volatility, jump-diffusion, stochastic volatility, etc.) may agree on the vanilla prices and totally disagree on the exotic prices and the hedging strategies. Smile dynamics are heuristically classified as 'sticky-delta' at one extreme, and 'sticky-strike' at the other, and the classification of models follows accordingly. The real question this distinction is hinging upon, however, is space homogeneity vs inhomogeneity. Local volatility models are inhomogeneous. The simplest stochastic volatility models are homogeneous. To be able to control the smile dynamics in stochastic volatility models, some authors have reintroduced some degree of inhomogeneity, or even worse, have proposed 'mixtures' of models. We show that this is not indispensable and that spot homogeneous models can reproduce any given smile dynamics, provided a step is taken into incomplete markets and the true variable ruling smile dynamics is recognized. We conclude with a general reflection on the smile problem and whether it can be solved.

1 Introduction

The smile problem has raised immense interest among practitioners and academics. Since the market crash in October 1987, the volatilities implied by the market prices of traded vanillas have been varying with strike and maturity, revealing inconsistency with the Black–Scholes (1973) model which assumes a constant volatility. Ever since, a multitude of volatility smile models have been developed. The earliest of the volatility models were the local volatility models.[1] They

Contact address: ITO33 SA, 39 rue L'homond, 75005 Paris, France.
Email: *NumberSix@ito33.com, **NewNumberTwo@ito33.com, †Sonia@ito33.com, ‡Wang@ito33.com.

inferred a volatility dependent on the stock price level and time that accommodates the market price of vanillas within the Black–Scholes framework (Dupire 1994, Derman and Kani 1994, Rubinstein 1994). Indeed, local volatility models postulate that the underlying follows a lognormal diffusion process equation

$$\frac{dS}{S} = \pi(t)\,dt + \sigma\,(S, t)\;dW$$

yielding the following partial differential equation (PDE) for derivative instruments:

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

They are so to speak an extension of the Black–Scholes lognormal diffusion process with constant volatility to a process where the volatility is dependent on both the share price level and time. Under these assumptions, the unique local volatility surface is backed out through forward induction from the smile of vanilla option prices. Once the local volatility surface is known, it is used to value and hedge any type of option on the same underlying. The implied volatility of an option with a given strike and a given maturity can be seen as an average over all local volatilities that the underlying may have as time evolves until the maturity date. Local volatility models accommodate the smile and are theoretically self-consistent as it is possible to hedge and as a matter of fact perfectly replicate options in order to price them, as done in the Black–Scholes framework. In other words, they retain the market completeness.

Unfortunately, as shown in Figure 2, the shape of the local volatility surface, inferred from the market vanilla smile represented in Figure 1, may sometimes look very surprising and unintuitive, with no easily explainable trend either along the underlying share price direction or in the time

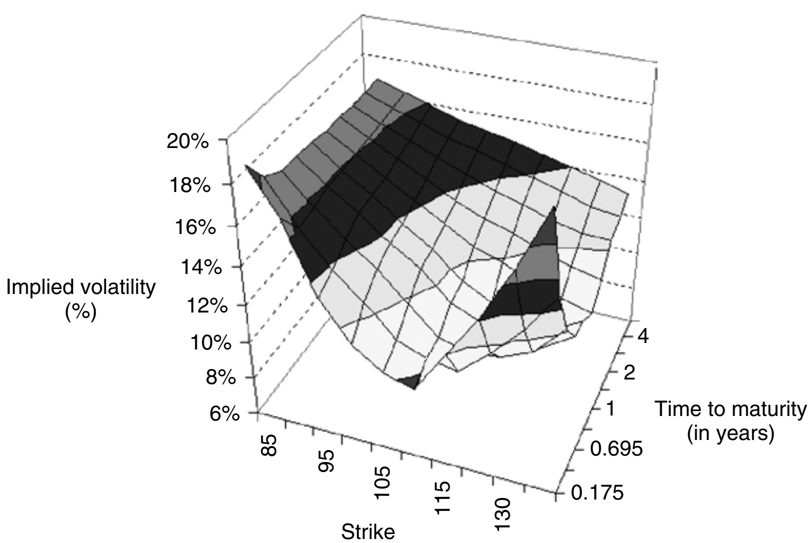


Figure 1: Implied volatility surface inferred from vanilla options market prices. ***Source:*** **S&P 500 index on October 1995**[1]

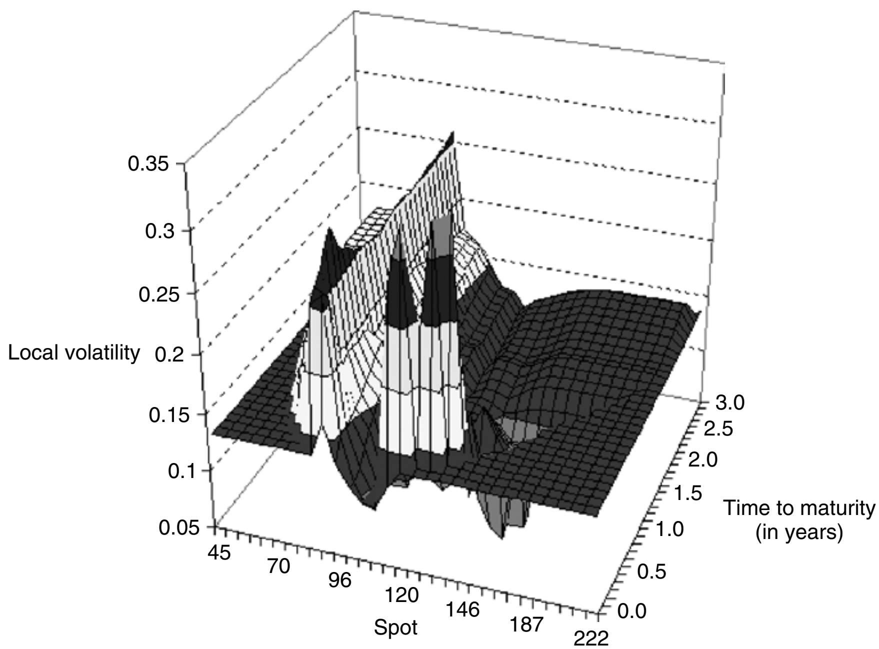


Figure 2: Local volatility surface inferred from vanilla options market prices

direction. For instance, far in the future, local volatilities are roughly constant, i.e. the model predicts a flattening of the smile, which seems inconsistent with the omnipresence of the skew or smile observed for the last 15 years. Not mentioning the numerical efforts in order to interpolate and extrapolate the sparse empirical smile data, then to smooth the surfaces of interest. This is computationally known as an 'ill-posed inverse problem'.

2 Is the local volatility model really a model?

2.1 The sirens of 'tweaking'

When you think about it, the local volatility models just provide numerical methods for finding a volatility surface $\sigma(S, t)$ that fits the market data of the options, $C(K, T)$, by exploiting the mechanics of the pricing equations or the PDEs. To our mind, they do not really provide a (physical) explanation of the smile phenomenon. Dupire has not discovered a smile model. His great discovery was the forward PDE for pricing vanilla options of different strikes and different maturities in one solve. Tweaking the diffusion coefficient in the Black–Scholes PDE in order to match a given set of vanilla option prices is reminiscent of the method of 'epicycles' which was the only way to account for the movement of celestial bodies when the real scientific explanation was lacking. (See Henrotte 2004 for a defence of homogeneous models against the dangers of 'tweaking' and Ayache 2001 for an early version of the argument.) Local volatility models do not intend to explain the volatility smile problem by introducing new dynamics for the underlying stock. And by 'new dynamics' we mean something original, like jumps or stochastic volatility or default. Suggesting that smiles are caused by jumps in the underlying or by stochastic volatility (or both) not only sounds realistic and informative, but may qualify as an explanation. Think how incredible it must sound, in comparison, that volatility should locally rise at a given point in time and space, then drop at some other point, for the sole purpose of matching today's option prices! It really sounds as if somebody was trying to force an interpretation in terms of local volatility on a phenomenon which has different and deeper origins. As a matter of fact, Jim Gatheral (2003) has provided what is to our mind the right interpretation of local volatility. He shows that local volatility is but the local expected variance of the underlying in general stochastic volatility models (that is to say, in 'realistic' models).

2.2 The 'natural' local volatility surface

Another reason why we should be suspicious of the local volatility model and why it falls in a class of its own (which may simply be the class of 'not being a model') is that it is non-parametric in essence or else arbitrarily parametric. Dupire's derivation essentially shows that any smile surface can be fitted by local volatility provided the model is non-parametric, and it basically provides the non-parametric formula. On the other hand, methods consisting in parameterizing the local volatility surface a priori (through spline functions or any other convenient representation), and in fitting the smile surface by minimization of a loss function (Coleman *et al*. 1999, Jackson *et al*. 1998), suffer from the arbitrariness of the representation, particularly the arbitrariness of the behaviour of local volatility at the boundaries of the domain. Proponents of such approaches are always at pains trying to justify their favourite representation of the local volatility surface on grounds of its intuitive appeal or physical realism or what have you. It is not uncommon that they maximize some entropy or some regularity criterion while minimizing their loss function, the underlying idea being that nature somehow favours smoothness and regularity. In a word, they look for the 'most natural local volatility function' matching the option prices. One wonders what that means.

2.3 Arbitrage-free interpolators

Jump-diffusion and stochastic volatility models, by contrast, lend themselves naturally to the routine of fitting the option prices by minimization of a loss function, as they are 'naturally parameterized' by the coefficients of the process (for instance the intensity of jumps and the parameters of the jump size distribution in the Merton model (1996); the volatility of volatility, its mean reversion, its correlation with the underlying in Heston 1993, etc.). As research on local volatility models was getting more and more entangled in issues purely computational (finding the smoothest arbitrage-free interpolation, maximizing the right regularity criterion, etc.; Andersen and Brotherton-Ratcliffe 1998, Avellaneda *et al*. 2000, Bodurtha and Jermakyan 1999, Coleman *et al*. 1999, Jackson *et al*. 1998, Kahale 2003, Lagnado and Osher 1997, Li 2001), and was drifting farther and farther away from the 'physics' of the problem, it so happened one day that our computational expert asked our financial theorist what to his mind the 'most natural local volatility function' could be, suited for a given smile. Undecided between many attractive numerical alternatives, our man was seeking guidance from the underlying 'physics'. Not surprisingly, the financial theorist suggested he looked at local volatility surfaces 'such as might have been produced by models of jumps in the underlying, or stochastic volatility, etc.' In other words, the suggestion was that the best solution to the *numerical problem* of inferring the smoothest, most regular, and arbitrage-free local volatility surface was to pretend that the option prices were generated by a jump-diffusion, stochastic volatility model! If you are so keen on local volatility, then indeed jump-diffusion/stochastic volatility models can be sold to you as 'financially meaningful, arbitrage-free, super-interpolators'. This is just the rehearsal of Gatheral's point. Only the question now becomes: If you go this far, why bother with local volatility any longer? For market completeness perhaps?

2.4 'Local' everything?

More to the point: Why hasn't anybody ever tried to fit a *non-parametric* jump-diffusion or stochastic volatility model to option data? Why is everybody busy searching for constant (or perhaps only time-dependent) parameters in Heston, Merton, SABR (Hagan *et al*. 2002), and nobody has proposed that both the diffusion coefficient and the jump coefficients, or both the volatility of volatility and the correlation coefficient, may become non-parametric functions of time and space? One possible answer is that the model would very rapidly become computationally infeasible; with the implication that the reason why non-parametric inference is actually done in the pure diffusion model and in no other model (or, in other words, the reason why local volatility models simply exist) is that it *can* be done. Hardly a proud conclusion. It means that local volatility models are just a temporary diversion outside the tracks of true progress. Another possible answer is that the continuum of vanilla call prices $C(K,T)$ will no longer be sufficient for calibration purposes when more than one parameter of the pricing equation are made a function of time and space. One would require an additional continuum of market prices, not redundant with the vanillas. Why not add, for instance, the continuum of prices of American one-touches $OT(B,T)$ of different barrier levels and maturity dates? As it happens, this might ensure agreement with the market prices of barrier options, an urgent problem for all exotic options trading desks.

We will have a lot more to say later about additional market information that we may require in the calibration phase. Enough to observe for the moment that the literature is not treating the showdown between local volatility and the other smile models properly. Like we said, local volatility is not a model, it is the tweaking of Black–Scholes. And the tweaking could equally be applied to Heston, or Merton, or any alternative smile model, if only we had the computational guts to do so. It seems the literature is standing at a methodological crossroads between the tough computational decision to involve additional instruments in the calibration—no matter the specific model or its parametric/non-parametric status—and the temptation to develop specific models just for their own sake and the sake of an original name, then to check whether they predict the right exotic option prices, or the right smile dynamics. At any rate, it is unfortunate that external issues, such as tractability, solvability, elegance of formulation, etc., should be the ultimate guides of scientific research. **We motivate our chapter by situating it precisely at this crossroads.**

As a matter of fact, an attempt could be made at the calibration of a jump-diffusion model with local diffusion component and local jump intensity. Indeed, a natural extension of the Black–Scholes diffusion model in the equity world is to include the risk of default in the pricing problem of equity derivatives subject to credit risk, like convertible bonds. This introduces the hazard rate function $\lambda(S,t)$ in the usual partial differential equation:

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

where X is the loss given default, and means we would have to calibrate the hazard rate function, on top of the volatility function, to available market data. The obvious candidates are the continuum of vanilla option prices $C(K, T)$ and the continuum of credit default swap spreads as a function of present stock price and maturity $CDS(S, T)$. See Andersen and Buffum (2003) for an example of such joint calibration. Note, however, that Andersen's procedure is parametric in that he proposes simple parametric representations of $\sigma(S, t)$ and $\lambda(S, t)$. But nothing stops us, in theory, from extending the forward induction argument of Dupire, or the Fokker–Planck equation approach of Klopfer and Tavella (2001), to the case where the probability density diffuses under the Brownian component as usual and 'leaks' into the state of default through the Poisson intensity of the default jump process, and from inferring $\sigma(S, t)$ and $\lambda(S, t)$ non-parametrically.

2.5 The mirage of the vanillas

The conclusion we draw from our first bash at local volatility models is twofold. First, local volatility is not a model. It is the 'corruption' of a model[2] and the corruption, for that matter, can spread over to all the other models. At best, local volatility can be seen as a shorthand or an interpretation: it is the local expected variance of some deeper and more realistic dynamics. (Think of Ehrenfest's theorem which interprets the classical mechanical variables as expectations of the 'true' quantum mechanical observables.) Second, when thinking about the other models (jump-diffusion, stochastic volatility, etc.), one should keep in mind that they can be made 'local' too. For once one recognizes that vanilla option prices will not be sufficient for calibration in that case, one realizes that there is nothing special about the vanillas anyway. The only reason why authors of jump-diffusion, stochastic volatility, or universal volatility models insist on fitting them to the vanillas is that they followed in the steps of the local volatility approach and vanillas were the obvious calibration candidates there.

We also fear the real reason might be that vanillas alone admit of analytical solutions in the models they propose, or even worse, that they have precisely grabbed the models which offered analytical solutions for the vanillas to begin with. We would love to see some of these authors calibrate their jump-diffusion, stochastic and universal volatility models, to a handful of options of *significantly different payoff structures*: vanillas, barriers, cliquets, credit default swaps, etc. As a matter of fact, vanilla options can be the poorest candidate for encapsulating the information about the stochastic process, when processes more general than a diffusion are considered. That our problem is called the 'smile problem' is no reason why the calibration of the model, or even its whole intention, should revolve around the vanillas. And that vanilla option trading is the ancestor of exotic option trading, or that traders are accustomed to envision alternative stochastic processes in terms of the vanilla smiles they generate, is an even worse excuse. But again, SABR would not be SABR if it did not allow the expansion of the Black–Scholes implied volatility (in other words the vanilla smile) in terms of the parameters of the process, and Heston would not be Heston, or Hull and White (1988) Hull and White, if...

3 Formulation of the smile problem

3.1 The real smile problem

Not only can we argue, *on a priori grounds or from a purely methodological point of view*, that the local volatility model is not a model, but it also demonstrably fails *as a model* of option smiles.

Indeed the real smile problem is not how to fit the vanillas or how to price them! Straightforward spline interpolation does that very nicely. The real smile problem is the pricing of exotic options and more generally the hedging of all kinds of options, including the vanillas, under dynamic assumptions at variance with the Black–Scholes model. As noted by almost everybody, the local volatility model fails miserably on both counts. Both the barrier option price structure and the dynamic behaviour of the smile predicted by a vanilla-calibrated local volatility model diverge from empirical observation (Lipton and McGhee 2002, Hagan *et al*. 2002). 'The failure of the local volatility model,' writes Hagan, 'means that we cannot use a Markovian model based on a single Brownian motion to manage our smile risk.' We need to assume an independent process for volatility. This opens the door to stochastic volatility models, and more generally, to all kinds of alternative dynamics that have been proposed over time as a replacement of Black–Scholes.

Perhaps the most important aspect of the smile problem today is to find a way of discriminating between all the alternative proposals to solve it. This is the symptom of a science in crisis, not just the symptom of a problem. Definitely the accurate pricing of exotics and the soundness of the hedging strategy are good selection criteria. To put it in Lipton's words (2002):

> We describe a series of increasingly complex models that can be used to price and hedge vanilla options consistently with the market. **We emphasize that, although all these models can be successfully calibrated to the market, they produce very different hedging strategies.** [...] A number of models have been proposed in the literature: the local volatility models of Dupire (1994), Derman & Kani (1994) and Rubinstein (1994); a jump-diffusion model of Merton (1976); stochastic volatility models of Hull and White (1988), Heston (1993) and others; mixed stochastic jump-diffusion models of Bates (1996) and others; universal volatility models of Dupire (1996), JP Morgan (1999), Lipton & McGhee (2001), Britten-Jones & Neuberger (2000), Blacher (2001) and others; regime switching models, etc. [...] Too often, these models are chosen ad hoc, for instance, on the grounds of their tractability and solvability. **However, the right criterion, as advocated by a number of practitioners and academics, is to choose a model that produces hedging strategies for both vanilla and exotic options resulting in profit and loss distributions that are sharply peaked at zero.**

This is the most cogent formulation of the smile problem we know of.

3.2 Indeterminateness of the conditionals

We shall quickly review the smile models which are most representative of today's smile literature, but let us first investigate the reason why smile models of different stochastic structure may not agree on exotic option pricing or the option hedging strategies (a.k.a. 'smile dynamics') even when calibrated to the same vanilla smile. The picture becomes clear when we have a look at the way the calibration is carried out. Denoting $A^{i,j}_{i_0,j_0}$ the price at state i_0 and time j_0 of a security paying off \$1 at state i and future time j (a.k.a. Arrow–Debreu security), it can be related to the vanilla call option prices in the following way:

$$A^{i,j}_{i_0,j_0} = \frac{C(K_{i+1}, T_j) - 2C(K_i, T_j) + C(K_{i-1}, T_j)}{\Delta K^2} \tag{1}$$

In continuous time and space this is expressed by

$$p(S,t;K,T)e^{-\int_t^T r(s)\,ds} = \frac{\partial^2 C(S,t;K,T)}{\partial K^2}$$

where $p(S,t;K,T)$ is the transition probability density from initial state and time (S,t) to (K,T). Introducing the vector notation:

$$A^{j}_{i_0,j_0} = \begin{bmatrix} A^{1,j}_{i_0,j_0} \\ A^{2,j}_{i_0,j_0} \\ \vdots \\ A^{N,j}_{i_0,j_0} \end{bmatrix} \tag{2}$$

and the matrix notation:

$$A^{j+1}_{j} = \begin{bmatrix} A^{1,j+1}_{1,j} & A^{2,j+1}_{1,j} & \cdots & A^{N,j+1}_{1,j} \\ A^{1,j+1}_{2,j} & A^{2,j+1}_{2,j} & & A^{N,j+1}_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ A^{1,j+1}_{N,j} & A^{2,j+1}_{N,j} & \cdots & A^{N,j+1}_{N,j} \end{bmatrix} \tag{3}$$

Up to a discounting factor, this is the matrix of conditional transition probabilities from states at date j to states at date $j+1$. (Crucially, the assumption here is that states of the world are just states of the underlying.)

The conditional probability rule yields the following equation:

$$\left(A^{j+1}_{i_0,j_0}\right)^T = \left(A^{j}_{i_0,j_0}\right)^T A^{j+1}_{j} \tag{4}$$

Without any further information about the structure of the stochastic process, this is the only constraint that the prices of vanilla options today impose on the matrix of conditional probabilities. Infinitely many matrices solve that equation of course. In a continuous diffusion framework this forward equation becomes

$$\frac{\partial p}{\partial T} + \frac{\partial (rKp)}{\partial K} - \frac{1}{2}\frac{\partial^2(\sigma^2 K^2 p)}{\partial K^2} = 0 \tag{5}$$

and shows why the knowledge of the prices of Arrow–Debreu securities maps the diffusion process $\sigma(K,T)$ completely.

3.3 Smile dynamics and model dependence

To repeat, the only information contained in the set of vanilla option prices $C(K,T)$ of different strikes and different maturities, independently of any model, is the map of transition probabilities from present day and present spot to whatever future time and future spot we are looking at. This says nothing about the conditional transition probabilities from a future date to a further future date. Additional information is needed to help determine those conditionals. In theory, we would need the knowledge of all 'forward smiles', in other words, the future prices of all vanilla options as seen from all possible states of the world, *not mentioning that the underlying stock price may not be the only state variable* (in stochastic volatility models, typically).

Choosing a particular model for the underlying dynamics definitely adds some structure. It is a form of parametrization of this totally non-parametric picture. The only 'structure' that the local volatility model adds consists in removing the need for market information beyond the vanilla option prices in the fully non-parametric case. The 'matrix' of conditionals is fully determined in that case, and there is no spatial state variable other than the underlying. Alternative models such as jump-diffusion, or stochastic volatility, or universal volatility models, also dramatically reduce the degrees of freedom in the choice of the conditionals, particularly so when the coefficients of the given process are constant, or time dependent, or assume some parametric form. Now think how different the structure of conditionals that they imply can be, compared to the pure diffusion case (e.g. the possibility of jumping and hitting a barrier in between future dates, the addition of another state variable indexing the forward smiles, etc.), yet their authors calibrate them to the vanillas just the same! In a sense, the local volatility model is more honest than the other models with regard to the conditionals. You just know there is nothing you can do. In the other models, by contrast, you calibrate a bunch of *constant parameters* in what seems to be a legitimate calibration move—typically you calibrate them to the vanillas—and this sets for you all the conditional structure. Hardly can a result be more model dependent!

3.4 Our preferred model

The reason why the local volatility model, the jump-diffusion models, the stochastic volatility models, or more generally the 'universal volatility models', may agree or not agree among each other or with the market on the prices of barrier options or forward starting options, is that each model imposes a specific smile dynamics, or structure of conditionals. We claim that this smile dynamics should not be imposed by the model, but inferred from the market. However, we have to pick a certain framework.

Calibration, pricing and dynamic hedging cannot be totally model independent, even though model independence should always act as a 'regulative ideal' in our research program. **We shall pick the framework with the features that everybody knows today are essential for explaining the smiles.** We know we need jumps (if only to account for shorter dated smiles and default risk) and we know we need stochastic volatility (to account for longer dated smiles and to acknowledge the very *raison d'être* of option markets and market-makers). Our discussion of local volatility and Henrotte's powerful statement[3] should steer us away from inhomogeneous models. **The coefficients of our stochastic process shall be constant.** However, we have learnt from the unhappy story of the conditionals that market option data, other than the vanillas, must be included in the calibration procedure. Under no circumstance shall we be prevented from doing so by what Henrotte describes, in other people's cases, as 'a very somber agenda': the need to produce closed form or quasi closed form pricing solutions. **Our pricing equations shall be solved by numerical algorithms.** For all these reasons, chiefly the fact that model names have traditionally been associated with the discovery of analytical solutions, our model shall bear no particular name. **We shall call it 'Nobody's model'.**

3.5 Including exotics in the calibration

On the calibration side, we have noted that the value of barrier options is sensitive to the flux of probability across the barrier (jumps, and volatility dynamics up to the barrier). The value of forward starting options, on the other hand, is directly linked to the conditional transition probabilities, or forward smiles. In other words, both depend on what extra structure the matrix of conditional transition probabilities may have, on top of the constraint given by the spot vanilla smile. This designates simple barrier options like the one-touch or American digital, and the forward starting options as the natural candidates for extending our calibration set and helping determine the smile dynamics.[4] Traders accustomed to Derman's (1999) classification of smile dynamics in terms of 'sticky-strike' or 'sticky-delta volatility regimes' know that the delta of the vanillas is very much dependent on the type of volatility regime the market is in. Derman's study produces evidence that both kinds of regimes have obtained over time within a single market. Depending on the regime you think the market is in, you make the following adjustment to your Black–Scholes hedge.

When $\sigma_{imp}(S, t, K, T)$ is the implied volatility for a European style option we have:

$$C(S, t, K, T) = C_{BS}\left(S, t, K, T, \sigma_{imp}(S, t, K, T)\right) \tag{6}$$

The delta hedge becomes a combination of Black–Scholes delta and a correction term due to the regime of movement of the smile with a moving underlying:

$$\Delta = \frac{\partial C}{\partial S} = \frac{\partial C_{BS}}{\partial S} + \frac{\partial C_{BS}}{\partial \sigma_{imp}} \cdot \frac{\partial \sigma_{imp}}{\partial S} \tag{7}$$

We claim that nobody should be in a position to decide which particular smile dynamics will prevail. It is really like guessing a price (as Marco Avellaneda once rightly observed in a financial workshop at NYU). Only the market can provide such information. We are saying that *your wrong guess about the smile dynamics can generate an immediate arbitrage opportunity against you, if somebody picks the right security to trade against you*. As a matter of fact, all FX option traders are aware of the existence of such a security! It is the barrier option, the simplest instance of which is the one-touch.

Different projected evolutions of the vanilla smile lead to different spot prices of barrier options in the FX traders' minds, because they think of the future cost of unwinding the vanilla static hedge that they have set up against the barrier option. This insight can be further refined and made rigorous in a fully dynamic hedging picture. (Indeed the vanilla static hedge that those FX exotic option traders have in mind is not always consistent with the smile dynamics they project. For instance they immunize the vega, the vanna and the volga of the barrier option with a static combination of vanillas, yet they derive their hedging ratios from the Black–Scholes model which assumes constant volatility.[5])

The price structure of the one-touches contains implicit information about the smile dynamics, therefore about the delta you should be using to hedge the vanilla options! So does the price structure of the forward starting options. This is why the one-touches and the forward starting options must be included in the calibration.

In conclusion, the exotic option pricing problem and the problem of smile dynamics are intimately linked, and the pricing/hedging model cannot dispense with including exotic options in the calibration.

4 A quick review of representative smile models

4.1 Stochastic volatility

In stochastic volatility models (Heston 1993, Hull and White 1998), volatility is itself stochastic and follows some mean reverting process with its own volatility and correlation with the underlying share. The stochastic volatility models can be seen as modelling the option price as an average of the Black–Scholes prices with respect to volatility. This model is essential for the pricing of longer-dated options which are most sensitive to volatility changes. It avoids the scale effect observed in long-term local volatilities. Least square fit is used to search for model parameters to match observed market prices.

The problem with stochastic volatility models is that the derivative instrument is exposed to volatility risk on top of market risk, and the underlying cannot hedge both Brownian motions.

The Heston model is, for instance, given by the following risk-neutral process:

$$\frac{dS}{S} = r\,dt + \sqrt{v}\,dW$$
$$dv = \kappa\,(\theta - v)\;dt + \varepsilon\sqrt{v}\,dZ$$

where the volatility process and the underlying process are correlated through a correlation coefficient ρ. And the pricing equation is given by:

$$\frac{\partial V}{\partial t} + \frac{1}{2}v\left(S^2\frac{\partial^2 V}{\partial S^2} + 2\rho\varepsilon S\frac{\partial^2 V}{\partial S\partial v} + \varepsilon^2\frac{\partial^2 V}{\partial v^2}\right) + rS\frac{\partial V}{\partial S} + \kappa\,(\theta - v)\,\frac{\partial V}{\partial v} = rV$$

The calibration of the model consists of finding parameters of the volatility process: κ (mean reversion), θ (long-term volatility), ϵ (volatility of volatility), ρ (correlation between the volatility process and the underlying process) as well as initial volatility state v_0, such that option market data is fitted.

4.2 Jump-diffusion

Jump-diffusion models (Merton 1996) add jumps and crashes to the standard diffusion process of the underlying. They intend to reproduce the underlying dynamics more realistically and to capture the strong smile exhibited by short-dated options. The underlying share price follows a risk-neutral process governed by the following equation:

$$\frac{dS}{S} = (r - \lambda m)\;dt + \sigma\,dW + \left(e_j - 1\right)\,dN$$

where N is a Poisson process with frequency λ, W is a Wiener process independent of N, j is a random logarithmic jump size with pdf $\phi(j)$ and m is the expected value of $e_j - 1$.

The problem again is that the Black–Scholes continuous hedging argument breaks down in the presence of jumps.

Some other models lay jumps on top of stochastic volatility models (Bates 1996).

4.3 Universal volatility

Blacher The universal volatility model of Blacher is described by the following risk-neutral process:

$$\frac{dS}{S} = rdt + \sigma\left(1 + \alpha(S - S_0) + \beta(S - S_0)^2\right) dW$$
$$d\sigma = \kappa\,(\theta - \sigma)\;dt + \varepsilon\sigma dZ$$

The volatility σ follows a mean reverting process to level θ, correlated with the underlying process via ρ.

It is worthy of note that Blacher motivates his universal volatility model for reasons almost opposite Hagan *et al.* (2002). Like Hagan, he speaks for stochastic volatility models. However, he notes that although the 'smile is stochastic, simple stochastic volatility models [such as Heston's] do not predict a systematic move of the relative smile when the spot changes'. 'Not what we observe in the market,' he says. 'This means hedging discrepancies, starting with a wrong delta.' In other words, Blacher is noting that space homogeneous models like Heston's follow the sticky-delta rule. The 'relative smile' they imply, i.e. the smile with respect to moneyness or delta of the option, is unchanged when the underlying spot changes. Yet Blacher wishes that the vanilla smile may not always move coincidentally with the underlying. He claims control over the smile dynamics. In order to achieve this, he has no choice but to reintroduce inhomogeneity in the spot homogeneous stochastic model.

He writes: 'α, the slope of the deterministic part, creates skew and governs the change of ATM implied vol with respect to change of underlying. β, the curvature of the deterministic part, creates smile curvature and governs the change of the slope of the smile curve with respect to change of underlying.'

Note that SABR also breaks the homogeneity of degree 1 by allowing values for β different from 1, in the risk-neutral process:

$$dF = \alpha F^{\beta} dW_1$$
$$d\alpha = \upsilon\alpha dW_2$$

F is the forward price, α its volatility, υ the volatility of volatility, and dW_1 and dW_2 are Wiener processes correlated through:

$$\langle dW_1, dW_2\rangle = \rho \cdot dt$$

Lipton Lipton (2002), on the other hand, argues for his universal volatility model on the grounds of its adequacy for pricing barrier options. He writes:

> A properly calibrated universal model matches the market [of barrier options] much closer than either local or stochastic volatility models, which tend to sandwich the market. [...] While both local and stochastic volatility models produce price corrections [for barrier options] in qualitative agreement with the market, only a universal volatility model is capable of matching the market properly. In our experience, this conclusion is valid for almost all path-dependent options.

By ‘properly calibrated universal model’ Lipton means ‘calibrated to the vanillas’. On the specific topic of calibration he otherwise notes: ‘Because of its complexity, the universal volatility model can be solved explicitly only in exceptional cases (which are of limited practical interest). [...] The model calibration, of course, is a different matter.’

Lipton’s risk-neutral stochastic process is given by:

$$\frac{dS}{S} = (r - \lambda m)\,dt + \sqrt{v}\sigma_L(t, S)\,dW + \left(e^j - 1\right)\,dN$$
$$dv = \kappa\,(\theta - v)\,dt + \varepsilon\sqrt{v}dZ$$

And the pricing equation is given by:

$$\frac{\partial V}{\partial t} + \frac{1}{2}v\left(\sigma_L^2(t, S)S^2\frac{\partial^2 V}{\partial S^2} + 2\rho\varepsilon\sigma_L^2(t, S)S\frac{\partial^2 V}{\partial S\partial v} + \varepsilon^2\frac{\partial^2 V}{\partial v^2}\right)$$
$$+(r - \lambda m)S\frac{\partial V}{\partial S} + \kappa\,(\theta - v)\,\frac{\partial V}{\partial v} + \lambda\int_{-\infty}^{+\infty} V\left(e^j S\right)\phi(j)\,dj = (r + \lambda)V$$

where $\sigma_L(t, S)$ is the local volatility part, κ the mean reversion of volatility, θ the long-term volatility, ϵ the volatility of volatility, ρ the correlation between the volatility process and the underlying process, λ the intensity of the Poisson jump process, $j > 0$ the random logarithmic jump size with PDF $\phi(j)$, and m the expected value of $e^j - 1$.

4.4 Conclusion

In conclusion of our review of existing smile models, let us retain the following fact. The local volatility model and the stochastic volatility model stand at opposite extremes. The first is inhomogeneous, the second is homogeneous. Neither one predicts the right smile dynamics or produces the right barrier options prices. Only the universal volatility model, which allows explicit control over the smile dynamics (by reintroducing inhomogeneity and by mixing local volatility behaviour with stochastic volatility behaviour), manages to fit the smile dynamics (Blacher 2001) and at the same time to fit the barrier option prices (Lipton and McGhee 2002).

Let us then solemnly pose the question: ‘*Is the recourse to inhomogeneity really indispensable?*’ Or again: ‘*Given our plea for inclusion of the exotics in the calibration and our credo in homogeneous models, can we also claim control over the smile dynamics?*’

5 Numerical illustrations of the smile problem

We will try to answer that big question by way of practical examples rather than fundamental theorizing. The examples will also serve the purpose of illustrating the smile problem, namely that models of different stochastic structure may very well agree on the vanilla smile yet completely disagree on the exotics and smile dynamics. Instead of solving Heston’s model, or Dupire’s model, or Lipton’s model, we will build up our series of examples from a simple instance of the ‘model with no name’, the model we have called ‘Nobody’s model’.

5.1 The calibration issue

Baby examples First, we consider a simple jump-diffusion model where the underlying diffuses with a constant Brownian volatility and may incur two jumps of fixed size and constant Poisson intensity. We call this simple stochastic structure ‘Baby1’.

TABLE 1: BABY1 PARAMETERS

Brownian diffusion Jump size	7.00% Jump intensity
−25%	0.2
10%	0.4

For illustration, we consider a Brownian volatility component of $v = 7\%$, an upward jump of size $y_1 = 10\%$ and intensity $\lambda_1 = 0.40$ and a downward jump of size $y_2 = -25\%$ and intensity $\lambda_2 = 0.2$. Table 1 summarizes the parameters of Baby1.

TABLE 2: VOLATILITY NUMBERS IMPLIED BY BABY1

Strike Maturity (years)	0.16	0.49	1
80	30.67%	22.20%	18.97%
85	27.41%	20.97%	18.33%
90	22.12%	18.47%	17.19%
95	15.47%	15.32%	15.70%
100	10.90%	12.96%	14.32%
105	11.69%	12.12%	13.37%
110	13.67%	12.16%	12.83%
115	14.48%	12.42%	12.58%
120	15.79%	12.73%	12.49%
130	17.37%	13.44%	12.56%
140	18.74%	14.08%	12.77%

The probabilities of jump are given in the risk-neutral measure. Consequently, we can compute the vanilla option prices generated by this process and re-express them in Black–Scholes implied volatility numbers (see Table 2), thus producing the smile. The interest rate is $r = 2\%$ and the underlying spot is $S = 100$.

Note that the smile is steepest for shorter dated options, and tends to flatten out for longer terms (see Figure 3). We can see this simple model as a discretization of the 'traditional' jump-diffusion models (e.g. Merton 1996) with a probability distribution of jump sizes.

Volatility smiles can alternatively be represented as a function of the option delta and maturity rather than its strike and maturity. This is the origin of the appellations 'sticky-strike' and 'sticky-delta'. Smiles that are a function of the moneyness of the option are sticky-delta. Their representation in the delta/maturity metric is invariant when the underlying moves. Figure 4 shows the alternative graph of our smile in that metric.

We recompute our smile for $S = 120$ (Figure 5). As our jump-diffusion model is homogeneous and volatility and jump sizes relate to proportional changes of the underlying, the resulting smile surface is sticky-delta. It is unchanged in the delta/maturity metric, and it moves along with the underlying in the strike/maturity metric.

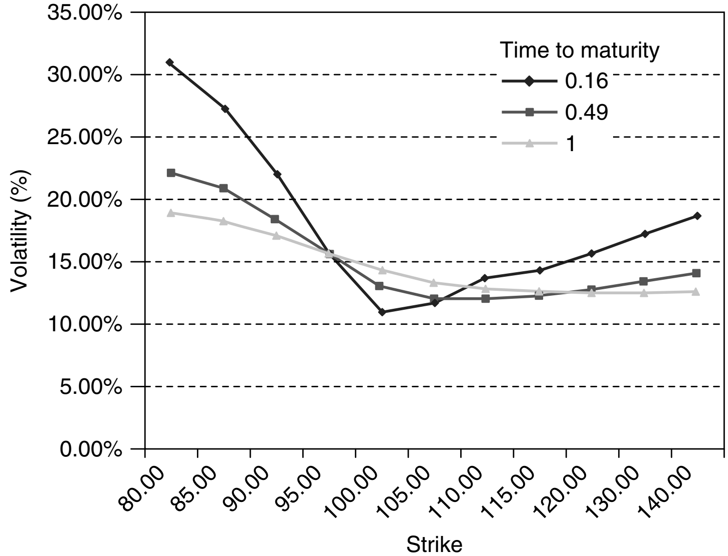


Figure 3: Volatility smile generated by Baby1 against strike price for three different expirations and underlying spot price of 100

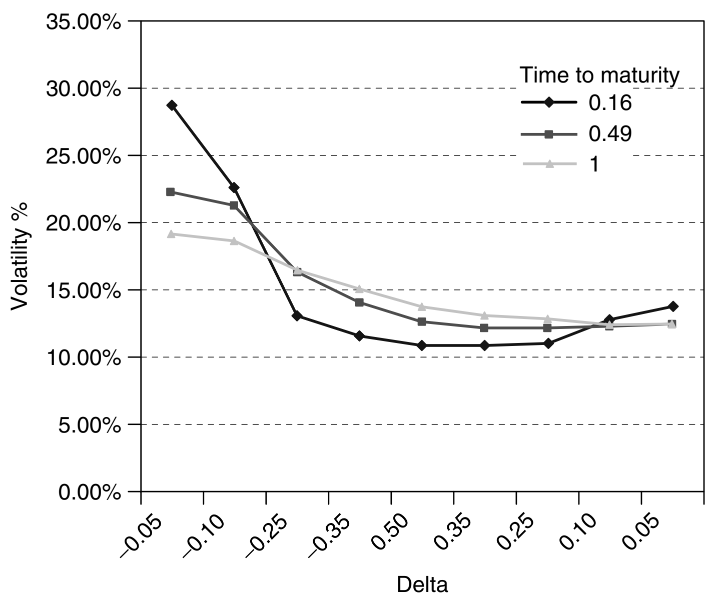


Figure 4: Volatility smile generated by Baby1 against delta for three different expirations and underlying spot price of 100

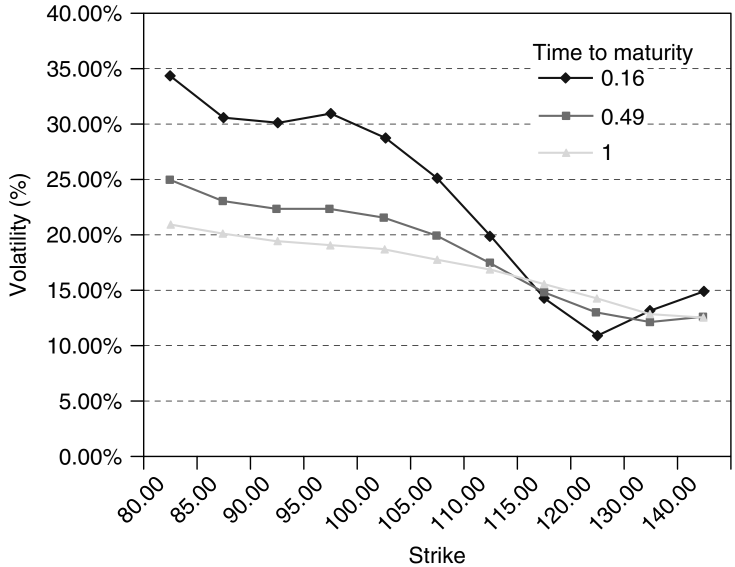


Figure 5: Smile produced by Baby1 against strike price for three different expirations and underlying spot price of 120

Next, we consider another simple stochastic structure that we call 'Baby2'. The volatility of the Brownian component is now stochastic and can assume two states, or regimes. The transitions, or jumps, between the two volatility states are caused by Poisson processes of constant intensity. At least two Poisson processes are needed to secure the transition from Regime 1 to Regime 2 and back. As Brownian volatility jumps between regimes, the underlying may simultaneously incur a jump of fixed size. This builds in correlation between jumps in the underlying (or return jumps) and volatility jumps. By convention, Regime 1 is the present regime. You can think of Baby2 as a simplification of stochastic volatility models with correlated return jumps and volatility jumps.

We then propose the following. We shall use Baby2 to try to fit the vanilla smile generated by Baby1. Note that Baby1 admits of five free parameters (the Brownian diffusion coefficient, the two jump sizes and the two jump intensities) and Baby2 of six (the diffusion coefficients in the two regimes, the two inter-regime jump sizes and the two jump intensities).

Calibration of Baby2 is achieved by searching for the six parameters by least squares fitting of the option prices produced by Baby1. The calibration results are shown in Figures 6 and 7 and the set of parameters is summarized in Table 3. Then we see how Baby1 and Baby2 price a given barrier option.

As seen in Table 4 and Table 5, Baby1 and Baby2 seem to be in agreement on the prices of the vanilla options and yet in disagreement on the price of the call 100 up and out at 107. You may think the discrepancy between the barrier option prices is due to the fact that Baby2 has not exactly matched the vanilla smile generated by Baby1. Indeed, Baby2 is structurally different from Baby1 in that it can only pick up a *single* return jump, when it starts in Regime 1. This jump takes it to Regime 2, and it is only then that it may incur a jump of a different nature. Notice how Baby2 has managed to decipher Baby1's downward jump (it finds a jump of size −28% and

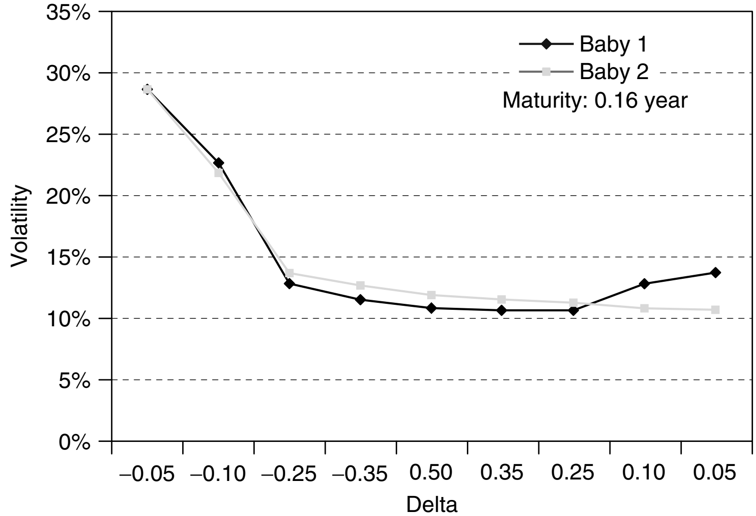


Figure 6: Comparison of the implied volatility curves of Baby2 and Baby1 for 0.16 year maturity

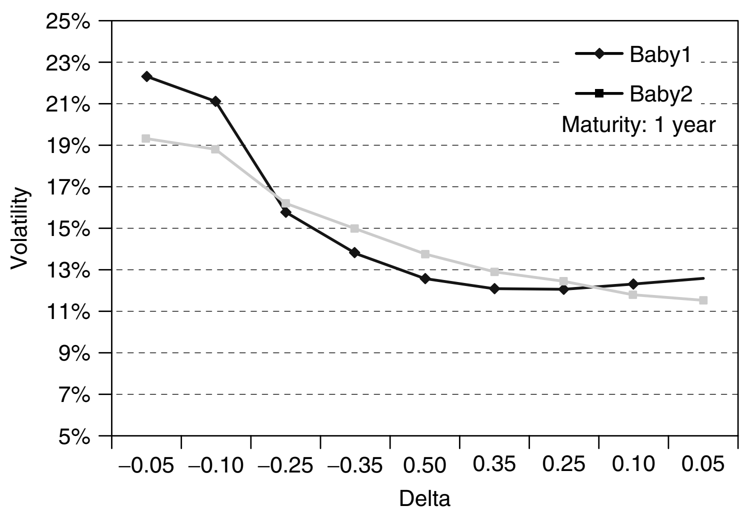


Figure 7: Comparison of the implied volatility curves of Baby2 and Baby1 for a maturity of 1 year

intensity 0.14 to account for the jump of size −25% and intensity 0.20), and how it has fudged Baby1's 7% Brownian and upward jump into a Brownian component of 10.02%.

However, total volatility in Regime 1 of Baby2 is very close to total volatility[6] in Baby1 (see Table 6). As a result, Baby2 performs better at fitting the out-of-the-money put skew of Baby1 than the out-of-the-money call skew. Still, it may look surprising that the difference between the

TABLE 3: BABY2 PARAMETERS WHICH BEST FIT THE VANILLA SMILE GENERATED BY BABY1 (TABLE 2)

	Brownian diffusion
Regime 1	10.02%
Regime 2	8.44%

	Jump size	Jump intensity
Regime 1 → Regime 2	−28.07%	0.1395
Regime 2 → Regime 1	0.24%	0.3947

TABLE 4: COMPARISON OF THE PRICES GENERATED BY BABY1 AND BABY2 FOR DIFFERENT 6-MONTH MATURITY OPTIONS

		Call 100	Call 107	Put 93
Baby1	**Price**	4.12	1.28	1.58
	Implied volatility	12.96 %	12.07%	16.58%
Baby2	**Price**	4.22	1.25	1.51
	Implied volatility	13.31%	11.93%	16.24%

TABLE 5: CALL 100 UP AND OUT AT 107, OF MATURITY SIX MONTHS PRICED BY BABY1 AND BABY2

	Price
Baby1	0.74
Baby2	0.49

barrier option prices produced by the two models should be so big, especially so when the prices of the calls of strike 100 and 107 are not that different.

Body examples To clear any remaining doubt, we move to the next stage and consider a more evolved model. The underlying can now find itself in *three* different regimes of Brownian volatility. Transition between the regimes is still carried out by a Markovian matrix of six inter-regime Poisson jumps. The model now involves 15 free parameters (three Brownian diffusion coefficients, six jump sizes and six jump intensities). We call this new stochastic structure ‘Body’.

Baby1 and Baby2 now appear as special cases of Body. Baby2 corresponds to Body with the transitions to Regime 3 disabled. And Baby1 corresponds to Body with the three diffusion coefficients set equal to 7% and the two Poisson jumps from any of the three regimes to any other set equal to Baby1’s Poisson jumps.

TABLE 6: TOTAL VOLATILITY IN THE REGIMES OF BABY1 AND BABY2

		Total volatility
Baby1	1	14.63%
Baby2	Regime1	14.50%
	Regime 2	8.44%

We then propose the following. We shall calibrate Body twice to a full vanilla smile, each time with a different initial guess on the 15 process parameters. And we shall pick a real vanilla smile this time (the one in Figure 1 that gave us the local volatility surface in the first section), not an artificially created one. Then we shall turn to the pricing of barrier options. The results of calibration are shown in Table 7 and the corresponding sets of parameters are shown in Tables 8 and 9.

Notice that two calibration instances, Body1 and Body2, match the given market vanilla smile fairly closely (see Table 7 and Figures 8, 9 and 10). Also note that we manage to fit a *whole surface* of options prices, with different strikes and different tenors, with one set of constant parameters, when other smile models typically require that the parameters become functions of time.[7] True, the reason for that may be that our parameters are many (15) and our 'Body' model is not so parsimonious after all. This also explains why the calibration procedure may produce multiple solutions and the loss function admits of several local minima. As far as barrier options are concerned, we first look at the one-touches. In market practice, one-touches are identified and quoted relative to Black–Scholes. The '30% one-touch' conventionally refers to the American digital option, paying out \$1 as soon as the barrier is hit from below, *that would be worth 30 cents in the Black–Scholes world*, when priced with the ATM implied volatility of corresponding maturity. ('−30% one-touch' conventionally means that the barrier is hit from above.) A market quote of −4.88% for that one-touch means that it is actually worth (30% − 4.88%) = 25.12% in the present market, or smile, conditions.

Table 10 describes the one-touch price structures given by Body1 and Body2. The differences are considerable. As a result, standard barrier options will also be priced very differently by the two models (see Table 11). Notice that it is the *same* model (Body) that is producing agreement on the vanillas and total disagreement on the barriers between two calibration instances. The situation is different from the case of agreement/disagreement between two *different* models, such as local volatility and stochastic volatility, or jump-diffusion. Those simpler models merely disagree with each other because of a big difference in what otherwise qualifies as *simple* stochastic structure. It is not even guaranteed that they can fit a complete vanilla smile surface. Their case is somewhat comparable to the agreement/disagreement we found between Baby1 and Baby2. When the stochastic structures become complex, however, and start combining stochastic volatility and correlated return jumps and volatility jumps (in models such as Body, or universal volatility, which seem to be imposed on us anyway by the natural course of events and by the evolution of the smile problem), we shall expect to witness increasingly frequent cases where a certain vanilla smile is perfectly matched, yet certain exotic options are very badly mispriced, or priced just by pure luck. In other words, we are way past the old debate on whether local volatility is better, or jump-diffusion is better, or stochastic volatility is better, on whether they agree or disagree on the exotics, and whether universal volatility should come and replace them all. Definitely

TABLE 7: COMPARISON OF THE IMPLIED VOLATILITY SURFACES GENERATED BY BODY1 AND BODY2 WITH THE ONE INFERRED FROM VANILLA MARKET PRICES. THE SPOT PRICE IS 100

						Strike					
Maturity (years)	80	85	90	95	100	105	110	115	120	130	140
	Market	19.00%	16.80%	13.30%	11.30%	10.20%	9.70%				
0.18	**Body1**	19.22%	16.38%	13.35%	11.69%	10.38%	10.29%				
	Body2	19.11%	17.14%	13.91%	10.93%	10.76%	10.00%				
	Market	17.70%	15.50%	13.80%	12.50%	10.90%	10.30%	10.00%	11.40%		
0.43	**Body1**	17.56%	15.85%	13.97%	12.43%	11.14%	10.08%	10.07%	11.53%		
	Body2	17.49%	15.89%	14.11%	12.22%	11.29%	10.35%	9.82%	10.30%		
	Market	17.20%	15.70%	14.40%	13.30%	11.80%	10.40%	10.00%	10.10%		
0.70	**Body1**	17.34%	15.90%	14.37%	13.00%	11.85%	10.87%	10.11%	10.20%		
	Body2	17.15%	15.86%	14.50%	12.96%	11.91%	10.95%	10.36%	10.37%		
	Market	17.10%	15.90%	14.90%	13.70%	12.70%	11.30%	10.60%	10.30%	10.00%	
0.94	**Body1**	17.22%	15.93%	14.60%	13.39%	12.36%	11.47%	10.69%	10.23%	11.04%	
	Body2	17.05%	15.94%	14.77%	13.42%	12.39%	11.44%	10.81%	10.64%	10.74%	
	Market	17.10%	15.90%	15.00%	13.80%	12.80%	11.50%	10.70%	10.30%	9.90%	
1.00	**Body1**	17.19%	15.93%	14.65%	13.48%	12.46%	11.60%	10.83%	10.32%	10.86%	
	Body2	17.04%	15.96%	14.82%	13.52%	12.50%	11.55%	10.91%	10.71%	10.74%	
	Market	16.90%	16.00%	15.10%	14.20%	13.30%	12.40%	11.90%	11.30%	10.70%	10.20%
1.50	**Body1**	16.99%	15.98%	14.97%	14.03%	13.19%	12.46%	11.80%	11.24%	10.56%	10.89%
	Body2	16.95%	16.08%	15.17%	14.13%	13.24%	12.38%	11.71%	11.34%	10.96%	10.96%

TABLE 7:

	Market	16.90%	16.10%	15.30%	14.50%	13.70%	13.00%	12.60%	11.90%	11.50%	11.10%
2.00	**Model**	16.87%	16.03%	15.20%	14.42%	13.71%	13.07%	12.48%	11.98%	11.17%	10.76%
	Body2	16.86%	16.13%	15.38%	14.53%	13.78%	13.02%	12.38%	11.94%	11.35%	11.11%
	Market	16.80%	16.10%	15.50%	14.90%	14.30%	13.70%	13.30%	12.80%	12.40%	12.30%
3.00	**Body1**	16.74%	16.12%	15.52%	14.94%	14.40%	13.89%	13.42%	12.99%	12.26%	11.67%
	Body2	16.70%	16.16%	15.61%	15.02%	14.47%	13.90%	13.37%	12.93%	12.21%	11.73%
	Market	16.80%	16.20%	15.70%	15.20%	14.80%	14.30%	13.90%	13.50%	13.00%	12.80%
4.00	**Body1**	16.68%	16.19%	15.72%	15.26%	14.83%	14.42%	14.03%	13.67%	13.03%	12.48%
	Body2	16.58%	16.15%	15.74%	15.29%	14.87%	14.44%	14.01%	13.64%	12.96%	12.41%
	Market	16.80%	16.40%	15.90%	15.40%	15.10%	14.80%	14.40%	14.00%	13.60%	13.20%
5.00	**Body1**	16.63%	16.24%	15.85%	15.48%	15.12%	14.78%	14.45%	14.14%	13.58%	13.09%
	Body2	16.49%	16.14%	15.81%	15.45%	15.12%	14.78%	14.44%	14.13%	13.53%	13.01%

TABLE 8: BODY1 PARAMETERS

	Brownian diffusion	Total volatility
Regime 1	9.57%	11.67%
Regime 2	6.24%	32.23%
Regime 3	2.25%	11.88%

	Jump size	Jump intensity
Regime 1 → Regime 2	−9.07%	0.2370
Regime 2 → Regime 1	62.67%	0.0855
Regime 1 → Regime 3	2.72%	3.3951
Regime 3 → Regime 1	−3.17%	2.9777
Regime 2 → Regime 3	24.63%	1.0944
Regime 3 → Regime 2	−22.66%	0.2040

TABLE 9: BODY2 PARAMETERS

	Brownian diffusion	Total volatility
Regime 1	7.77%	11.63%
Regime 2	19.11%	25.08%
Regime 3	3.98%	7.45%

	Jump size	Jump intensity
Regime 1 → Regime 2	−9.02%	0.6254
Regime 2 → Regime 1	15.85%	0.5124
Regime 1 → Regime 3	5.24%	0.8750
Regime 3 → Regime 1	2.19%	0.7163
Regime 2 → Regime 3	17.17%	0.4589
Regime 3 → Regime 2	−11.20%	0.2891

universal volatility is the answer and Lipton's model has somewhat outgrown Lipton's article. As universal volatility models or SVJ models (stochastic volatility + jumps) seem unavoidable, the preoccupying issue today is how to avoid a dilemma, occurring within the *same* universal volatility model, such as embodied by Body1 and Body2.

You can easily imagine what the obvious trap would be. 'How shall we distinguish between multiple local minima, such as Body1 and Body2, and pick the right one?' and you may be tempted to answer: 'Let us pick the solution that fits the vanillas best, down to the last penny!' This is what a well-known analytics vendor seems to be proposing. Their way out of the dilemma is that a simulated annealing algorithm shall find *the global minimum of the loss function involving the vanillas only!* Has anyone worried where that would leave the exotics? We live in a very dangerous world indeed.

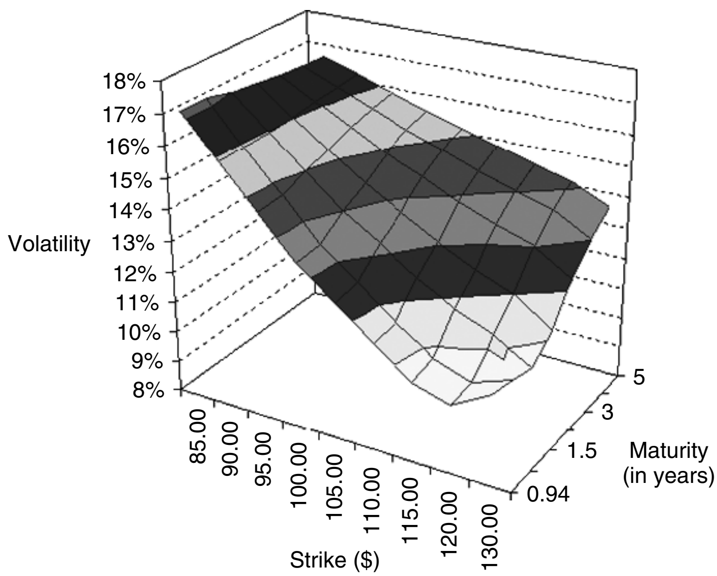


Figure 8: Body1 implied volatility surface

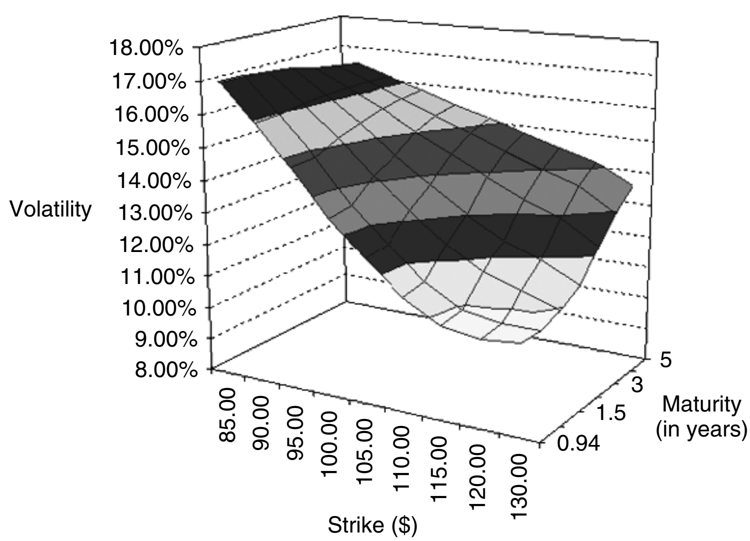


Figure 9: Body2 implied volatility surface

We know what the right proposal should be. *Include the one-touches, or other relevant exotic options, in the calibration procedure.* As a matter of fact, calibrating to the one-touches together with the vanillas transforms the ill-posed problem into a well-posed one. We will no longer try to reach for the global minimum among many local minima, but for a unique global minimum, full stop.

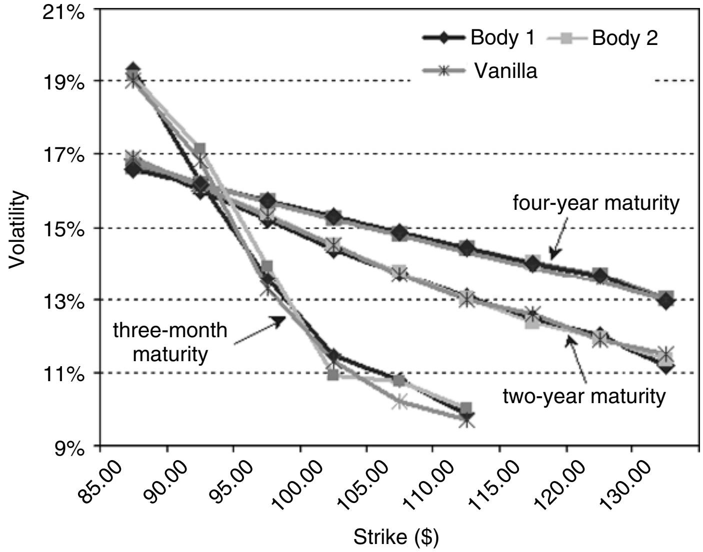


Figure 10: Cross-sections of the implied volatility surfaces shown in Figures 8 and 9 at three different maturities

To illustrate that, we calibrate Body to the vanilla smile and to *the whole collection of one-touches produced by Body1* (Table 10), yet we select as initial guess of parameters *the solution produced by Body2* (Table 9). This way we can see whether the one-touches will pull us out of what used to be the wrong local minimum. The calibration result is summarized in Table 12. We call it 'Body1Double', and check it against Body1. Our minimization routine is a standard Newton method.

Notice the following interesting phenomenon. Within an acceptable numerical tolerance, Body1Double and Body1 seem to agree on the Brownian diffusion in all three regimes and on the Poisson jump sizes and intensities taking us from Regime 1 to Regimes 2 and 3. They also agree on the Poisson jumps leading from Regime 3 to Regimes 1 and 2. However, Body1Double and Body1 seem to have switched the Poisson jumps leading from Regime 2 to Regimes 1 and 3. The explanation is that total volatility is roughly the same in Regime 1 and Regime 3 (while it is much higher in Regime 2), and that the only things that the underlying can 'see', once in Regime 2, are the *total volatility* of the Regime it will visit next and the Poisson jumps of course. While formally different, Body1 and Body1Double are in fact perfectly equivalent solutions (as when you permutate the regimes). As a matter of fact, we can check how well they agree on the pricing of the Put 103 knocked out at 95, for different spot prices and different regimes (Figure 11).

Full Body, anybody, and nobody You may wonder what is so special about the stochastic structure of Body. Nothing really, except that it has the minimum features that seem to be required to capture the phenomenology of smile and smile dynamics. As far as we are concerned, this is the only thing that counts. The question whether volatility should be diffusing rather than jumping in between discrete states, whether the Poisson jump distribution should be continuous rather than discrete, is in the last resort an aesthetic question (and often driven by the desire

TABLE 10: ONE-TOUCH PRICES INFERRED BY BODY1 AND BODY2

		One-touches									
Maturity (year fraction)		−5%	−10%	−20%	−30%	−50%	50%	30%	20%	10%	5%
0.175	**Body1**	0.51%	−1.26%	−3.81%	−5.37%	−6.44%	−6.13%	−7.81%	−8,36%	−6.08%	−3.58%
	Body2	3.99%	0.51%	−5.80%	−10.45%	−14.78%	−7.72%	−7.01%	−5.91%	−4.18%	−2.66%
1.5	**Body1**	7.15%	6.23%	2.44%	−1.70%	−8.19%	−3.04%	−6.64%	−7.89%	−6.67%	−4.04%
	Body2	8.78%	8.94%	6.63%	3.08%	−4.88%	−3.62%	−7.76%	−8.16%	−5.98%	−3.55%
5	**Body1**	8.12%	8.74%	7.56%	5.17%	−0.87%	−0.02%	−2.63%	−4.10%	−4.45%	−3.30%
	Body2	8.06%	9.12%	8.74%	7.10%	2.43%	−0.11%	−3.14%	−4.65%	−4.59%	−3.17%

TABLE 11: PRICING BY BODY1 AND BODY2 OF A PUT 103, KNOCKED OUT AT 95, WITH A 90-DAY MATURITY

	Price
Body1	0.99
Body2	1.29

TABLE 12: COMPARISON OF THE PARAMETERS AND TOTAL VOLATILITY NUMBERS OF BODY1DOUBLE AND BODY1

	Brownian diffusion		Total volatility	
	Body1Double	Body1	Body1Double	Body1
Regime 1	9.55%	9.57%	11.69%	11.67%
Regime 2	6.44%	6.24%	32.23%	32.50%
Regime 3	2.41%	2.253%	11.88%	11.76%

	Jump size		Jump intensity	
	Body1Double	Body1	Body1Double	Body1
Regime 1 → Regime 2	−9.05%	−9.07%	0.2405	0.2370
Regime 2 → Regime 1	25.02%	62.67%	1.1279	0.0855
Regime 1 → Regime 3	2.79%	2.72%	3.3208	3.3951
Regime 3 → Regime 1	−3.07%	−3.17%	2.9882	2.9777
Regime 2 → Regime 3	65.12%	24.63%	0.0729	1.0944
Regime 3 → Regime 2	−22.68%	−22.66%	0.2025	0.2040

of analytical solutions). And there is just no way we could discriminate between the probability distributions of such models, by looking at the time series of the underlying. Volatility of volatility is hardly measurable. Not mentioning that every continuous model turns ‘discrete’ when solved numerically.

To the aesthetically minded, however, we can always suggest that Body can be further worked out into a full-bodied version that we call ‘Full Body’. There is no limitation to the number of volatility regimes we may want to consider, so a continuum of regimes is in theory possible. And there is no limitation either to the number of Poisson jumps occurring between regimes or within regimes. As we shift between Regime 1 and Regime 2, it could be a random draw whether the concurrent return jump is positive or negative, and of what size. And Regime 1 could be characterized, not just by a Brownian diffusion, but also by a collection of Poisson jumps occurring within that regime. Body is very flexible and can mimic any given model. Body is really anybody’s model. Or it can be everybody’s model at the same time (for instance Regime 1 can harbour a full local volatility model, Regime 2 a full Heston model, Regime 3 a full Merton

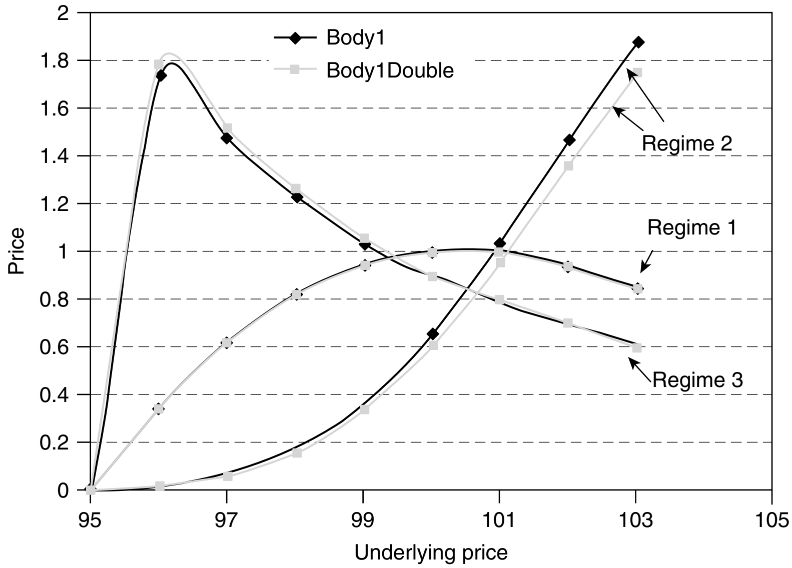


Figure 11: Price of the Put 103 down-and-out at 95 against the underlying price using Body1 and Body1Double parameters, in all three regimes

model, etc.). Yet Body will always be the dynamic, perfectly inter-temporally consistent, version of such 'mixings', by contrast to what has come to be known as the 'mixture' or 'ensemble' approach (Gatarek 2003, Johnson and Lee 2003). We should really be talking of 'superposition models' in our case rather than 'mixtures' (if we may borrow this crucial distinction from quantum mechanics), in order to distance ourselves from the unhappy 'ensemble' approach.

Full Body is in fact a general structure, a family of models rather than a model. The way people are used to think about regimes is in temporal succession. A regime of 'sticky-strike' smile behaviour can follow a regime of 'sticky-delta', etc. In the limit, we propose that you wake up every day in a state of stochastic superposition of such regimes (yet, we repeat, with total inter-temporal consistency and homogeneity), and that you watch for the market prices (one-touches, forward starting options, etc.) that will best determine the superposition. This may sound as the end of modelling to some people: 'Black–Scholes, Merton, Heston, SABR, Bates, sticky-strike, sticky-delta, etc., those are *models*, those are good names!' Indeed so. Our model deserves no name.

5.2 The hedging issue: optimal hedging

Let us now explore the other side of the smile problem, which we said was intimately linked to the pricing of exotic options, namely the discrepancy that may occur between the hedging strategies of two different models despite their being calibrated to the same vanilla smile. Before we do so, however, we have to introduce a fundamental concept. In all the smile models we've been considering (jump-diffusion, stochastic volatility, universal volatility) markets are incomplete. In other words, contingent claims cannot be replicated with the underlying alone. Indeed the Black–Scholes argument of self-financing, perfect dynamic hedging breaks down in the presence of jumps and/or stochastic volatility. Local volatility smile models try desperately to save the complete market paradigm, but are unrealistic precisely for this reason. They imply, for instance, that a barrier option is perfectly hedgeable with the underlying, no matter the volatility smile.

The other models evade the hedging issue altogether. They lay the stochastic process of the underlying in the risk-neutral world directly, and assume that option value is the discounted expectation of payoff under the risk-neutral measure.[8] While this guarantees that their option prices do not create instant arbitrage opportunities, they offer no guarantee that the option value is 'arbitraged' against the process of the underlying, in the Black–Scholes sense of 'volatility arbitrage'. In other words, you cannot hedge the option with the underlying, and 'lock' the option value at the inception of the trade, through subsequent dynamic action on the underlying. All you are offered in terms of hedging is the partial derivative with respect to underlying—never a hedge in the presence of jumps—or some 'external' bucketing of the volatility surface, which almost certainly contradicts the assumptions of the model.

What is needed is a theory of option pricing and hedging in incomplete markets. We will introduce the concept of 'optimal dynamic hedging'. By that we mean a self-financing dynamic portfolio, involving the underlying and the money account, which optimally replicates the derivative instrument, in some sense of 'optimality'. Our choice of criterion is the minimization of the variance of the P&L of the total portfolio. In other words, we draw on stochastic control theory to propose a self-financing dynamic hedging strategy for the derivative that lets you break even on average and guarantees that the distribution of your P&L is the most 'sharply peaked at zero' that can be. We *then* propose as a definition of 'derivative instrument value' the initial cost of the self-financing optimal hedging strategy. And we find that the initial cost of the optimal self-financing replicating portfolio has the property of a pricing operator, it therefore behaves like a risk-neutral probability (Henrotte 2002).

Because our optimal hedging takes place in the real world, and our risk-neutral probability measure is associated with optimal hedging, we are able to link our risk-neutral probability with the real probability. Calibration and pricing can take place in the risk-neutral world. Since our process parameters are inferred from the market prices of options, it is as if we were reverse-engineering the pricing operator from those traded prices, and reapplying it to find the unknown prices of some other options. However, when we start worrying about hedging the option, this can only take place in the real world and necessitates the transformation of the probability measure. This transformation requires an independent input: the market price of risk of the underlying, or its Sharpe ratio.

We also define the variable HERO (*Hedging Error at Replicating Optimum*) as the minimized standard deviation of the hedged portfolio. HERO is the measure of market incompleteness with regard to the given instrument. It may be large either because the underlying is 'incomplete' (large jumps, stochastic volatility...) or because the payoff is complex (exotics...). In the absence of jumps and stochastic volatility, our optimal hedge would indeed coincide with the Black–Scholes perfect hedge, and HERO would collapse to zero. Alternatively, the HERO of the underlying is trivially zero, no matter the stochastic process.

5.3 The 'true' smile dynamics

Let us now go back to our solemn question: 'Can we have control over the smile dynamics in homogeneous models?' At first blush, it seems the answer is no. Indeed, in space homogeneous models, Euler's theorem implies the following relation:

$$C = S(\partial C/\partial S) + K(\partial C/\partial K) \tag{1}$$

where C is the vanilla option price, S the underlying price and K the option strike.

C, S, K and $\partial C/\partial K$ being fixed for a fixed smile surface, this implies $\partial C/\partial S$, or Δ, is fixed. So it seems that two homogeneous models will agree on the option delta when they are calibrated to the same smile, no matter their respective stochastic structures. The Merton model, the Heston model, the Bates model, the SABR model when $\beta = 1$, will all produce the same vanilla option delta. Only space inhomogeneous models (like local volatility or universal volatility which involve an explicit relation between the diffusion coefficient and the underlying), can yield a different delta, because of the corrective term they introduce (see Equation 7).

But we wonder. Is $\Delta = \partial C/\partial S$ the right measure of smile dynamics? The answer is clearly 'yes' in the local volatility case where the underlying is the sole driving variable. However, in models involving another state variable, typically in stochastic volatility or universal volatility models, one cannot realistically move the underlying over an infinitesimal time interval and freeze the other variable. As volatility is correlated with the underlying, it is very likely that it moves too. Partial derivatives, such as $\partial C/\partial S$ and $\partial C/\partial \sigma$, capture the smile dynamics only partially. What we really need is the real time dynamics of the option price. In the local volatility case, we were able to apply the chain rule to get the real time delta. The question is, how can we apply the chain rule when volatility is an *indeterministic function* of the underlying, i.e. is correlated with it?

Before we try to answer what seems to be a challenging mathematical question, let us ask why we need the information on smile dynamics in the first place. Obviously in order to determine the number of underlying shares that should be held against the derivative, or in other words, to hedge. Only in the local volatility model does the notion of hedge coincide with the *mathematical* derivative with respect to underlying. In incomplete market models, there is no mathematically ready, i.e. *non-financial*, notion of hedge. We need to form the financial notion of hedge first (for instance optimal hedging in the sense of minimum variance), then work out the mathematics.

We claim that our 'optimal hedge' is the substitute of the notion of smile dynamics in incomplete market models. As a matter of fact, the whole notion of 'smile dynamics' appears to be muddled once the problem is set in the right frame. It is but a heritage of the local volatility model—the only place where it finds its meaning—and the whole comparison of smile behaviours between local volatility and stochastic volatility models appears to be ill-founded for that matter (you are not comparing apples to apples), if all that is meant is the partial derivative with respect to the underlying. So we might as well drop the whole notion of smile dynamics and get down to the hedge directly. What good is the notion of smile dynamics in jump-diffusion models anyway?

Recall that as the market is incomplete, we can only hedge optimally, and the HERO reflects how imperfect the hedge is. The optimal hedge that we produce already factors in the fact that the underlying may diffuse and jump, and that volatility may be stochastically varying, correlated with the underlying. In other words, it captures precisely the sense of 'total derivative' that mathematics alone was unable to give us. What seemed to be a purely mathematical question (How do we generalize the chain rule when the functions are indeterministic?) receives a *financial* answer once the real purpose of the question is recognized (i.e. hedging).

However, if your only interest in smile dynamics is to predict the future shape of the smile surface, and not necessarily to hedge, then your question may admit of a *probabilistic* answer—and a probabilistic answer only—outside the one-factor framework. Conditionally on the underlying trading at some level S at some future date t, you may want to know what the *expected* value of the vanilla options may be at that time, or in other words, what the smile surface may be *expected* to look like. *Expectation* here means probabilistic averaging (either risk-neutral or real) over the possible states of the other state variables(s), conditionally on the underlying being in state S. You should bear in mind, though, that this *expected value of the option* is a different notion to its *future price*, as it is purely mathematical and unrelated to replication.

Therefore the big question really becomes: 'Can two homogeneous models agree on the vanilla option prices, yet disagree on their *optimal hedging strategies*?' The answer is a resounding 'yes', as will be seen from the same Body examples as before. Recall the two instances of our calibration of Body to a full vanilla smile which had resulted in two different local minima, and consequently, in two different one-touch price structures. We weren't sure at the time whether the two solutions implied *different* smile dynamics, as they agreed on the option delta by homogeneity and by Euler's theorem. That they agree on the option price and delta, yet disagree on the optimal hedge (and HERO), can now be made explicit (see Table 13).

TABLE 13: BODY1 AND BODY2 OUTPUTS FOR A 107 CALL

Sharpe ratio	0.1		0.5		0.9	
	Body1	Body2	Body1	Body2	Body1	Body2
Price	1.0131	1.0189	1.0132	1.0189	1.0132	1.0189
HERO	1.4429	1.2609	1.4429	1.2608	1.2811	1.1238
Optimal hedge	**0.2217**	**0.1543**	**0.2177**	**0.1543**	**0.2409**	**0.1803**
Delta	0.2894	0.2774	0.2895	0.2774	0.2894	0.2774
Gamma	0.0531	0.0540	0.0531	0.0540	0.0531	0.0540

Only when additional information is included in the calibration, that is to say, information constraining the conditional transition probabilities, will the models agree on the 'smile dynamics'. And this is now meant both in the sense that they will agree on the exotic option pricing and that they will agree on the (optimal) hedging strategy. 'How do we gain control over the smile dynamics?' is therefore simply answered by controlling some exotic option price structures, typically the one-touches or forward starting options.

This is a general answer, not just specific to homogeneous models. Indeed, optimal hedging in incomplete markets is a general idea. It is just that the homogeneous models have helped us make our point more sharply, thanks to the 'surprising' feature due to Euler's theorem and to what seemed to be a loss of control over the option deltas. Also recall that Hagan and Blacher, who were arguing for control of the smile dynamics in inhomogeneous models, were not really taking into account what we have called the *true* smile dynamics.

In conclusion, there is no need to reintroduce inhomogeneity just for the sake of fitting a desired smile dynamics or a desired barrier option price structure. Henrotte's principle can thus be reiterated: **any departure from homogeneity should be the cause of great concern and should therefore be strongly motivated.**

We also find interesting that the answer to what seemed at first an 'innocent' yet very relevant question ('How do I control the smile dynamics in my smile model?') should require *the theory*

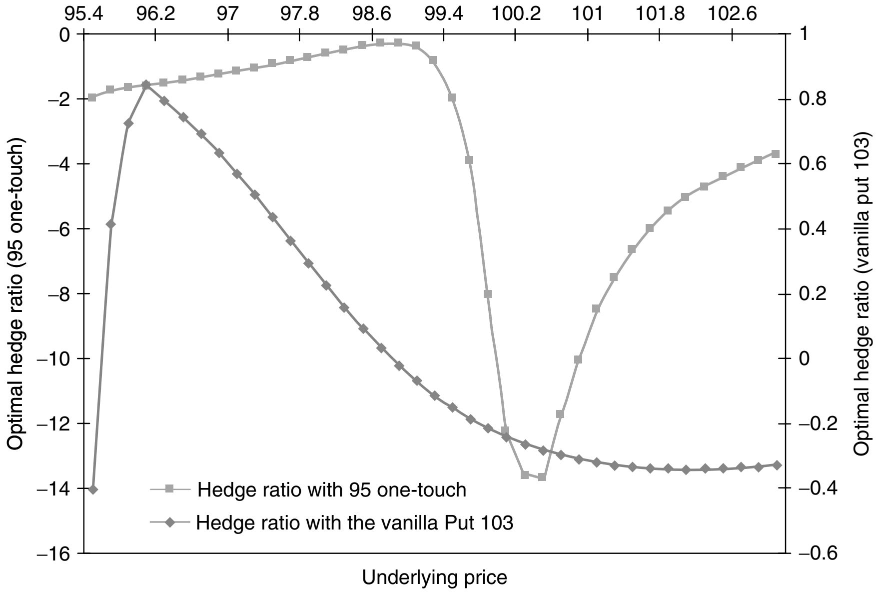


Figure 12: Optimal hedging ratios of the Put 103 KO 95 when either of the 95 one-touch or the vanilla Put 103 are used for dynamic hedging in combination with the underlying. The HERO (for $S = 100$) is 0.96 when no additional hedging instruments are used. It is 0.44 when the one-touch is used and 0.73 when the vanilla Put is used

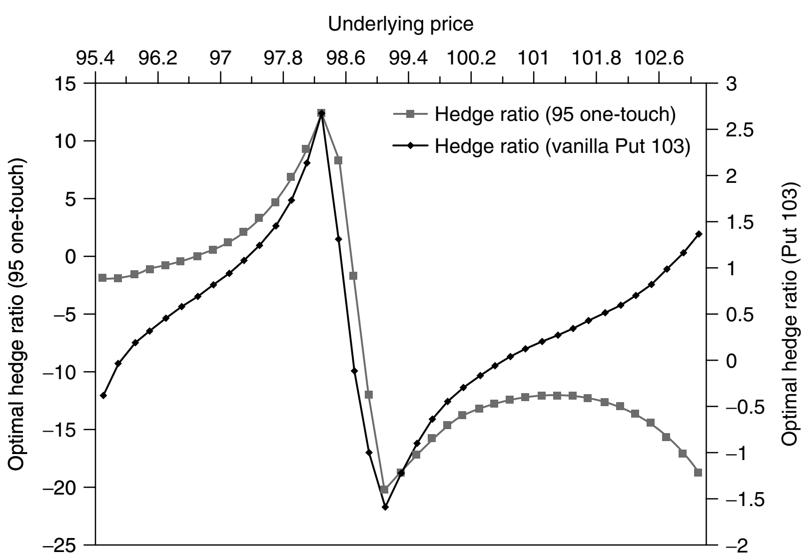


Figure 13: Optimal hedging ratios of the Put 103 KO 95 when both the 95 one-touch and the vanilla Put 103 are used for dynamic hedging in combination with the underlying. The HERO is now nearly zero over the whole range of spot prices

of hedging and pricing in incomplete markets as an indispensable intermediary step. Financially relevant questions can only be answered by relevant *financial* theory. The need to go back to the 'basics' is a very welcome conclusion, to say the least, at a time when quantitative finance seems to be wasting itself in sophisticated mathematical exercise, or even worse, in sophistical pseudo-models imported from foreign domains (e.g. the 'mixture of models', or 'ensemble', approach which cannot even afford an inter-temporal process, let alone a hedging rationale).[9]

6 Conclusion: generalizing Black–Scholes

We have made the case for the necessity of introducing exotic options in the calibration phase of the smile model, and the necessity of thinking in incomplete markets. *Smile dynamics* is more important than *smiles* as pricing and hedging are essentially *dynamic* concepts, and incomplete markets are omnipresent as smiles are essentially a departure from Black–Scholes. As a matter of fact, the smile problem really *begins* with the question of the smile dynamics and the question of the hedging rationale.[10] These questions had remained hidden from us as long as we remained blind to the degree of model dependence in the traditional models. Calibration to the exotics not only validates the right guess about the smile dynamics, but it allows us, thanks to an extension of the argument of optimal dynamic hedging in incomplete markets, to further *lock* the implied smile dynamics.

Indeed, stochastic control theory can be invoked again and our optimal dynamic, self-financing, hedging portfolios can be generalized to include other hedging instruments beside the underlying (see Figures 12 and 13). The price processes of the hedging instruments are independently available to us as the initial costs of their respective optimal hedging strategies involving the underlying alone. This guarantees that the price of the hedged derivative instrument can still be defined as the initial cost of the composite hedging portfolio, and be independent of the particular choice of hedging instruments other than the underlying. Dynamic multi-hedging of a derivative instrument allows the resulting HERO to be even smaller and the market to approach completeness. Typically a barrier option will be dynamically hedged with a combination of the underlying, a vanilla option, and a one-touch. A convertible bond will be hedged with a combination of the underlying, an equity option and a credit default swap. A complex cliquet will be hedged with the underlying and a combination of simple forward starting options.

Calibration should be calibration with a point. It achieves nothing on its own. Treating the vanillas, the one-touches, the forward starting options, or the credit default swaps, as alternative liquid instruments underlying our jump-diffusion/stochastic volatility process, and using them in the dynamic hedging of the given derivative instrument the same way that the underlying stock is traditionally used in Black–Scholes, is the right way to generalize Black–Scholes to the case of smiles. Making sure that the smile model prices the 'underlyings' in agreement with the market, and that it is calibrated to their dynamics, is in the end no different from saying that the Black–Scholes model prices *the* underlying in agreement with the market and is calibrated to its Brownian volatility.

When the hedging instruments are appropriately chosen, we expect the hedge ratios to be robust. Our hope is that they may even not depend on the particular model. In the end, a model is just a piece of machinery, 'cogs and wheels' that allow us to dynamically glue together the appropriate derivative instruments. If the relevant dynamics is properly captured (in other words, if the model is calibrated to the maximum relevant information), and if the hedging instruments are properly chosen, then the hedging strategy should more or less impose itself naturally. As a matter of fact, we found that it very often corresponded to the trader's, model-independent, intuition.

Thus we conclude with the *disappearance of the model*. If solving the smile problem means finding *the right tool*, then the directions we have suggested are indeed the right directions to pursue. This goes hand in hand with a constant awareness of the perfectibility and relativity of the tool. What we have proposed in this chapter is not so much the 'definitive smile model' as it is the definitive way to think *critically* about any model.

But if solving the smile problem means finding the absolutely true process and the absolute pricing algorithm, then we can safely declare: 'Nobody can solve the smile problem!'

FOOTNOTES & REFERENCES

1. We will later refer to the local volatility model(s) in the singular or the plural depending on whether we mean the theoretical principle or the particular numerical techniques.
2. As it was once argued in one thread of the Wilmott forums—Skew and forward volatilities (http://www.wilmott.com/messageview.cfm?catid=4&threadid=2551&FTVAR_MSGDBTABLE=).
3. See Henrotte (2004).
4. Of course, one-touches and forward starting options will not, in general, determine the smile dynamics completely (as Peter Carr once objected in a private communication). Think how large the number of degrees of freedom would be in the matrix of conditionals, if the problem were left completely non-parametric. Not mentioning the multiplication of that number by the number of spatial state variables. When we say the one-touches and the forward starting options help determine the smile dynamics, we mean it only relatively. Indeed, we, too, will have to depend on our particular choice of model for imposing the missing constraining structure. We need, however, to strike the right balance between the degree of structure imposed by the model and its ability to match the prices of contingent claims with very different payoff structures. Our solution is original both in the sense that it avoids the trap of non-parametric inference and that it is more flexible than the traditional parametric models.
5. See Lipton and McGhee (2002).
6. Total volatility includes the Brownian volatility and the volatility due to jumps, it is expressed by $V_i^2 = v_i^2 + \sum_k \lambda_i^k \left(y_i^k\right)^2 + \sum_j \lambda_{i\to j} \left(y_{i\to j}\right)^2$ where i denotes the regime for which the total volatility is calculated, j denotes the regimes i the underlying can migrate to, k denotes the jumps occuring within regime i. The rest of the notation is self-explanatory.
7. e.g. Dynamic SABR.
8. Typically, Lipton (2002) writes: '*As always, we can evaluate the price of an option as the discounted expectation of its payout under a risk-neutral measure. We set aside many important issues related to the incompleteness of the market in the presence of jumps and stochastic volatility, and use the risk-neutralized dynamics [...] throughout.*'
9. See Piterbag (2003) for a sweeping criticism of the ensemble approach.
10. See Ayache (2004).

■ Andersen, L. and Brotherton-Ratcliffe, R. (1998) The equity option volatility smile: an implicit finite-difference approach. *The Journal of Computational Finance*, 1(2), Winter.

■ Andersen, L. and Buffum, D. (2003) Calibration and implementation of convertible bond models. Working paper,Bank of America Securities.
■ Avellaneda, M. Carelli, A. and Stella, F. (2000) A Bayesian approach for constructing implied volatility surfaces through neural networks. *The Journal of Computational Finance*, 4(1), 83–107, Fall.
■ Ayache, E. (2001) A good smile is vital. *FOW*, May.
■ Ayache, E. (2004) The philosophy of quantitative finance. *Wilmott*.
■ Bakshi, G. and Cao, C. (2002) Risk-neutral jumps, kurtosis, and option pricing. Working paper, December.
■ Bates, D. (1996) Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options. *Review of Financial Studies*, 9(1), 69.
■ Blacher, G. (2001) A new approach for designing and calibrating stochastic volatility models for optimal delta-vega hedging of exotic options. Conference presentation at Global Derivatives, Juan-les-Pins.
■ Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637.
■ Bodurtha Jr., J. and Jermakyan, M. (1999) Nonparametric estimation of an implied volatility surface. *The Journal of Computational Finance*, 2(4), 29–60, Summer.
■ Coleman, T., Li, Y. and Verma, A. (1999) Reconstructing the unknown local volatility function. *The Journal of Computational Finance*, 2(3), 77–102, Spring.
■ Derman, E. (1999) Regimes of volatility. *Risk*, April.
■ Derman, E. and Kani, I. (1994) The volatility smile and its implied tree. *Quantitative Strategies Research,* Goldman Sachs.
■ Dupire, B. (1994) Pricing with a smile. *Risk*, 7(1), 18.
■ Gatarek, D. (2003) Libor market models with stochastic volatility. March.
■ Gatheral, J. (2003) Stochastic volatility and local volatility. *Lecture Notes*, Fall Term.
■ Hagan, P., Kumar, D., Lesniewski, A. and Woodward, D. (2002) Managing smile risk. *Wilmott*, p. 84.
■ Henrotte, P. (2002a) Dynamic mean variance analysis. Working paper, http://www.ito33.com/theory/henrotte_philippe-portfolios.pdf, July.
■ Henrotte, P. (2002b) Pricing kernels and dynamic portfolios. Working paper, http://www.ito33.com/theory/henrotte_philippe-kernels.pdf, August.
■ Henrotte, P. (2004) The case for time homogeneity. *Wilmott*, January.
■ Heston, S. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6: 327–43.
■ 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: 29–61.
■ Jackson, N., Sueli, E. and Howison, S. (1998) Computation of deterministic volatility surfaces. *The Journal of Computational Finance*, 2(2), 5–32, Winter.
■ Johnson, S. and Lee, H. (2003) Capturing the smile. *Risk*, March.
■ Kahale, N. (2003) An arbitrage-free interpolation of volatilities. Working paper, Hiram Finance, May.
■ Lagnado, R. and Osher, S. (1997) A technique for calibrating derivative security pricing models: numerical solution of an inverse problem. *The Journal of Computational Finance*, 1(1), 13, Fall.

■ Li, Y. (2001) A new algorithm for constructing implied binomial trees: does the implied model fit any volatility smile? *The Journal of Computational Finance*, 4(2), 69–95, Winter.
■ Lipton, A. (2002) The volatility smile problem. *Risk*, February.
■ Lipton, A. and McGhee, W. (2002) An efficient implementation of the universal volatility model. *Risk*, May.
■ Merton, R. (1996) Options pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1).
■ Piterbarg, V. (2003) Mixture of models: a simple recipe for a...hangover? Working paper, Bank of America, July.
■ Rubinstein, M. (1994) Implied binomial trees. *Journal of Finance*, (49), 771.
■ Tavella, D. and Klopfer, W. (2001) Implying local volatilities. *Wilmott*, August.

21
Philosophy of Finance: Definitive Smile Model: Part I

Elie Ayache

Why should we write about smile models? This is the question behind the question. For if the definitive smile model is not yet in sight, perhaps a definitive smile story is possible.

What is there more to say on the subject of smiles, and what is there to expect from reflection on the smile problem today? Could the answer be the further elaboration of the existing models? That is, could the future of our story be purely technological and one of taking up the technical complications one after the other, trying out jump-diffusion after the diffusion or stochastic volatility after local, deterministic volatility? Should one become a specialist in Laplace and Fourier transforms, and rank the models by classes of integrability, carefully selecting the functional form that promises the most exciting analytical gymnastics? And shouldn't then every quantitative analyst start worrying about the best way to promote his model, and how to best argue that his model must be the right one? Jump-diffusion may be better than diffusion because of the existence of large and rare moves in the underlying. Moreover, 'the ability of infinite-activity jump processes to capture both frequent small moves and rare large moves' may give us a further reason, as argued by Carr *et al*. (2002), to discard the diffusion component altogether in the light of statistical evidence for the fine structure of asset returns. Or perhaps the quant should worry about explaining market option prices as instantly observed rather than analyzing the underlying time series, and feel confident that his smile model is the right one when it is able to match the prices of, say, the barrier options, on top of the vanillas. This is the point of Lipton (2002b), and his defense of his 'universal volatility model' which mixes jump-diffusion and stochastic volatility.

Lipton's progress

'Why should we write about smiles anymore?' The answer may be that the only thing worth writing today is a review of existing smile models and their classification, *à bestiaire*, like the

French say. This is what Lipton (2002a) has attempted. A roadmap may indeed become desirable when the territory keeps expanding and the beasts look stranger and stranger, if only because it has the virtue of listing the known obstacles and the dark alleys. You read here and there that closed-form solutions cannot be had when there is correlation between the underlying and its volatility, or that calibration becomes a formidable task when the underlying is jumping and volatility is stochastic. A roadmap, however, is only as good as the vehicle that it is intended for, and it is clear that Lipton's intended vehicle is the closed-form, or semi-closed-form solution, when it can be had. On the other hand, there is something disheartening about the very idea of a 'complete guide', and that is that such a guide is only as good as its vintage. Apart from proposing a smile model for every taste and culture (jump-diffusion, stochastic volatility, local volatility), and updating us on the last fashionable trend, what is to be gained from such a listing over and above its comprehensiveness and good taste? What is the real advance? And when the 'universal volatility model', which Lipton offers for the finale of his catalog, is itself interpreted back into the series as the latest model produced, or in other words, the last model of the list which naturally beats all the others in terms of complexity and number of parameters, might we not fear that the truly different argument that Lipton brings up at some point, namely the capacity of this model to match the market price of barrier options, may look very remote? If matching the barrier option prices is such a definitive argument, then why bother with the history and lineage of smile models anymore? Lipton's dramatic build-up makes it all sound as if smile modeling finally reached an age when jumps can be safely combined with stochastic volatility and the appropriate Fourier transform successfully obtained, and as if—surprise!—the market concurred, in celebration of that age and in acknowledgement of that maturity, with the gift of his agreement on the barrier option prices. Are we to believe that empirical agreement with the barrier option prices was just waiting for this last advance in smile theory and smile model design, and for an advance with precisely that parametric form? Or was the 'universal volatility model' somehow encoded in the market? What other reasons are we offered for this agreement apart from pure luck, or just the supernatural argument that the 'universal volatility model' is next on our list and has got to address, for that sole reason, the next unsolved problem which is the matching of barrier option prices? Instead of showing us what it takes for a smile model to match the barrier option prices over and above its matching the vanillas, and why the 'universal volatility model' has it—for that would provide a real next stage for our thinking about smiles—Lipton lapses into metaphysics and retains as the only benefit from agreement with barrier option prices the fact that his model must somehow be distinguished on that list and be true; period. To put it differently: what if barrier option prices contained additional information that has to be calibrated in the model independently of the vanillas? When Lipton's point is precisely that two smile models can agree on the vanillas and disagree on the barriers, might we not fear that the market dynamics may evolve the next day and imply a different price structure for the barriers, given a certain price structure of the vanillas? Might the 'universal volatility model' not fall into disgrace itself despite its superlative name, and the market favors shift to a more encompassing model still, or perhaps revert to an older and simpler model? Again, what is missing here is a theory of that extra step, or new frontier in smile intelligence, which the barrier options represent, and empirical evidence is just not good enough an argument.

A meta-model

It may sound as if I am hinting at some kind of superior model, or meta-model, which could see what is happening when the 'universal volatility model' manages to match the barriers and the

Heston model, or the local volatility model, does not. It would be a meta-model both in the sense that it embeds the models of lower rank as specific instances and that it provides a critique of those models. But then the 'universal volatility model' was supposed to be just that! As a matter of fact, 'universal volatility' is not just a model, but a whole family of jump-diffusion models combined with stochastic volatility and it can reproduce, at one extreme, a local volatility model or a pure diffusion with variable diffusion coefficient, and at the other, pure stochastic volatility or a Heston-like model. It can even assume a pure jump process. So while Lipton has proposed the all-encompassing, overarching model we are looking for, he has not provided the critique. And the reason is that he paused at the meta-level only to rush down into the one instance of his meta-model which afforded an analytical solution, yet differed enough from Heston or local volatility to deserve the name of 'universal volatility model'. More importantly, Lipton has not taken the extra step of calibrating his model to the barrier options a priori. The vanilla implied volatility surface is all we have to go along with in order to establish the parameters of the model, and agreement with the barriers is then checked a posteriori. We are left with the flat conclusion that his vanilla-calibrated 'universal volatility model' predicts the right price for the barrier options, for example the double-no-touch, for no other reason than that it hits the right balance between the local volatility model which underestimates it and the Heston model which overestimates it.

Beginning of the smile problem

So at best, Lipton's 'universal volatility model' looks like an adjustment or a refinement of pre-existing models. 'The market is subtler than you think, so the story goes. It doesn't exactly behave like any of the standard smile models you've been using, local volatility, stochastic volatility, pure jump, but somewhat in the middle. And what else did you expect? The road to barrier options has been concealed from the known tracks, but it definitely exists on our roadmap. This is precisely the road that you can see now opening up in the middle. It may be a little harder to journey because of the additional parameters and the tougher Fourier transform, but it is there alright.' The reason I dispute this statement is that the 'universal volatility model' is not in the middle, but is supposed to be above. It shouldn't really belong on the roadmap, but in the bureau revising the roadmap. And the barrier option pricing problem is supposed to be the key to our real thinking about smiles, and not just fall as an additional item on the list of things that one model can do and the others cannot. As long as the smile problem was one of accounting for the implied volatility smile of the vanillas, alternative explanations could compete on the same level and their relative advantages be compared. One explanation, for instance, proposed that the coefficient of the Brownian diffusion was not constant in the plane but varied according to Dupire's formula. Another claimed that the diffusion process was overlaid by Poisson jumps, whose size and intensity we would have to determine by calibration. Yet another assumed that volatility was stochastic itself and correlated with the underlying. Or indeed an explanation mixing all three kinds of process, diffusion, jumps, and stochastic volatility, could be considered in turn. Any of these explanations was as good as another as long as the challenge was to describe a certain way that reality should be, for a vanilla smile to be the consequence. You may have had issues like overfitting or underfitting and questions about the right number of degrees of freedom and whether or not you should allow for term structure of the parameters, but these were technical issues.

However, the smile problem enters a new phase—or rather, it rises to a new level—when it becomes one of trusting the proposed model for the hedging strategy one should follow. In fact, the vanilla option deltas produced by the competing models differ largely from one model to another. For instance, the local volatility model predicts that the option price will evolve with the underlying in such a way that the smile moves in the opposite direction to the underlying movement. See Hagan (2002) for the analysis and criticism of that phenomenon. By contrast, a stochastic volatility model, like Heston or SABR, predicts that the smile evolves in the same direction as the underlying, or in other words, that implied volatility is a function of the option moneyness. From descriptive metaphysics the problem has now moved to speculative metaphysics. The question is no longer to explain the present smile, but to predict its evolution. In fact, the smile problem, as I like to call it, really begins here. Indeed, any of the static descriptive explanations of the vanilla smile is as good as another, and for that matter, no better than straightforward spline interpolation! No one would have a problem with the smile, and no one would need a smile model, if the problem was just the pricing of vanilla options under implied volatility smiles. Similarly, the smile problem really begins with the question of pricing the barrier options. Since there is no way we could interpolate a Black–Scholes implied volatility number for the barrier option from the vanilla implied volatility surface—should we interpolate at the strike of the barrier option or at its barrier?—we definitely need a smile model to form its price. And, surely enough, the vanilla-calibrated competing smile models yield different barrier option prices, just as they yield different vanilla option deltas. Speculative metaphysics back again.

The term 'metaphysics', however, seems to suggest that the truth must be lying somewhere behind the phenomenon, only we have no other way to get hold of it at present but to speculate about it. And now Lipton's article on 'universal volatility' and Hagan's article on SABR appear as ways of re-embedding speculative metaphysics into descriptive metaphysics, by enlarging the view. Both authors argue that their model describes reality accurately, only they draw a more comprehensive picture of reality. Their picture now includes, beside the vanilla smile, the observed barrier option prices in Lipton's case, and the observed vanilla option deltas, in Hagan's. Both authors seem to ignore the possibility that the barrier pricing problem, or the vanilla delta problem, may be adding a new dimension to the smile problem rather than a new side to reality, and that both the barrier price structure and the vanilla delta structure may change, for a fixed vanilla smile. What would Lipton do if empirical barrier option prices moved closer to the pattern predicted by a local volatility model and away from his 'universal volatility model'? And what would Hagan do if empirical vanilla option deltas started reflecting a sticky-strike situation rather than sticky-delta? Would they discard their models? As a matter of fact, different delta behaviors and different barrier price structures have been empirically observed at different times and at different places. See Derman's paper on volatility regimes. In the end, Lipton and Hagan may be just reflecting a reality specific to their particular market, foreign exchange options in Lipton's case, and interest rate options in Hagan's. (Even worse, they may be reflecting a self-fulfilling prophecy.) In other words, it may very well be that the vanilla option deltas have to be calibrated into the model independently, the same way the barrier option prices should be. Indeed, we show in another paper that the two problems are intimately linked, and that they hinge on the dynamics of the smile.

'What is there more to say about smiles?' And the answer should be: everything! Any smile model leaving untouched the question of the hedging strategy of the vanillas, or the question of the pricing rationale of the barrier options, has not even begun to address the smile problem. And it will not do to argue that the vanilla hedges have consistently been observed to be such and such in my market, or that the barrier option prices happen to be such and such. The fallacy which consists in arguing for the validity of a given smile model ('universal volatility', SABR) on the grounds of the empirical confirmation of the option delta or the barrier price it produces, is worse than leaving these problems untouched. For it suggests that all there is to expect from the delta or the barrier price is a distinction and a confirmation in retrospect, and that agreement with the market delta or the market barrier price is the last word in the smile model contest. It suggests that the problem is over, when we claim that it has only begun and that the delta and the barrier are the first things we should really be writing about.

A departure from Black–Scholes

We may essentially define smiles as a radical departure from Black–Scholes. And we do not mean it in the sense that the observed vanilla prices differ from the Black–Scholes uniform implied volatility. For all we know, the Black–Scholes formula may have never existed. It may have been altogether unimaginable that rehedging could take place continuously or that transactions could be costless. And Black and Scholes, for that matter, may have had to come up with a more complex formula, which implied itself a 'volatility smile' relative to the usual formula. What we mean when we say that smiles are a radical departure from Black–Scholes, is that smiles really begin when we are no longer able to apply what is really important in Black–Scholes. And what is really important in Black–Scholes is not the formula or the usual simplifying assumptions (continuous, frictionless trading) but the following two things: the dynamic hedging idea and the idea of translating the option price into an implied volatility number. These are the true inventions which have revolutionized our way of dealing with options.

Now translating the vanilla option price into an implied volatility number is still possible under smiles: interpolation does that nicely. Therefore the smile problem doesn't begin here. The smile problem begins as soon as we depart from Black–Scholes and no longer have a fix on either the hedge or the representative volatility number. It begins with the problem of the vanilla option delta and the problem of the barrier option price representation. This is the reason why any smile model that manages to match the market prices of the vanilla options, but offers no guarantee that it will match their market deltas, or that it will match the market prices of barrier options, really ends before the beginning of the smile problem. Lipton and Hagan offer no such guarantee. They are just lucky enough that their model agrees with their market reality. The only way to offer the guarantee is to build it into the model. This is a call to a voluntarist and active attitude. And now we can understand why Lipton and Hagan, who had no means of controlling the barrier option price structure, or the vanilla option delta structure, beyond the matching of the vanilla option prices, could offer no other guarantee than just the passive belief in the existence of a truth out there and the correspondence of their models with that truth.

Thinking after Black–Scholes

It will be my contention that Lipton and Hagan are the last representatives of a philosophical tradition that misinterpreted the meaning of the Black–Scholes model and the significance of its teaching. Philosophy and interpretation wouldn't worry us much if they had no effect on the science and remained confined in the preserve of reflection and meditation. It doesn't really matter to the Black–Scholes model how we interpret it or philosophize about it. The philosophy of Black–Scholes (and more generally, the philosophy of derivative pricing) will be shown to matter, however, to the science and practice that followed Black–Scholes, namely the smiles. The smile problem, as we face it today and insofar as it begins today, is essentially a philosophical problem. Or so I will argue. To really think about smiles, one has first to learn to think about Black–Scholes, and only then will one know how to think after Black–Scholes. Since smiles are the radical departure from Black–Scholes, anyone misinterpreting Black–Scholes will misconstrue the way of departing from it, and therefore will misunderstand smiles.

'Departure from Black–Scholes' and 'thinking after Black–Scholes' have to be understood in the two senses of the terms. Smiles depart from Black–Scholes in the sense that they radically differ from it and that they take in, basically, anything that constitutes a breach of the Black–Scholes paradigm. (And it is a big world out there! Jumps can induce volatility smiles, but so can stochastic volatility, and default risk, and firm leverage, and discrete hedging, and transactions costs. Any realistic derivative pricing model is a smile model, really.) And smiles depart from Black–Scholes in the sense that they issue from it and that they are its generalization. Or rather, they will strike us as the true generalization of Black–Scholes, once we identify the strands in Black–Scholes that should really be generalized. Likewise, thinking about smiles is thinking after Black–Scholes: thinking what is next and taking up where Black–Scholes has left off. And it is thinking after Black–Scholes: thinking in the style of Black–Scholes and following its teaching.

Now the reason why the tradition that followed Black–Scholes has misinterpreted it and missed the thrust of the whole new science that was being born, is that it thought of the Black–Scholes model as the description of some physical reality. It thought Black and Scholes were literally after the lognormal distribution of asset returns and presumed that the Black–Scholes model was false when it was faced with the first deviation from the predicted option prices, i.e. smiles. Yet this tradition had nothing to say about the widespread continued use of the Black–Scholes pricing formula in spite of the obvious inaccuracy of the underlying theoretical model, or about the apparent ease with which traders just went ahead and plugged in a different implied volatility number for every different option they wished to price. This phenomenon is sometimes referred to as 'the robustness, or the resilience, of the Black–Scholes model'. The traditional criticism explained it away as being just a consequence of the simplicity and intuitive appeal of the Black–Scholes model. And while it set out to find the theoretical substitute of Black–Scholes, it argued that people using Black–Scholes were doing something they shouldn't really do. The situation was one of essential tension between the longevity and increasing popularity of the Black–Scholes model (still the textbook model, still the option pricing benchmark) and the increasingly smaller odds that the 'true' model may finally be found. For once correspondence to truth had become a requirement and once the alternative to a false Black–Scholes had been philosophically reduced to the quest for the true smile model and nothing but the true smile model, this quest could not just stop at the first step and simply match the vanilla smile. The true model had to tell all the truth: it had to match the barrier options, it had to produce the right hedges (witness the arguments from Lipton and Hagan), and last but not least, it had to appeal to practitioners, not just academics, and satisfy them that it was every bit as robust and functional as Black–Scholes.

Never before in the sciences had we witnessed such a big gap and such a great conflict between the endeavor of the theorist looking for the true model and the behavior of the practitioner using the model. While the continued 'falsification' (to use a Popperian term) of every successive model had done nothing but excite the theorist and exacerbate his belief that the truth must be lying ahead—forever lying ahead, never in the present model, always in the next—and while it had done nothing but precipitate an escalation of arguments from his part instead of making him consider a radical alternative, the practitioner had no such exacting concerns and enjoyed a much greater freedom of movement, literally making the truth rather than finding it, and making the market in the vanillas and the exotics. Not mentioning that the exotic structures themselves were being made up every day and that they created new markets every day. So are we to believe that truth is just sitting there, waiting for the true model to find it, and that this moment of truth will then at once embrace all the exotic structures that have come about or will have to come about? Or might the theorist argue that truth is itself a relative and forever shifting notion and that he doesn't mind reiterating the whole nested sequence of models every time a new class of exotic structures is introduced, no matter whether the new sequence and the new 'history of science' contradicted the previous ones? And how would we account for the transition regimes, where truth is not yet itself an established notion and the only truth-maker is everybody's guess about what to count as an arbitrage?

REFERENCES

■ Carr, P., Géman, H., Madan, D., and Yor, M. (2002) The fine structure of asset returns: An empirical investigation. *Journal of Business*, 75(2), 305–32.
■ Derman, E. (1999) Regimes of volatility. *Risk*, April.
■ Hagan, P., Kumar, D., Lesniewski, A., and Woodward, D. (2002) Managing smile risk. *Wilmott*, p. 84.
■ Lipton, A. (2002a) The volatility smile problem. *Risk*, February.
■ Lipton, A. and McGhee, W. (2002b) An efficient implementation of the universal volatility model. *Risk*, May.

22
Philosophy of Finance: Definitive Smile Model: Part II

Elie Ayache

Black–Scholes is right and significant only to the extent that it is not true. In this chapter we look at what arises from a discussion of true versus right.

The so-called *nesting of models* seems to be the most recent fashionable exercise with respect to the truth project in quantitative analysis. For instance Bakshi and Cao (2003) argue in a recent empirical study that a double-jump option pricing model taken from Duffie *et al*. (2000), which improves on the previous model (Bates 1996, which in turn improved on the model before (Heston 1993) in adding underlying jumps to stochastic volatility) in offering the possibility of adding volatility jumps correlated to the underlying jumps, performs better both in matching the in-sample vanilla options and in pricing the out-of-sample options. Not forgetting Lipton, who argues that the 'universal volatility model' which improves on all of the local volatility, jump-diffusion and stochastic volatility models in mixing all their characteristics, performs better in terms of pricing the exotic options. The impression one gets from this argumentative zeal is one of a converging sequence of models, bound to reach the final nest where truth must be lying. I wonder how the exotics would fare in the double-jump model, and whether a bit of its vanilla explanatory power should not be sacrificed in order to account for the barriers. At any rate, turning one's attention to the exotics would imply a break in the thrust of the argument of Bakshi and Cao and in the push to the truth about the vanillas. The same break occurs in Lipton, if one starts worrying about the possibility of a change in the price structure of the barriers for a given vanilla price structure.

The right alternative to a false Black–Scholes model, I think, is not to look for the true substitute but to drop the whole metaphysical notion of *truth in an option pricing model*. Although Black–Scholes is clearly false in the sense of not corresponding to empirical fact about option prices, I want to argue that it is *valid*, in a new enlarged sense of validity. If the Black–Scholes model is still being used by traders and practitioners all round, then it has got to be valid, and this validity has got to be independent of the true–false dichotomy. In Chapter 21 I stressed that the two important things in Black–Scholes are the notion of dynamic hedging and the synthesizing of option prices in the implied volatility number. What the first really did is allow the traders to link option value to a concrete rule of action. The necessity to update the option *delta* with the

Black–Scholes formula and to rebalance the hedge every time the underlying moved was the real reason why the Black–Scholes model was used in effect. Given the freedom that the option trader enjoyed in setting up the implied volatility number, it should have been suspicious from the start that the Black–Scholes formula might have had another motivation. Option valuation was made *effective* through the concrete link that the delta provided with the underlying. And valuing options effectively was no longer a matter of applying the pricing formula punctually and theoretically, but required from the trader that he consistently monitored and *followed* his option trade. Surely enough he could lock the option value by neutralizing his delta exposure, but this very move suggested that he should get back to his option trade every now and then, and gave meaning to this constant revisiting. Following the rule of delta hedging and delta rebalancing inscribed the option value in a chain of coordinated actions instead of leaving it as a theoretical result on the trader's spreadsheet. It turned the option into a *relational* concept which now involved the whole functional relationship with the underlying and no longer stood alone in abstraction.[1]

As for the second important thing—the expression of option prices in terms of implied volatility numbers—it provided the option traders with a new and very efficient language. Traders were able to relate to the (implied) volatility they were buying, selling, or trading off, more easily than they did to the naked option prices, and the Black–Scholes model which inspired all this with its flat volatility assumption never was an impediment to the actual multiplication of implied volatility numbers across the option chains, and to the capacity of the language to adapt itself to situations pretty much at variance with the original Black–Scholes world.

The philosophical point I am trying to make, which will help banish truth altogether as an irrelevant category in our case, is that the Black–Scholes model has bestowed *meaning* on options and on option trading through the *algorithm of delta hedging* and the *language of implied volatility*, and that meaning is not of such nature as to fall under the scrutiny of metaphysical truth or to be deemed true or false. The realm of meaning, also known as the realm of validity, is philosophically distinct from the realm of truth. And I claim that Black–Scholes is valid because meaning is a much richer category than truth. Think that we can use language meaningfully, and for that matter compose poems and create metaphors, or propose scientific theories and advance wild interpretations of the physical world, without necessarily speaking truthfully.

True, the trader may be flying in the face of the theoretical Black–Scholes model when he updates both the underlying price *and* the implied volatility number in his formula and rebalances his hedge accordingly, still it cannot be claimed that he therefore represents a falsity. On the contrary, delta hedging is the right thing to do—this is the main lesson from Black–Scholes—and the trader doing it shows a perfect understanding of the meaning of options, even though he may not know the truth about them. Now Black–Scholes may be the *wrong* model to use for delta hedging in the presence of smiles (given jumps and stochastic volatility), and we may be willing to start looking for the *right* model. The fact remains that the valid dichotomy here is the right–wrong dichotomy, not the true–false. Right and wrong do not partition the space of reasons the same way that true and false partition the space of facts. You may be doing or thinking the right thing for the right kind of reason, without there necessarily being a fixed reference against which you can justify your action or your thinking.

When Lipton and Hagan argue that their model is the right model because it gives the right barrier option prices or produces the right option hedges, their argument is a truth claim in disguise, not a validity claim. Hence our criticism. Indeed the barrier option prices and the vanilla option hedges are the fixed reference they relate to and the ultimate truth-maker they seek. By contrast, what would be a valid and much richer model (valid in our extended sense of validity, richer in the sense that validity and meaning precisely exceed truth)—in a word, what we would call the *right* model—is a model where you would explicitly include the barrier option prices and the vanilla option deltas in the calibration. And we say that the model is right (and not just true) because it depends on no external, 'fixed' reference which may very well vary the next day, but incorporates the *variability* of the reference itself. It turns the concept of the *right smile model* into a relational and relative concept: the model will give the *right* barrier option prices, or the right delta hedges, simply because it relies on a law of logic (even a syllogism) not on a matter of fact; it will give the right price for the barrier options when it is calibrated to the right barrier options, and it will produce the right delta hedges when it is calibrated to produce them.

The significance of Black–Scholes

To really assess the significance of the Black–Scholes model and what it meant to both the science and the history of the science, and to fully appreciate what it takes to really think about Black–Scholes, think what our thinking would look like if Black–Scholes were true. If hedging were continuous and if we lived in a world of underlying Brownian motion with constant (non-stochastic) volatility, options would be redundant. They wouldn't exist except by name. All that would remain to do is to buy or sell the underlying (and you would definitely find somebody prepared to take the opposite bet, in this perfectly random world), or to invest an initial fee in a certain combination of the underlying and the riskless bond, to be able to run a self-financing dynamic trading strategy which may result, for instance, in being long the underlying at a certain level, at a certain date, if it trades above that level at that date, or in being short it at a certain level, if it trades below that level. Conversely, you may sell that combination for a certain fee, and run the opposite self-financing dynamic trading strategy in order to preserve that fee, no matter the outcome of the underlying at maturity. Options would exist only by name, and the underlying would be the only thing worth buying or selling or trading in ever more sophisticated strategies. And should it turn out that options must exist, by some metaphysical decree, beyond the mere naming of those self-financing dynamic strategies, why would anyone buy them or sell them? Wouldn't everybody agree on their initial value and their outcome? Since you can personally perfectly replicate any contingent payoff, all you would need is a party to your trades in the underlying. No option market *per se* would come to exist.

What we are really saying is that if Black–Scholes were true, what Black–Scholes would really have to say ('Options exist and they can be traded. You can buy them, sell them and even hedge them, etc.') would not be true or false, or right or wrong. It would really be unsayable. Black–Scholes would really have nothing to say. Fortunately, Black–Scholes is not true, and this is why we have something instead of nothing. As Alberto Coffa would say, '*the unsayable is not true, but there is something it is right about*'. And what Black–Scholes is right about is precisely this, in Black–Scholes, which looks outside the closed formula and outside the complete market paradigm and its tautological consequence for options. Black–Scholes is precisely right in having bestowed on options and option markets the meaning that we have been talking about. And what is so amazing about the Black–Scholes model, and definitely distinguishes it, and the history of the science that will follow from it, from any other history of science, is the extraordinary philosophical pressure that is exerted on it the minute it is subjected to reflection. Never before has a model or a theory or a framework been so finished and so closed on paper and so eager to crack open under philosophical questioning. *Black–Scholes is right insofar as it is not true*. Anything meaningful, and historical, and thought-provoking that Black–Scholes may have to say, has nothing to do with Black–Scholes and everything to do with smiles. Options exist (independently of their hedging strategies of course: otherwise how could we even start talking of hedging *them*?) only insofar as the hedge is not perfect and there is leeway in the choice of the hedging strategy. And option markets exist only insofar as the language of implied volatility has got more than one word.

The process of objectification and the true science in Black–Scholes

Now we can see why the two most significant strands in Black–Scholes, the dynamic hedging story and the implied volatility story, are the true things worth generalizing and reflecting upon. Once the philosophical picture is set in the right frame, and the Black–Scholes model is no longer followed for the something *true* but for the something *right* that it has to say, we understand where all the robustness comes from. Black–Scholes seems so inseparable from options and option talk because it was first to insert the option value into the algorithm of delta hedging and the language of implied volatility. It thereby granted options a special kind of being: a 'being *objective*' which is at once more significant than 'being a name' and far more robust (far less risky and unstable) than 'being true'. The Black–Scholes model turned options into scientific and linguistic *objects*. The original theory may be simplistic and we may have abused of the original single-worded language, the fact remains that the delta hedging algorithm has contributed to the process of 'objectification' of the option (as a neo-Kantian would say), or in other words to 'the construction of its being *as object* through conceptual determination'. As for the implied volatility language, it has provided an effective translation of option prices and option markets. When traders relate to the Black–Scholes model, they do not really care whether the model, as model, is true, and whether it relates to some transcendent reality. All they care about are the *objects* and the functional relations between them.[2] They care about the option and the delta as inter-operative concepts. What I am trying to say is that the scientific moment that one should try to capture in Black–Scholes is the moment of the *sending* of the strands (Brownian motion as the simplest way of breathing life and time value in the option, implied volatility as the single knob to calibrate the model with, and dynamic replication as the operative rule), not the moment when the strands meshed with each other in a single fateful knot, and gave us the closed-form formula and the complete market, thus ending philosophical thinking before it even started.

The science that we would like to capture and nurture in Black–Scholes is not the bit that argues from Brownian motion to the continuous perfect replication to the Black–Scholes PDE to the analytical formula. For this is only a clever mixture of stochastic calculus and no-arbitrage principle, which takes advantage of the continuous-path property of Brownian motion and the ability of a continuously rebalanced self-financing portfolio to be immune against the Brownian innovation. This 'pencil-and-paper' Black–Scholes does not really interest us. What science we see in Black–Scholes is the part that gave birth to the history of the science. It is the part concerned with the objectification that we talked about earlier. As our neo-Kantian philosopher of science would go on to say: 'The *fact of science* is the fact of objectification at its most developed stage, and philosophy's task is to grasp the categories of objectification governing scientific development.'[3] The part in Black–Scholes corresponding to the 'fact of science' is no doubt the part that makes options objective, not the one that makes them redundant—the part that initiates philosophical thinking, not the one that evacuates it. It is the part which literally occurs outside the closed-form formula and speaks distinctively of options, of option hedges, and option implied volatility smile. It wouldn't cross the mind of the first option trader anyhow that options may be redundant, and that they may not have their own market, quite independently of the underlying.

The history of the science

Now think that the original motivation of Black–Scholes and Merton was to provide the traders with tools to rationally price and possibly arbitrage those options! Surely enough, the assumption of lognormal distribution of asset returns must have seemed to them the most attractive initial step to get the problem going. And how surprised Black and Scholes must have been to find, as a result of this single step, that options and option markets were being dismissed completely! If the history of the science were to be rewritten, Black and Scholes would really have to keep their paper hidden from the eyes of the public. Any option pricing and hedging model would have been good for publishing, except the original Black–Scholes! This is why we've been urging that, although the Black–Scholes model is undoubtedly a historic finding and although the Black–Scholes language still permeates the totality of our conceptual dealings with options—even the word 'smiles' implicitly refers to Black–Scholes—we should really think of options as if Black–Scholes had never existed. This means we should not try to save the complete market paradigm at all costs, or look preferentially for models which result in analytical pricing formulae. All these things, all these worries and the research programs that they spawned, should really disappear from our sight when we interpreters set new eyes on the science and the history of the science. Now that we know about jump-diffusion and stochastic volatility and discrete hedging and transactions costs and incomplete markets, and now that the actual history of the science has shown us the *necessity* to know about all this, how could a thin coincidence such as perfect replication under Brownian motion and the analytical tractability of the Black–Scholes model matter any longer? How could such a contingent fact even strike us as something worth mentioning in our rewriting process? History may originate from a degenerate case, but the history of a science, in the sense of the philosophical rewriting and grounding of the science, may not.

The trouble with Black–Scholes, however, is that history (real history, not the philosopher's) could not have been written otherwise, and perhaps this singular fate is the most interesting part of the interpretive story. Indeed, how could Black and Scholes resist publishing their paper, and how could the public not welcome it instantly,[4] when it allowed the exact pricing (and hedging) of European options, and freed the valuation of contingent claims from the question of risk preferences? And how could option traders resist talking of implied volatility instead of option price, when Black–Scholes had shown them how to get rid of any other determinant of value through delta hedging, and left them with volatility as the only measure of cheapness and dearness of options? Or rather, once delta hedging had eliminated first-order market risk, the option trader was left with a sense of option cheapness and dearness directly related to the risks he knew Black–Scholes *could not cover in reality*: gamma risk and vega risk. And here you can see the creation of Black–Scholes starting to act contrary to Black–Scholes. For what did the option traders do once they got hold of the Black–Scholes formula and measured the ease with which it allowed them to connect the value of an option and a volatility number? Create volatility smiles! So what Black–Scholes has done in the end is provide the option traders with the best way to talk and to act outside Black–Scholes!

The option language

And what would it matter anyhow if traders spoke an unruly and 'unregimented' language? Isn't that always the case with natural languages? Accusing the traders of inconsistency on the grounds of their multi-volatility talk is the same as arguing that every competent speaker, in every natural language, can sooner or later be forced into a contradiction, if the questioner pushes her strictly from antecedent to consequent and the black and white logic of truth tables is applied to her utterances. Is it not precisely the lesson of the philosophy of language (at least after Wittgenstein) that logic shall not be the judge of language but the other way around, and that both the notions of logic and 'matter of fact', so dear to the heart of the empiricist, shall themselves be relative to a language? Must there not be, as Richardson says, 'a structure inherent in any language that provides the framework within which that language can *first* express any matters of fact'? Is the whole notion of 'matter of fact' not itself 'internal to a logico-linguistic framework'? And 'but for a prior specification of a logical structure', wouldn't the very notion of 'fact' be itself without sense? Language is robust precisely in the sense that one should not hold reality (or logic) fixed and try to vary the propositions of the language in order to come up with a falsity (or a contradiction) which would invalidate the language. On the contrary, one should hold language to be valid no matter what—for it is language that makes the world not the world that makes the language—and come to accept the fact that the contexts of utterance and their background logic may themselves be changing, in a word, that the world may itself be changing and that every speaker may be tacitly aware of it, every time some surface utterance strikes one as false or other-worldly. Language is not true or false, and it is not supposed to be a faithful picture of the facts. 'Our words do not carve up nature at the joints' and nature does not care with how many tenses we may conjugate our verbs. Language is robust in the sense that it allowed us to travel safely through our thousands of years of evolution and to survive its many changing worlds. It is robust in the sense that we are able to have revolutions which overturn our most deeply entrenched conceptual schemes (such as Gödel's theorem, Quantum Mechanics), yet we make sense of them *with language*. It is robust in the sense that we are able to do philosophy, to be reflective, etc.

Black–Scholes is valid and robust precisely in the sense that natural language is. Once we agree that what is meaningful and significant in Black–Scholes does not lie on the side of the lognormal assumption and the Black–Scholes formula—not on the side of complete markets and perfect attainability of the contingent payoffs—but on the side of the dynamic relations that Black–Scholes has helped establish between the option, the hedge, the implied volatility representation and the movements of the underlying, we stop thinking of Black–Scholes as a theoretical model and start thinking about it as language. So long as the trader knows what he is doing, it doesn't matter whether he changes the implied volatility parameter between two option trades, or between two delta readjustments. He is competent in that language. The option has first to exist, and second we have to start thinking of hedging it. It is the privilege of no option pricing model to bestow existence on options, even less so to rob them of their existence like the theoretical Black–Scholes does. No option pricing model[5] is even entitled to establish the prices of the vanilla options in place of their own market. We're not even sure that a smile model may be entitled to price the *exotic* options without somehow relying on their own market. All an option pricing model is welcome to do is provide the trader with a language, or in other words, a coherent way of travelling across the vagaries of the option world and of surviving its overturns. A language: that is, a conceptual scheme, a *Weltanschauung*.

And this general remark applies to the Black–Scholes model as well! Not the theoretical, vacuous Black–Scholes, but the meaningful, critical[6] Black–Scholes. There is indeed a sense in which Black–Scholes is the first smile model! Don't the option traders speak of Black–Scholes implied volatilities, and use Black–Scholes hedges, in real-life option markets? And aren't they confident of what they're doing because they know everybody speaks the same language? The only practical use of Black–Scholes, after all, is to let you travel from point A to point B. And you are basically OK travelling with Black–Scholes so long as the delta (possibly adjusted to account for the change in implied volatility) takes care of first-order market risk,[7] and so long as you are confident that everybody will still be speaking the Black–Scholes language at point B (having made the same implied volatility correction that you did). Black–Scholes, and for that matter any option pricing model, is only here after all to ensure safe travel through a price difference, not to quote an absolute price. Physics is essentially differential. And the key concept in every option model should be the option delta, not the option theoretical value.

The option delta

Delta is the *critical* concept here, in the two senses of the term. The trader's risk *critically* depends on the delta of his option position; in other words delta is the one important variable he will have to worry about after the inception of the trade. And delta is a critical concept in the sense that the entirety of our philosophical *critique* of option models has hinged on it so far, from our first contending that the smile problem really begins with the problem of the delta (or equivalently the problem of barrier option pricing), to our firm belief that the rule of delta hedging and rebalancing is the dispenser of scientific objectivity, to our conclusion that Black–Scholes is right and valid and meaningful to the extent that the Black–Scholes delta *should not* make the option redundant. Delta is the philosophically fertile notion and the entry point to all the different strands we've been exploring. First of all, it is the delta hedging idea which has made the language of implied volatility effective. Second, you can look at the delta from any side you wish, depending on your philosophical inclination. When it is part of the Black–Scholes derivation and formal theory is your sole concern, delta hedging leads to the strict option pricing formula that you know: it gives you the *law* that option prices obey. When it is viewed against the neo-Kantian background of relational concepts and the priority of objectivity over truth, delta embodies the operative *rule* which *conceptually determines* the option. When it is reinserted in the pragmatic context of actual hedging, which necessitates a real-time trader and his actual sense of opportunity, delta is your pathway to freedom: you can decide to over-hedge or under-hedge, optimally hedge, hedge discretely, not hedge at all, etc.

All of this hints at the idea that, once the options and their market are given and firmly given (contrary to their evaporation by Black–Scholes magic), we should first and foremost preoccupy ourselves with the hedge. Hedging is the key; option value is only a derivative notion. As for the option price, it is the purely opportunistic, almost political, variation of the option value. Hedging is the critical concept. For instance, we will show later that proposals to correlate default risk with the process of the underlying equity, which may sometimes go as far as invoking grandiloquent structural models of the firm, have as sole motivation the ability to produce higher equity deltas for the convertible bonds than in standard models, or indeed to generate such deltas for the straight debt, exactly like the trader would expect in real life. In this case as in many others, it is matching the delta that is the heart of the matter. Nobody really cares about the full underlying process, or the even less observable capital structure of the issuing firm.

FOOTNOTES & REFERENCES

1. We are here reiterating the neo-Kantian view of concept formation. In Alan Richardson's (1998) words: 'Perhaps the most important aspect of the neo-Kantian project [...] is the lesson it took from the development of pure mathematics and mathematical physics in the nineteenth century. For the neo-Kantians, this development exhibits a new type of concept formation that makes evident the functional nature of objective concepts and stands opposed to the traditional notion of concept formation via the process of abstraction.'

2. Again, we are echoing the neo-Kantian view of scientific objects as individuated via their relations to one another. They are neither bundles of subjective impressions (following the philosophical doctrine of idealism) nor pieces of an absolute reality (following the philosophical doctrine of realism). 'This view', writes Richardson, 'clearly contrasts with any naïve realism that speaks of objective knowledge as objective not because of the systematic interrelations of the objects in the system but by relations to transcendent objects outside the system. Similarly, it is inconsistent with any idealism that founds objectivity in the subjective experience of any one individual, or that denies objectivity to knowledge in general.'

3. Steven Galt Crowell, *Husserl, Heidegger and the Space of Meaning*, Northwestern University Press, 2001.

4. I am being guilty of history-rewriting, even here. For it appears that Black and Scholes had difficulty getting their 1973 paper accepted for publication. But this serves my interpretive point exactly. What I have called the 'significance' and the 'meaning' of Black–Scholes was not first apparent to the editor's eye. He could not have guessed the history that was to follow—the history of volatility trading—and the generations of volatility traders that were to come, from what looked, on the surface, like a simple analytical formula. In a word, he could not have guessed about the later philosophy of Black–Scholes, the part which came after Black–Scholes and that we have aptly identified with the smiles. Like I said, the 'fact of science' in Black–Scholes does not belong to the 1973 Black–Scholes.

5. From now on, 'option pricing model' will mean 'smile model', because we said Black–Scholes shouldn't really exist and smiles are the only thing there is.

6. Critical in the sense of the Kantian critique of metaphysics, and the subsequent construction of the objectivity of scientific theories.

7. Of course you will not be OK if jumps in the underlying occur between A and B. But we group jump risk under 'gamma risk' and it is second-order in this sense, not in the sense of the magnitude of the loss.

■ Bakshi, G. and Cao, C. (2003) Risk-neutral kurtosis, jumps and option pricing: Evidence from 100 most actively traded firms on the CBOE. Working Paper, Smith School of Business, University of Maryland.

■ Bates, D.S. (1996) Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies*, 9(1), Winter, 69–107.

■ Crowell, S.G. (2001) *Husserl, Heidegger and the Space of Meaning*, Northwestern University Press.

■ Duffie D., Pan, J. and Singleton, K.J. (2000) Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68, 1343–1376.

■ 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(2), 327–343.

■ Richardson, A. (1998) *Carnap's construction of the world: The aufbau and the emergence of logical empiricism*, Cambridge University Press.

23
A Perfect Calibration! Now What?

Wim Schoutens,* Erwin Simons and Jurgen Tistaert****

We show that several advanced equity option models incorporating stochastic volatility can be calibrated very nicely to a realistic option surface. More specifically, we focus on the Heston Stochastic Volatility model (with and without jumps in the stock price process), the Barndorff-Nielsen–Shephard model and Lévy models with stochastic time. All these models are capable of accurately describing the marginal distribution of stock prices or indices and hence lead to almost identical European vanilla option prices. As such, we can hardly discriminate between the different processes on the basis of their smile-conform pricing characteristics. We therefore are tempted to apply them to a range of exotics. However, due to the different structure in path behaviour between these models, the resulting exotics prices can vary significantly. It motivates a further study on how to model the fine stochastic behaviour of assets over time.

1 Introduction

Since the seminal publication of the Black–Scholes model in 1973, we have witnessed a vast effort to relax a number of its restrictive assumptions. Empirical data show evidence for non-normal distributed log-returns together with the presence of stochastic volatility. Nowadays, a battery of models is available which captures non-normality and integrates stochastic volatility. We focus on the following advanced models: the Heston Stochastic Volatility Model (Heston 1993) and its generalization allowing for jumps in the stock price process (see e.g. Bakshi *et al*. 1997), the Barndorff-Nielsen–Shephard model introduced in Barndorff-Nielsen and Shephard (2001) and Lévy models with stochastic time introduced by Carr (*et al*. 2001). This class of models is built

Contact addresses: *K.U. Leuven, Celestijnenlaan 200 B, B-3001 Leuven, Belgium and **ING SWE, Financial Modeling, Marnixlaan 24, B-1000 Brussels, Belgium.
E-mail: Wim.Schoutens@wis.kuleuven.ac.be, Erwin.Simons@ing.be, Jurgen.Tistaert@ing.be
The views expressed in this chapter are those of the authors and do not necessarily reflect the positions of their employers.

out of a Lévy process which is time-changed by a stochastic clock. The latter induces the desired stochastic volatility effect.

Section 2 elaborates on the technical details of the models and we state each of the closed-form characteristic functions. The latter are the necessary ingredients for a calibration procedure, which is tackled in section 3. The pricing of the options in that framework is based on the analytical formula of Carr and Madan (1998). We will show that all of the above models can be calibrated very well to a representative set of European call options. Section 4 describes the simulation algorithms for the stochastic processes involved. Armed with good calibration results and powerful simulation tools, we will price a range of exotics. Section 5 presents the computational results for digital barriers, one-touch barriers, lookbacks and cliquet options under the different models. While the European vanilla option prices hardly differ across all models considered, we obtain significant differences in the prices of the exotics. The chapter concludes with a formal discussion and gives some directions for further research.

2 The models

We consider the risk-neutral dynamics of the different models. Let us shortly define some concepts and introduce their notation.

Let $S = \{S_t, 0 \leq t \leq T\}$ denote the stock price process and $\phi(u, t)$ the characteristic function of the random variable $\log S_t$, i.e.,

$$\phi(u, t) = E[\exp(\mathrm{i}u \log(S_t))].$$

If for every integer n, $\phi(u, t)$ is also the nth power of a characteristic function, we say that the distribution is *infinitely divisible*. A Lévy process $X = \{X_t, t \geq 0\}$ is a stochastic process which starts at zero and has independent and stationary increments such that the distribution of the increment is an infinitely divisible distribution. A *subordinator* is a non-negative non-decreasing Lévy process. A general reference on Lévy processes is Bertoin (1996), for applications in finance see Schoutens (2003).

The risk-free continuously compounded interest rate is assumed to be constant and denoted by r. The dividend yield is also assumed to be constant and denoted by q.

2.1 The Heston Stochastic Volatility model

The stock price process in the Heston Stochastic Volatility model (HEST) follows the Black–Scholes SDE in which the volatility is behaving stochastically over time:

$$\frac{\mathrm{d}S_t}{S_t} = (r - q)\mathrm{d}t + \sigma_t \mathrm{d}W_t, \quad S_0 \geq 0,$$

with the (squared) volatility following the classical Cox–Ingersoll–Ross (CIR) process:

$$\mathrm{d}\sigma_t^2 = \kappa(\eta - \sigma_t^2)\mathrm{d}t + \theta\sigma_t \mathrm{d}\tilde{W}_t, \quad \sigma_0 \geq 0,$$

where $W = \{W_t, t \geq 0\}$ and $\tilde{W} = \{\tilde{W}_t, t \geq 0\}$ are two correlated standard Brownian motions such that $\mathrm{Cov}[\mathrm{d}W_t \mathrm{d}\tilde{W}_t] = \rho \mathrm{d}t$.

The characteristic function $\phi(u,t)$ is in this case given by Heston (1993) or Bakshi *et al.* (1997):

$$\begin{aligned}\phi(u,t) &= E[\exp(\mathrm{i}u\log(S_t))|S_0,\sigma_0^2]\\ &= \exp(\mathrm{i}u(\log S_0+(r-q)t))\\ &\quad\times\exp(\eta\kappa\theta^{-2}((\kappa-\rho\theta u\mathrm{i}-d)t-2\log((1-g\mathrm{e}^{-dt})/(1-g))))\\ &\quad\times\exp(\sigma_0^2\theta^{-2}(\kappa-\rho\theta\mathrm{i}u-d)(1-\mathrm{e}^{-dt})/(1-g\mathrm{e}^{-dt})),\end{aligned}$$

where

$$d=((\rho\theta u\mathrm{i}-\kappa)^2-\theta^2(-\mathrm{i}u-u^2))^{1/2}, \tag{1}$$

$$g=(\kappa-\rho\theta u\mathrm{i}-d)/(\kappa-\rho\theta u\mathrm{i}+d). \tag{2}$$

2.2 The Heston Stochastic Volatility model with jumps

An extension of HEST introduces jumps in the asset price (Bakshi *et al.* 1997). Jumps occur as a Poisson process and the percentage jump-sizes are lognormally distributed. An extension also allowing jumps in the volatility was described in Knudsen and Nguyen-Ngoc (2003). We opt to focus on the continuous version and the one with jumps in the stock price process only.

In the Heston Stochastic Volatility model with jumps (HESJ), the SDE of the stock price process is extended to yield:

$$\frac{\mathrm{d}S_t}{S_t}=(r-q-\lambda\mu_J)\mathrm{d}t+\sigma_t\mathrm{d}W_t+J_t\mathrm{d}N_t,\quad S_0\geq 0,$$

where $N=\{N_t,t\geq 0\}$ is an independent Poisson process with intensity parameter $\lambda>0$, i.e. $E[N_t]=\lambda t$. J_t is the percentage jump size (conditional on a jump occurring) that is assumed to be lognormally, identically and independently distributed over time, with unconditional mean μ_J. The standard deviation of $\log(1+J_t)$ is σ_J:

$$\log(1+J_t)\sim\mathrm{Normal}\left(\log(1+\mu_J)-\frac{\sigma_J^2}{2},\sigma_J^2\right).$$

The SDE of (squared) volatility process remains unchanged:

$$\mathrm{d}\sigma_t^2=\kappa(\eta-\sigma_t^2)\mathrm{d}t+\theta\sigma_t\mathrm{d}\tilde{W}_t,\quad \sigma_0\geq 0,$$

where $W=\{W_t,t\geq 0\}$ and $\tilde{W}=\{\tilde{W}_t,t\geq 0\}$ are two correlated standard Brownian motions such that $\mathrm{Cov}[\mathrm{d}W_t\mathrm{d}\tilde{W}_t]=\rho\mathrm{d}t$. Finally, J_t and N are independent, as well as of W and of $\tilde{W}$.

The characteristic function $\phi(u,t)$ is in this case given by:

$$\begin{aligned}\phi(u,t) &= E[\exp(\mathrm{i}u\log(S_t))|S_0,\sigma_0^2]\\ &= \exp(\mathrm{i}u(\log S_0+(r-q)t))\\ &\quad\times\exp(\eta\kappa\theta^{-2}((\kappa-\rho\theta u\mathrm{i}-d)t-2\log((1-g\mathrm{e}^{-dt})/(1-g))))\\ &\quad\times\exp(\sigma_0^2\theta^{-2}(\kappa-\rho\theta\mathrm{i}u-d)(1-\mathrm{e}^{-dt})/(1-g\mathrm{e}^{-dt})),\\ &\quad\times\exp(-\lambda\mu_J\mathrm{i}ut+\lambda t((1+\mu_J)^{\mathrm{i}u}\exp(\sigma_J^2(\mathrm{i}u/2)(\mathrm{i}u-1))-1)),\end{aligned}$$

where d and g are as in (1) and (2).

2.3 The Barndorff-Nielsen–Shephard model

This class of models, denoted by BN–S, were introduced in Barndorff-Nielsen and Shephard (2001) and have a comparable structure to HEST. The volatility is now modeled by an Ornstein Uhlenbeck (OU) process driven by a subordinator. We use the classical and tractable example of the Gamma–OU process. The marginal law of the volatility is Gamma-distributed. Volatility can only jump upwards and then it will decay exponentially. A co-movement effect between up-jumps in volatility and (down)-jumps in the stock price is also incorporated. The price of the asset will jump downwards when an up-jump in volatility takes place. In the absence of a jump, the asset price process moves continuously and the volatility decays also continuously. Other choices for OU-processes can be made, we mention especially the Inverse Gaussian OU process, leading also to a tractable model.

The squared volatility now follows an SDE of the form:

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + dz_{\lambda t}, \tag{3}$$

where $\lambda > 0$ and $z = \{z_t, t \geq 0\}$ is a subordinator as introduced before.

The risk-neutral dynamics of the log-price $Z_t = \log S_t$ are given by

$$dZ_t = (r - q - \lambda k(-\rho) - \sigma_t^2/2)dt + \sigma_t dW_t + \rho dz_{\lambda t}, \qquad Z_0 = \log S_0,$$

where $W = \{W_t, t \geq 0\}$ is a Brownian motion independent of $z = \{z_t, t \geq 0\}$ and where $k(u) = \log E[\exp(-uz_1)]$ is the cumulant function of z_1. Note that the parameter ρ is introducing a co-movement effect between the volatility and the asset price process.

As stated above, we chose the Gamma–OU process. For this process $z = \{z_t, t \geq 0\}$ is a compound–Poisson process:

$$z_t = \sum_{n=1}^{N_t} x_n, \tag{4}$$

where $N = \{N_t, t \geq 0\}$ is a Poisson process with intensity parameter a, i.e. $E[N_t] = at$ and $\{x_n, n = 1, 2, \ldots\}$ is an independent and identically distributed sequence, and each x_n follows an exponential law with mean $1/b$. One can show that the process $\sigma^2 = \{\sigma_t^2, t \geq 0\}$ is a stationary process with a marginal law that follows a Gamma distribution with mean a and variance a/b. This means that if one starts the process with an initial value sampled from this Gamma distribution, at each future time point t, σ_t^2 is also following that Gamma distribution. Under this law, the cumulant function reduces to:

$$k(u) = \log E[\exp(-uz_1)] = -au(b + u)^{-1}.$$

In this case, one can write the characteristic function (Barndorff-Nielsen *et al.* 2002) of the log price in the form:

$$\begin{aligned}\phi(u, t) &= E[\exp(\mathrm{i}u \log S_t)|S_0, \sigma_0] \\ &= \exp\left(\mathrm{i}u(\log(S_0) + (r - q - a\lambda\rho(b - \rho)^{-1})t)\right)\end{aligned}$$

$$\times \exp\left(-\lambda^{-1}(u^2 + \mathrm{i}u)(1 - \exp(-\lambda t))\sigma_0^2/2\right)$$
$$\times \exp\left(a(b - f_2)^{-1}\left(b \log\left(\frac{b - f_1}{b - \mathrm{i}u\rho}\right) + f_2\lambda t\right)\right),$$

where

$$f_1 = f_1(u) = \mathrm{i}u\rho - \lambda^{-1}(u^2 + \mathrm{i}u)(1 - \exp(-\lambda t))/2,$$
$$f_2 = f_2(u) = \mathrm{i}u\rho - \lambda^{-1}(u^2 + \mathrm{i}u)/2.$$

2.4 Lévy models with stochastic time

Another way to build in stochastic volatility effects is by making time stochastic. Periods with high volatility can be looked at as if time runs faster than in periods with low volatility. Applications of stochastic time change to asset pricing go back to Clark (1973), who models the asset price as a geometric Brownian motion time-changed by an independent Lévy subordinator.

The Lévy models with stochastic time considered in this chapter are built out of two independent stochastic processes. The first process is a Lévy process. The behaviour of the asset price will be modeled by the exponential of the Lévy process suitably time-changed. Typical examples are the Normal distribution, leading to the Brownian motion, the Normal Inverse Gaussian (NIG) distribution, the Variance Gamma (VG) distribution, the (generalized) hyperbolic distribution, the Meixner distribution, the CGMY distribution and many others. An overview can be found in Schoutens (2003). We opt to work with the VG and NIG processes for which simulation issues become quite standard.

The second process is a stochastic clock that builds in a stochastic volatility effect by making time stochastic. The above-mentioned (first) Lévy process will be subordinated (or time-changed) by this stochastic clock. By definition of a subordinator, the time needs to increase and the process modeling the rate of time change $y = \{y_t, t \geq 0\}$ needs also to be positive. The economic time elapsed in t units of calendar time is then given by the integrated process $Y = \{Y_t, t \geq 0\}$ where

$$Y_t = \int_0^t y_s \mathrm{d}s.$$

Since y is a positive process, Y is an increasing process. We investigate two processes y which can serve for the rate of time change: the CIR process (continuous) and the Gamma–OU process (jump process).

We first discuss NIG and VG and subsequently introduce the stochastic clocks CIR and Gamma–OU. In order to model the stock price process as a time-changed Lévy process, one needs the link between the stochastic clock and the Lévy process. This role will be fulfilled by the characteristic function enclosing both independent processes as described at the end of this section.

2.4.1 NIG Lévy process An NIG process is based on the Normal Inverse Gaussian (NIG) distribution, NIG(α, β, δ), with parameters $\alpha > 0$, $-\alpha < \beta < \alpha$ and $\delta > 0$. Its characteristic function is given by:

$$\phi_{NIG}(u; \alpha, \beta, \delta) = \exp\left(-\delta\left(\sqrt{\alpha^2 - (\beta + \mathrm{i}u)^2} - \sqrt{\alpha^2 - \beta^2}\right)\right).$$

Since this is an infinitely divisible characteristic function, one can define the NIG process $X^{(NIG)} = \{X_t^{(NIG)}, t \geq 0\}$, with $X_0^{(NIG)} = 0$, as the process having stationary and independent NIG distributed increments. So, an increment over the time interval $[s, s+t]$ follows a NIG$(\alpha, \beta, \delta t)$ law. An NIG process is a pure jump process. One can relate the NIG process to an Inverse Gaussian time-changed Brownian motion, which is particularly useful for simulation issues (see section 4.1).

2.4.2 VG Lévy process The characteristic function of the VG(C, G, M), with parameters $C > 0$, $G > 0$ and $M > 0$ is given by:

$$\phi_{VG}(u; C, G, M) = \left(\frac{GM}{GM + (M-G)\mathrm{i}u + u^2}\right)^C.$$

This distribution is infinitely divisible and one can define the VG process $X^{(VG)} = \{X_t^{(VG)}, t \geq 0\}$ as the process which starts at zero, has independent and stationary increments and where the increment $X_{s+t}^{(VG)} - X_s^{(VG)}$ over the time interval $[s, s+t]$ follows a VG(Ct, G, M) law. In Madan *et al*. (1998) it was shown that the VG process may also be expressed as the difference of two independent Gamma processes, which is helpful for simulation issues (see section 4.2).

2.4.3 CIR stochastic clock Carr *et al*. (2001) use as the rate of time change the CIR process that solves the SDE:

$$\mathrm{d}y_t = \kappa(\eta - y_t)\mathrm{d}t + \lambda y_t^{1/2}\mathrm{d}W_t,$$

where $W = \{W_t, t \geq 0\}$ is a standard Brownian motion. The characteristic function of Y_t (given y_0) is explicitly known (see Cox *et al*. 1985):

$$\begin{aligned}\varphi_{CIR}(u, t; \kappa, \eta, \lambda, y_0) &= E[\exp(\mathrm{i}uY_t)|y_0] \\ &= \frac{\exp(\kappa^2\eta t/\lambda^2)\exp(2y_0\mathrm{i}u/(\kappa + \gamma\coth(\gamma t/2)))}{(\cosh(\gamma t/2) + \kappa\sinh(\gamma t/2)/\gamma)^{2\kappa\eta/\lambda^2}},\end{aligned}$$

where

$$\gamma = \sqrt{\kappa^2 - 2\lambda^2\mathrm{i}u}.$$

2.4.4 Gamma–OU stochastic clock The rate of time change is now a solution of the SDE:

$$\mathrm{d}y_t = -\lambda y_t\mathrm{d}t + \mathrm{d}z_{\lambda t}, \tag{5}$$

where the process $z = \{z_t, t \geq 0\}$ is as in (4) a compound Poisson process. In the Gamma–OU case the characteristic function of Y_t (given y_0) can be given explicitly.

$$\begin{aligned}\varphi_{\Gamma-OU}(u; t, \lambda, a, b, y_0) &= E[\exp(\mathrm{i}uY_t)|y_0] \\ &= \exp\left(\mathrm{i}uy_0\lambda^{-1}(1 - e^{-\lambda t})\right. \\ &\qquad \left. + \frac{\lambda a}{\mathrm{i}u - \lambda b}\left(b\log\left(\frac{b}{b - \mathrm{i}u\lambda^{-1}(1 - e^{-\lambda t})}\right) - \mathrm{i}ut\right)\right).\end{aligned}$$

2.4.5 Time-changed Lévy process Let $Y = \{Y_t, t \geq 0\}$ be the process we choose to model our business time (remember that Y is the integrated process of y). Let us denote by $\varphi(u; t, y_0)$ the characteristic function of Y_t given y_0. The (risk-neutral) price process $S = \{S_t, t \geq 0\}$ is now modeled as follows:

$$S_t = S_0 \frac{\exp((r-q)t)}{E[\exp(X_{Y_t})|y_0]} \exp(X_{Y_t}), \tag{6}$$

where $X = \{X_t, t \geq 0\}$ is a Lévy process. The factor $\exp((r-q)t)/E[\exp(X_{Y_t})|y_0]$ puts us immediately into the risk-neutral world by a mean-correcting argument. Basically, we model the stock price process as the ordinary exponential of a time-changed Lévy process. The process incorporates jumps (through the Lévy process X_t) and stochastic volatility (through the time change Y_t). The characteristic function $\phi(u, t)$ for the log of our stock price is given by:

$$\begin{aligned}\phi(u,t) &= E[\exp(\mathrm{i}u \log(S_t))|S_0, y_0] \\ &= \exp(\mathrm{i}u((r-q)t + \log S_0)) \frac{\varphi(-\mathrm{i}\psi_X(u); t, y_0)}{\varphi(-\mathrm{i}\psi_X(-\mathrm{i}); t, y_0)^{\mathrm{i}u}},\end{aligned} \tag{7}$$

where

$$\psi_X(u) = \log E[\exp(\mathrm{i}uX_1)];$$

$\psi_X(u)$ is called the characteristic exponent of the Lévy process.

Since we consider two Lévy processes (VG and NIG) and two stochastic clocks (CIR and Gamma–OU), we will finally end up with four resulting models abbreviated as VG–CIR, VG–OUΓ, NIG–CIR and NIG–OUΓ.

Because of (time)-scaling effects, one can set $y_0 = 1$, and scale the present rate of time change to one. More precisely, we have that the characteristic function $\phi(u, t)$ of (7) satisfies:

$$\begin{aligned}\phi_{NIG-CIR}(u, t; \alpha, \beta, \delta, \kappa, \eta, \lambda, y_0) &= \phi_{NIG-CIR}(u, t; \alpha, \beta, \delta y_0, \kappa, \eta/y_0, \lambda/\sqrt{y_0}, 1), \\ \phi_{NIG-\Gamma OU}(u, t; \alpha, \beta, \delta, \lambda, a, b, y_0) &= \phi_{NIG-\Gamma OU}(u, t; \alpha, \beta, \delta y_0, \lambda, a, b y_0, 1), \\ \phi_{VG-CIR}(u, t; C, G, M, \kappa, \eta, \lambda, y_0) &= \phi_{VG-CIR}(u, t; C y_0, G, M, \kappa, \eta/y_0, \lambda/\sqrt{y_0}, 1), \\ \phi_{VG-\Gamma OU}(u, t; C, G, M, \lambda, a, b, y_0) &= \phi_{VG-\Gamma OU}(u, t; C y_0, G, M, \lambda, a, b y_0, 1).\end{aligned}$$

Actually, this time-scaling effect lies at the heart of the idea of incorporating stochastic volatility through making time stochastic. Here, it comes down to the fact that instead of making the volatility parameter (of the Black–Scholes model) stochastic, we are making the parameter δ in the NIG case and the parameter C in the VG case stochastic (via the time). Note that this effect does not only influence the standard deviation (or volatility) of the processes, also the skewness and the kurtosis are now fluctuating stochastically.

3 Calibration

Carr and Madan (1998) developed pricing methods for the classical vanilla options which can be applied in general when the characteristic function of the risk-neutral stock price process is known.

Let α be a positive constant such that the αth moment of the stock price exists. For all stock price models encountered here, typically a value of $\alpha = 0.75$ will do fine. Carr and Madan then showed that the price $C(K, T)$ of a European call option with strike K and time to maturity T is given by:

$$C(K, T) = \frac{\exp(-\alpha \log(K))}{\pi} \int_0^{+\infty} \exp(-\mathrm{i}v \log(K))\varrho(v)\mathrm{d}v, \tag{8}$$

where

$$\varrho(v) = \frac{\exp(-rT)E[\exp(\mathrm{i}(v - (\alpha + 1)\mathrm{i}) \log(S_T))]}{\alpha^2 + \alpha - v^2 + \mathrm{i}(2\alpha + 1)v} \tag{9}$$

$$= \frac{\exp(-rT)\phi(v - (\alpha + 1)\mathrm{i}, T)}{\alpha^2 + \alpha - v^2 + \mathrm{i}(2\alpha + 1)v}. \tag{10}$$

Using fast Fourier transforms, one can compute within a second the complete option surface on an ordinary computer. We apply the above calculation method in our calibration procedure and estimate the model parameters by minimizing the difference between market prices and model prices in a least-squares sense.

The data set consists of 144 plain vanilla call option prices with maturities ranging from less than one month up to 5.16 years. These prices are based on the volatility surface of the Eurostoxx 50 index, having a value of 2461.44 on October 7th, 2003. The volatilities can be found in Table 4. For the sake of simplicity and to focus on the essence of the stochastic behaviour of the asset, we set the risk-free interest rate equal to 3% and the dividend yield to zero. The results of the calibration are visualized in Figure 1 and Figure 2 for the NIG–CIR and the BNS model respectively; the other models give rise to completely similar figures. Here, the circles are the market prices and the plus signs are the analytical prices (calculated through formula (8) using the respective characteristic functions and obtained parameters).

In Table 1 one finds the risk-neutral parameters for the different models. For comparative purposes, one computes several global measures of fit. We consider the root mean square error ($rmse$), the average absolute error as a percentage of the mean price (ape), the average absolute error (aae) and the average relative percentage error ($arpe$):

$$rmse = \sqrt{\sum_{options} \frac{(\text{Market price} - \text{Model price})^2}{\text{number of options}}}$$

$$ape = \frac{1}{\text{mean option price}} \sum_{options} \frac{|\text{Market price} - \text{Model price}|}{\text{number of options}}$$

$$aae = \sum_{options} \frac{|\text{Market price} - \text{Model price}|}{\text{number of options}}$$

$$arpe = \frac{1}{\text{number of options}} \sum_{options} \frac{|\text{Market price} - \text{Model price}|}{\text{Market price}}$$

In Table 2 an overview of these measures of fit is given.

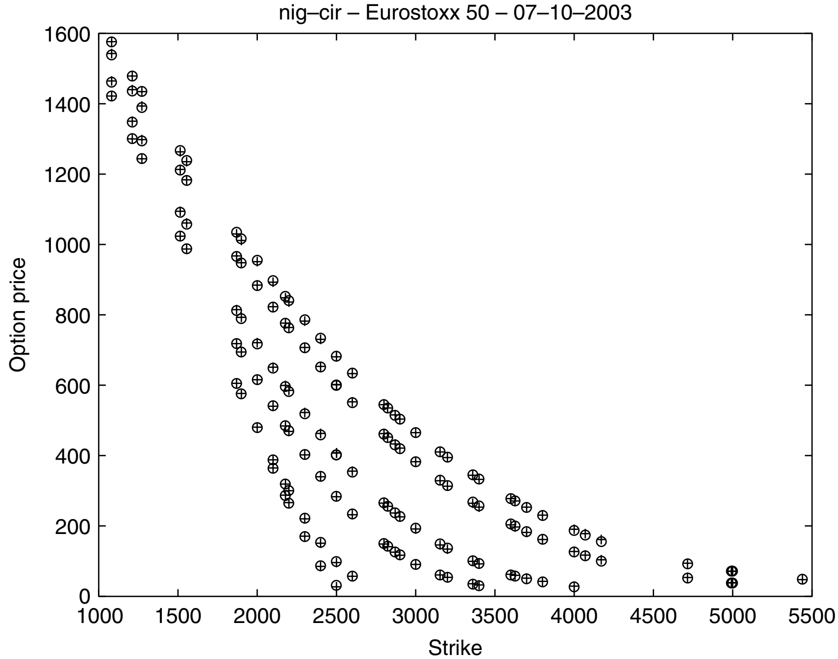


Figure 1: Calibration of NIG−CIR model

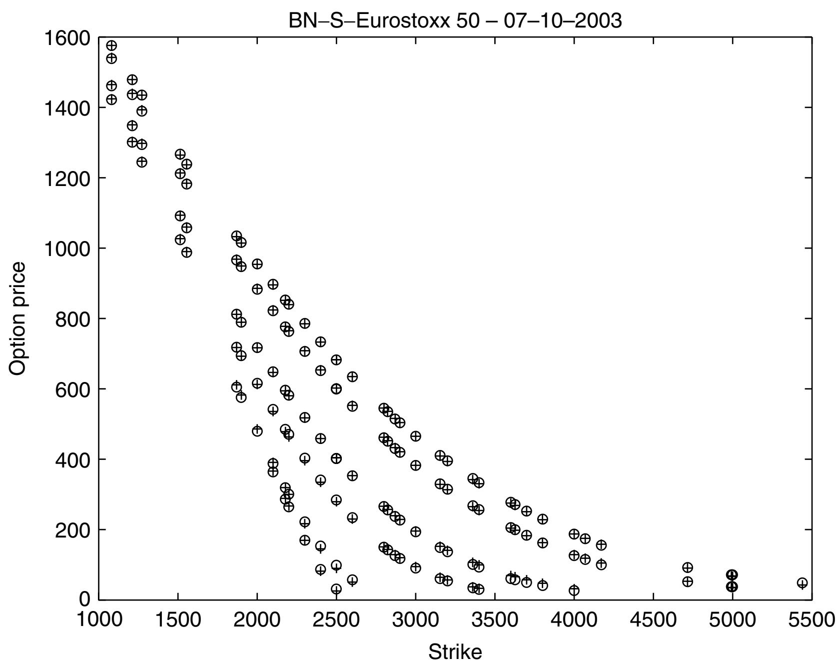


Figure 2: Calibration of Barndorff-Nielsen−Shephard model

TABLE 1: RISK-NEUTRAL PARAMETERS

HEST
$\sigma_0^2 = 0.0654, \kappa = 0.6067, \eta = 0.0707, \theta = 0.2928, \rho = -0.7571$

HESJ
$\sigma_0^2 = 0.0576, \kappa = 0.4963, \eta = 0.0650, \theta = 0.2286, \rho = -0.9900, \mu_j = 0.1791,$
$\sigma_j = 0.1346, \lambda = 0.1382$

BN−S
$\rho = -4.6750, \lambda = 0.5474, b = 18.6075, a = 0.6069, \sigma_0^2 = 0.0433$

VG−CIR
$C = 18.0968, G = 20.0276, M = 26.3971, \kappa = 1.2145, \eta = 0.5501,$
$\lambda = 1.7913, y_0 = 1$

VG−OUΓ
$C = 6.1610, G = 9.6443, M = 16.0260, \lambda = 1.6790, a = 0.3484,$
$b = 0.7664, y_0 = 1$

NIG−CIR
$\alpha = 16.1975, \beta = -3.1804, \delta = 1.0867, \kappa = 1.2101, \eta = 0.5507,$
$\lambda = 1.7864, y_0 = 1$

NIG−OUΓ
$\alpha = 8.8914, \beta = -3.1634, \delta = 0.6728, \lambda = 1.7478, a = 0.3442,$
$b = 0.7628, y_0 = 1$

TABLE 2: GLOBAL FIT ERROR MEASURES

Model:	rmse	ape	aae	arpe
HEST	3.0281	0.0048	2.4264	0.0174
HESJ	2.8101	0.0045	2.2469	0.0126
BN−S	3.5156	0.0056	2.8194	0.0221
VG−CIR	2.3823	0.0038	1.9337	0.0106
VG−OUΓ	3.4351	0.0056	2.8238	0.0190
NIG−CIR	2.3485	0.0038	1.9194	0.0099
NIG−OUΓ	3.2737	0.0054	2.7385	0.0175

4 Simulation

In the current section we describe in some detail how the particular processes presented in section 2 can be implemented in practice in a Monte Carlo simulation pricing framework. For this we first discuss the numerical implementation of the four building block processes which drive them. This will be followed by an explanation of how one assembles a time-changed Lévy process.

4.1 NIG Lévy process

To simulate a NIG process, we first describe how to simulate NIG(α, β, δ) random numbers. NIG random numbers can be obtained by mixing Inverse Gaussian (IG) random numbers and standard Normal numbers in the following manner. An IG(a, b) random variable X has a characteristic function given by:

$$E[\exp(\mathrm{i}uX)] = \exp(-a\sqrt{-2u\mathrm{i} + b^2} - b)$$

First simulate IG($1, \delta\sqrt{\alpha^2 - \beta^2}$) random numbers i_k, for example using the Inverse Gaussian generator of Michael, Schucany and Haas (Devroye 1986). Then sample a sequence of standard Normal random variables u_k. NIG random numbers n_k are then obtained via:

$$n_k = \delta^2 \beta i_k + \delta\sqrt{i_k} u_k.$$

Finally the sample paths of an NIG(α, β, δ) process $X = \{X_t, t \geq 0\}$ in the time points $t_n = n\Delta t$, $n = 0, 1, 2, \ldots$ can be generated by using the independent NIG($\alpha, \beta, \delta\Delta t$) random numbers n_k as follows:

$$X_0 = 0, \qquad X_{t_k} = X_{t_{k-1}} + n_k, \quad k \geq 1.$$

4.2 VG Lévy process

Since a VG process can be viewed as the difference of two independent Gamma processes, the simulation of a VG process becomes straightforward. A Gamma process with parameters $a, b > 0$ is a Lévy process with Gamma(a, b) distributed increments, i.e. following a Gamma distribution with mean a/b and variance a/b^2. A VG process $X^{(VG)} = \{X_t^{(VG)}, t \geq 0\}$ with parameters $C, G, M > 0$ can be decomposed as $X_t^{(VG)} = G_t^{(1)} - G_t^{(2)}$, where $G^{(1)} = \{G_t^{(1)}, t \geq 0\}$ is a Gamma process with parameters $a = C$ and $b = M$ and $G^{(2)} = \{G_t^{(2)}, t \geq 0\}$ is a Gamma process with parameters $a = C$ and $b = G$. The generation of Gamma numbers is quite standard. Possible generators are Johnk's gamma generator and Berman's gamma generator (Devroye 1986).

4.3 CIR stochastic clock

The simulation of a CIR process $y = \{y_t, t \geq 0\}$ is straightforward. Basically, we discretize the SDE:

$$\mathrm{d}y_t = \kappa(\eta - y_t)\mathrm{d}t + \lambda y_t^{1/2}\mathrm{d}W_t, \quad y_0 \geq 0,$$

where W_t is a standard Brownian motion. Using a first-order accurate explicit differencing scheme in time the sample path of the CIR process $y = \{y_t, t \geq 0\}$ in the time points $t = n\Delta t$, $n = 0, 1, 2, \ldots$, is then given by:

$$y_{t_n} = y_{t_{n-1}} + \kappa(\eta - y_{t_{n-1}})\Delta t + \lambda y_{t_{n-1}}^{1/2}\sqrt{\Delta t}\; v_n,$$

where $\{v_n, n = 1, 2, \ldots\}$ is a series of independent standard Normally distributed random numbers. For other more involved simulation schemes, like the Milstein scheme, resulting in a higher-order discretization in time, we refer to Jäckel (2002).

4.4 Gamma–OU stochastic clock

Recall that for the particular choice of a OU–Gamma process the subordinator $z = \{z_t, t \geq 0\}$ in equation (3) is given by the compound Poisson process (4).

To simulate a Gamma(a, b)–OU process $y = \{y_t, t \geq 0\}$ in the time points $t_n = n\Delta t$, $n = 0, 1, 2, \ldots$, we first simulate in the same time points a Poisson process $N = \{N_t, t \geq 0\}$ with intensity parameter $a\lambda$. Then (with the convention that an empty sum equals zero)

$$y_{t_n} = (1 - \lambda\Delta t) y_{t_{n-1}} + \sum_{k=N_{t_{n-1}}+1}^{N_{t_n}} x_k \exp(-\lambda\Delta t \tilde{u}_k),$$

where $\tilde{u}_k$ are a series of independent uniformly distributed random numbers and x_k can be obtained from your preferred uniform random number generator via $x_k = -\log(u_k)/b$.

4.5 Path generation for time-changed Lévy process

The explanation of the building block processes above allows us next to assemble all the parts of the time-changed Lévy process simulation puzzle. For this one can proceed through the following five steps (Schoutens 2003):

(i) simulate the rate of time change process $y = \{y_t, 0 \leq t \leq T\}$;
(ii) calculate from (i) the time change $Y = \{Y_t = \int_0^t y_s \mathrm{d}s, 0 \leq t \leq T\}$;
(iii) simulate the Lévy process $X = \{X_t, 0 \leq t \leq Y_T\}$;
(iv) calculate the time-changed Lévy process X_{Y_t}, for $0 \leq t \leq T$;
(v) calculate the stock price process using (6). The mean correcting factor is calculated as:

$$\frac{\exp((r-q)t)}{E[\exp(X_{Y_t})|y_0]} = \frac{\exp((r-q)t)}{\varphi(-\mathrm{i}\psi_X(-\mathrm{i}); t, 1)}.$$

5 Pricing of exotic options

As evidenced by the quality of the calibration on a set of European call options in section 3, we can hardly discriminate between the different processes on the basis of their smile–conform pricing characteristics. We therefore put the models further to the test by applying them to a range of more exotic options. These range from digital barriers, one-touch barrier options, lookback options and finally cliquet options with local as well as global parameters. These first generation exotics with path-dependent payoffs were selected since they shed more light on the dynamics of the stock processes. At the same time, the pricings of the cliquet options are highly sensitive to the forward smile characteristics induced by the models.

5.1 Exotic options

Let us consider contracts of duration T, and denote the maximum and minimum process, resp., of a process $Y = \{Y_t, 0 \leq t \leq T\}$ as

$$M_t^Y = \sup\{Y_u; 0 \leq u \leq t\} \text{ and } m_t^Y = \inf\{Y_u; 0 \leq u \leq t\}, \quad 0 \leq t \leq T.$$

5.1.1 Digital barriers We first consider digital barrier options. These options remain worthless unless the stock price hits some predefined barrier level $H > S_0$, in which case they pay at expiry a fixed amount D, normalized to 1 in the current settings. Using risk-neutral valuation, assuming no dividends and a constant interest rate r, the time $t = 0$ price is therefore given by:

$$\text{digital} = e^{-rT} E_Q[1(M_T^S \geq H)],$$

where the expectation is taken under the risk-neutral measure Q.

Observe that with the current definition of digital barriers their pricing reflects exactly the chance of hitting the barrier prior to expiry. The behaviour of the stock after the barrier has been hit does not influence the result, in contrast with the classic barrier options defined below.

5.1.2 One-touch barrier options For one-touch barrier call options, we focus on the following four types:

- The down-and-out barrier call is worthless unless its minimum remains above some 'low barrier' H, in which case it retains the structure of a European call with strike K. Its initial price is given by:

$$DOB = e^{-rT} E_Q[(S_T - K)^+ 1(m_T^S > H)]$$

- The down-and-in barrier is a normal European call with strike K, if its minimum went below some 'low barrier' H. If this barrier was never reached during the lifetime of the option, the option remains worthless. Its initial price is given by:

$$DIB = e^{-rT} E_Q[(S_T - K)^+ 1(m_T^S \leq H)]$$

- The up-and-in barrier is worthless unless its maximum crossed some 'high barrier' H, in which case it obtains the structure of a European call with strike K. Its price is given by:

$$UIB = e^{-rT} E_Q[(S_T - K)^+ 1(M_T^S \geq H)]$$

- The up-and-out barrier is worthless unless its maximum remains below some 'high barrier' H, in which case it retains the structure of a European call with strike K. Its price is given by:

$$UOB = e^{-rT} E_Q[(S_T - K)^+ 1(M_T^S < H)]$$

5.1.3 Lookback options The payoff of a lookback call option corresponds to the difference between the stock price level at expiry S_T and the lowest level it has reached during its lifetime. The time $t = 0$ price of a lookback call option is therefore given by:

$$LC = e^{-rT} E_Q[S_T - m_T^S].$$

Clearly, of the three path-dependent options introduced so far, the lookback option depends the most on the precise path dynamics.

5.1.4 Cliquet options Finally we also test the proposed models on the pricing of cliquet options. These still are very popular options in the equity derivatives world that allow the investor to participate (partially) in the performance of an underlying over a series of consecutive time periods $[t_i, t_{i+1}]$ by 'clicking in' the sum of these local performances. The local performances are measured relative to the stock level S_{t_i} attained at the start of each new subperiod, and each of the local performances is floored and/or capped to establish whatever desirable mix of positive and/or negative payoff combination. Generally on the final sum an additional global floor (cap) is applied to guarantee a minimum (maximum) overall payoff. This can all be summarized through the following payoff formula:

$$\min\left(cap_{glob}, \max\left(floor_{glob}, \sum_{i=1}^{N} \min\left(cap_{loc}, \max\left(floor_{loc}, \frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}}\right)\right)\right)\right)$$

Observe that the local floor and cap parameters effectively border the relevant 'local' price ranges by centering them around the future, and therefore unknown, spot levels S_{t_i}. The pricing will therefore depend in a non-trivial subtle manner on the forward volatility smile dynamics of the respective models, further complicated by the global parameters of the contract. For an in-depth account of the related volatility issues we refer to the contribution of Wilmott (2003) in one of the previous issues.

5.2 Exotic option prices

We price all exotic options through Monte Carlo simulation. We consistently average over 1 000 000 simulated paths. All options have a lifetime of three years. In order to check the accuracy of our simulation algorithm we simulated option prices for all European calls available in the calibration set. All algorithms gave a very satisfactory result, with pricing differences with respect to their analytic calibration values less than 0.5%.

An important issue for the path-dependent lookback, barrier and digital barrier options above is the frequency at which the stock price is observed for purposes of determining whether the barrier or its minimum level has been reached. In the numerical calculations below, we have assumed a discrete number of observations, namely at the close of each trading day. Moreover, we have assumed that a year consists of 250 trading days.

In Figure 3 we present simulation results with models for the digital barrier call option as a function of the barrier level (ranging from $1.05S_0$ to $1.5S_0$). As mentioned before, aside from the discounting factor e^{-rT}, the premiums can be interpreted as the chance of hitting the barrier during the option lifetime. In Figures 4–6, we show prices for all one-touch barrier options (as a percentage of the spot). The strike K was always taken equal to the spot S_0. For reference we summarize in Table 5 all option prices for the above discussed exotics. One can check that the barrier results agree well with the identity DIB + DOB = vanilla call = UIB + UOB, suggesting that the simulation results are well converged. Lookback prices are presented in Table 3.

Consistently over all figures the Heston prices suggest that this model (for the current calibration) results in path dynamics that are more *volatile*, breaching more frequently the imposed barriers. The results for the Lévy models with stochastic time change seem to move in pairs, with the choice of stochastic clock dominating over the details of the Lévy model upon which the stochastic time change is applied. The first couple, VG–Γ and NIG–Γ show very similar results, overall showing the least *volatile* path dynamics, whereas the VG–CIR and NIG–CIR prices consistently fall midway in the pack. Finally the OU–Γ results without stochastic clock typically fall between the Heston and the VG–CIR and NIG–CIR prices.

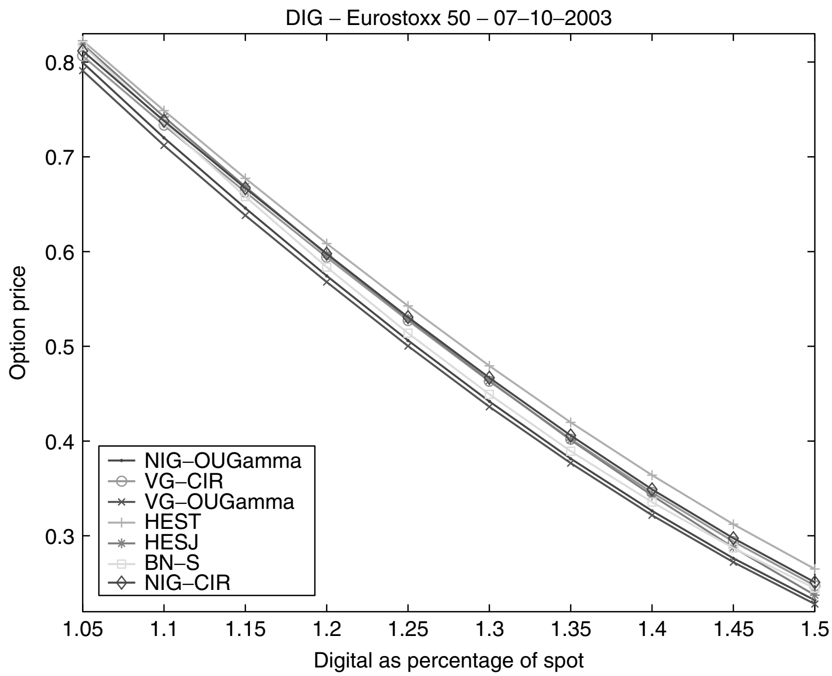


Figure 3: Digital barrier prices

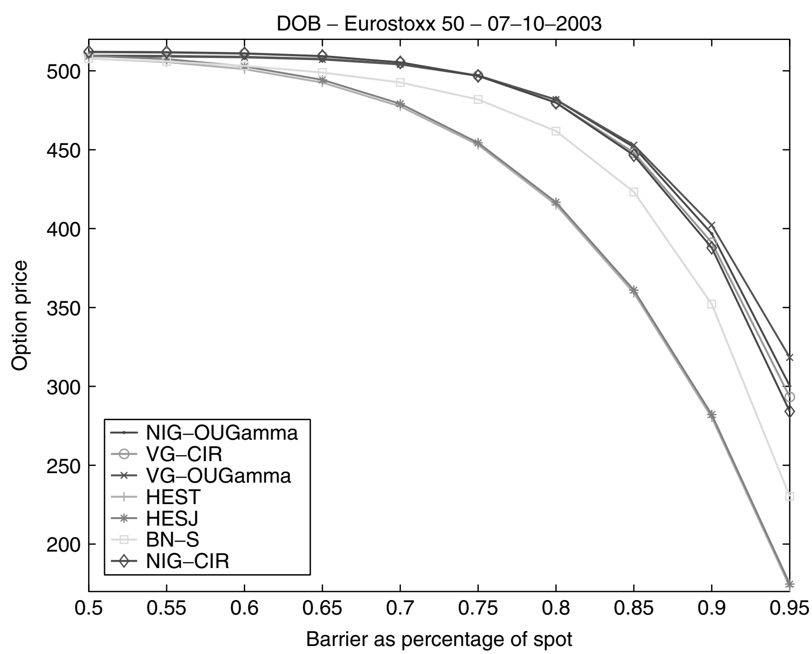


Figure 4: DOB prices

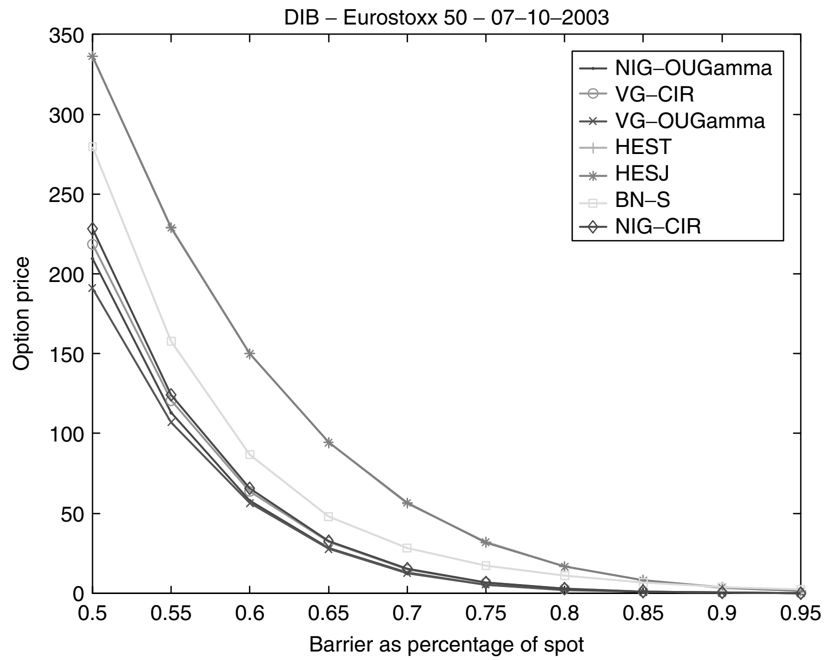


Figure 5: DIB prices

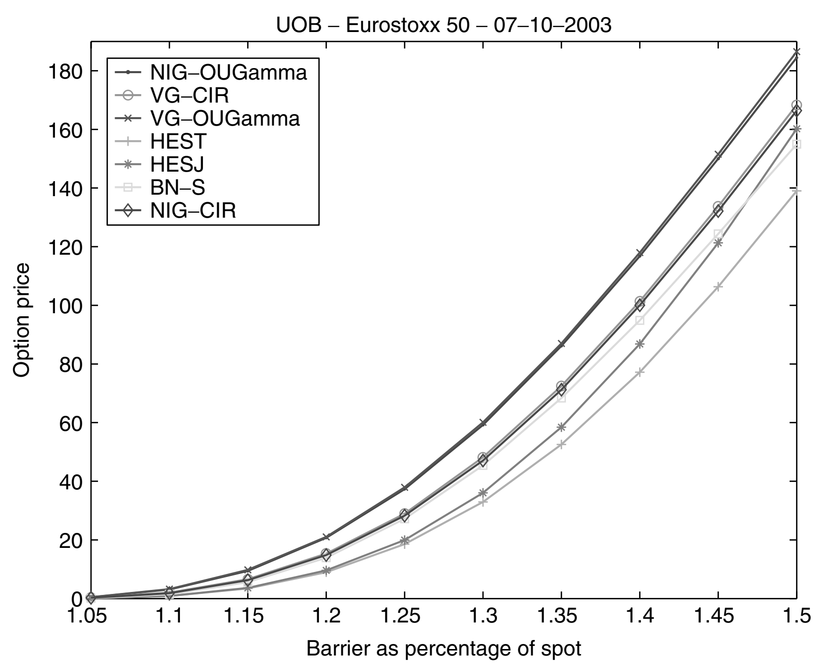


Figure 6: UOB prices

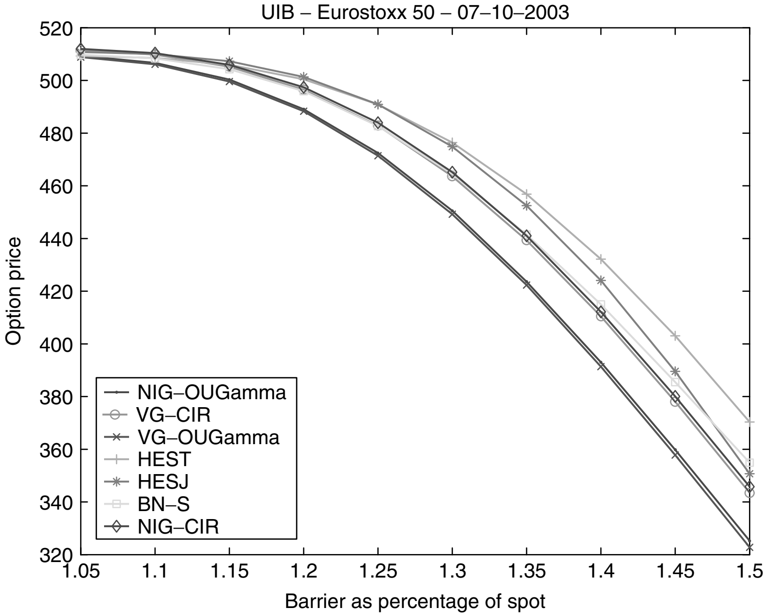


Figure 7: UIB prices

TABLE 3: LOOKBACK OPTION PRICES

HEST	HESJ	BN−S	VG−CIR	VG−OUΓ	NIG−CIR	NIG−OUΓ
844.51	845.19	771.28	724.80	713.49	730.84	722.34

Besides these qualitative observations it is important to note the magnitude of the observed differences. Lookback prices vary over about 15%, the one-touch barriers over 200%, whereas for the digital barriers we found price differences of over 10%. Finally for the cliquet premiums a variation of over 40% was noted.

For the Cliquet options, the prices are shown in Figures 8–9 for two different combinations. The numerical values can be found in Tables 6 and 7. These results are in line with the previous observations.

6 Conclusion

We have looked at different models, all reflecting non-normal returns and stochastic volatility. Empirical work has generally supported the need for both ingredients.

We have demonstrated the clear ability of all proposed processes to produce a very convincing fit to a market-conform volatility surface. At the same time we have shown that this calibration could be achieved in a timely manner using a very fast computational procedure based on FFT.

TABLE 4: IMPLIED VOLATILITY SURFACE EUROSTOXX 50, OCTOBER 7TH, 2003

Maturity (year fraction)	0.0361	0.2000	1.1944	2.1916	4.2056	5.1639
Strike						
1081.82			0.3804	0.3451	0.3150	0.3137
1212.12			0.3667	0.3350	0.3082	0.3073
1272.73			0.3603	0.3303	0.3050	0.3043
1514.24			0.3348	0.3116	0.2920	0.2921
1555.15			0.3305	0.3084	0.2899	0.2901
1870.30		0.3105	0.2973	0.2840	0.2730	0.2742
1900.00		0.3076	0.2946	0.2817	0.2714	0.2727
2000.00		0.2976	0.2858	0.2739	0.2660	0.2676
2100.00	0.3175	0.2877	0.2775	0.2672	0.2615	0.2634
2178.18	0.3030	0.2800	0.2709	0.2619	0.2580	0.2600
2200.00	0.2990	0.2778	0.2691	0.2604	0.2570	0.2591
2300.00	0.2800	0.2678	0.2608	0.2536	0.2525	0.2548
2400.00	0.2650	0.2580	0.2524	0.2468	0.2480	0.2505
2499.76	0.2472	0.2493	0.2446	0.2400	0.2435	0.2463
2500.00	0.2471	0.2493	0.2446	0.2400	0.2435	0.2463
2600.00		0.2405	0.2381	0.2358	0.2397	0.2426
2800.00			0.2251	0.2273	0.2322	0.2354
2822.73			0.2240	0.2263	0.2313	0.2346
2870.83			0.2213	0.2242	0.2295	0.2328
2900.00			0.2198	0.2230	0.2288	0.2321
3000.00			0.2148	0.2195	0.2263	0.2296
3153.64			0.2113	0.2141	0.2224	0.2258
3200.00			0.2103	0.2125	0.2212	0.2246
3360.00			0.2069	0.2065	0.2172	0.2206
3400.00			0.2060	0.2050	0.2162	0.2196
3600.00				0.1975	0.2112	0.2148
3626.79				0.1972	0.2105	0.2142
3700.00				0.1964	0.2086	0.2124
3800.00				0.1953	0.2059	0.2099
4000.00				0.1931	0.2006	0.2050
4070.00					0.1988	0.2032
4170.81					0.1961	0.2008
4714.83					0.1910	0.1957
4990.91					0.1904	0.1949
5000.00					0.1903	0.1949
5440.18						0.1938

Note that an almost identical calibration means that at the time-points of the maturities of the calibration data set the marginal distribution is fitted accurately to the risk-neutral distribution implied by the market. If we have different models leading all to such almost perfect calibrations, all models have almost the same marginal distributions. It should, however, be clear that even if

TABLE 5: EXOTIC OPTION PRICES

	H/S_0	NIG–OUΓ	VG–CIR	VG–OUΓ	HEST	HESJ	BN–S	NIG–CIR
Call		509.76	511.80	509.33	509.39	510.89	509.89	512.21
DOB	0.95	300.25	293.28	318.35	173.03	173.85	230.25	284.10
DOB	0.9	396.80	391.17	402.24	280.30	280.79	352.14	387.83
DOB	0.85	451.61	448.10	452.97	359.27	359.05	423.21	446.52
DOB	0.8	481.65	479.83	481.74	415.06	414.65	461.82	479.77
DOB	0.75	497.00	496.95	496.80	453.13	452.76	481.85	496.78
DOB	0.7	504.31	505.24	504.05	477.47	477.37	492.62	505.38
DOB	0.65	507.53	509.10	507.21	492.52	492.76	498.93	509.34
DOB	0.6	508.88	510.75	508.53	501.09	501.74	503.17	511.09
DOB	0.55	509.43	511.40	509.06	505.55	506.46	505.93	511.80
DOB	0.5	509.64	511.67	509.24	507.78	508.91	507.68	512.08
DIB	0.95	209.51	218.51	190.98	336.35	337.04	279.61	228.10
DIB	0.9	112.95	120.62	107.08	229.08	230.09	157.72	124.37
DIB	0.85	58.14	63.69	56.35	150.11	151.83	86.65	65.68
DIB	0.8	28.11	31.96	27.59	94.32	96.24	48.04	32.43
DIB	0.75	12.76	14.84	12.53	56.26	58.13	28.01	15.42
DIB	0.7	5.45	6.55	5.28	31.91	33.51	17.24	6.83
DIB	0.65	2.23	2.70	2.11	16.86	18.12	10.94	2.87
DIB	0.6	0.88	1.04	0.79	8.29	9.14	6.69	1.11
DIB	0.55	0.33	0.39	0.26	3.83	4.42	3.94	0.40
DIB	0.5	0.12	0.13	0.09	1.60	1.98	2.19	0.13
UIB	1.05	509.32	511.52	508.84	509.30	510.78	509.73	511.98
UIB	1.1	506.68	509.80	506.11	508.52	509.90	508.38	510.37
UIB	1.15	500.33	505.21	499.56	505.96	507.08	504.28	505.93
UIB	1.2	489.05	496.50	488.30	500.42	501.04	495.95	497.41
UIB	1.25	472.47	482.84	471.39	490.85	490.73	482.66	483.94
UIB	1.3	450.54	463.62	449.23	476.43	475.30	464.48	465.16
UIB	1.35	423.62	439.32	422.32	456.83	454.79	441.48	441.00
UIB	1.4	393.01	410.46	391.36	432.17	428.96	414.98	412.16
UIB	1.45	359.77	378.05	357.80	403.03	399.24	385.50	380.04
UIB	1.5	325.25	343.46	322.79	370.33	365.57	354.90	345.79
UOB	1.05	0.44	0.27	0.49	0.09	0.10	0.13	0.23
UOB	1.1	3.08	2.00	3.22	0.87	0.98	1.48	1.84
UOB	1.15	9.43	6.59	9.77	3.42	3.80	5.58	6.27
UOB	1.2	20.71	15.29	21.03	8.96	9.85	9.85	14.80
UOB	1.25	37.29	28.95	37.94	18.53	20.15	27.20	28.26
UOB	1.3	59.22	48.17	60.10	32.95	35.58	45.38	47.04
UOB	1.35	86.14	72.47	87.00	52.55	56.10	68.39	71.21
UOB	1.4	116.75	101.33	117.96	77.20	81.93	94.88	100.04
UOB	1.45	149.98	133.74	151.52	106.35	111.65	124.36	132.16
UOB	1.5	184.50	168.33	186.53	139.04	145.31	154.96	166.41

TABLE 5: (*continued*)

	H/S_0	NIG–OUΓ	VG–CIR	VG–OUΓ	HEST	HESJ	BN–S	NIG–CIR
DIG	1.05	0.7995	0.8064	0.7909	0.8226	0.8218	0.8173	0.8118
DIG	1.1	0.7201	0.7334	0.7120	0.7487	0.7478	0.7360	0.7380
DIG	1.15	0.6458	0.6628	0.6382	0.6774	0.6762	0.6580	0.6670
DIG	1.2	0.5744	0.5940	0.5678	0.6084	0.6069	0.5836	0.5977
DIG	1.25	0.5062	0.5273	0.5003	0.5427	0.5408	0.5138	0.5308
DIG	1.3	0.4418	0.4630	0.4363	0.4794	0.4770	0.4493	0.4668
DIG	1.35	0.3816	0.4021	0.3767	0.4198	0.4169	0.3893	0.4059
DIG	1.4	0.3264	0.3456	0.3217	0.3640	0.3603	0.3355	0.3490
DIG	1.45	0.2763	0.2940	0.2722	0.3122	0.3087	0.2870	0.2975
DIG	1.5	0.2321	0.2474	0.2280	0.2649	0.2610	0.2446	0.2510

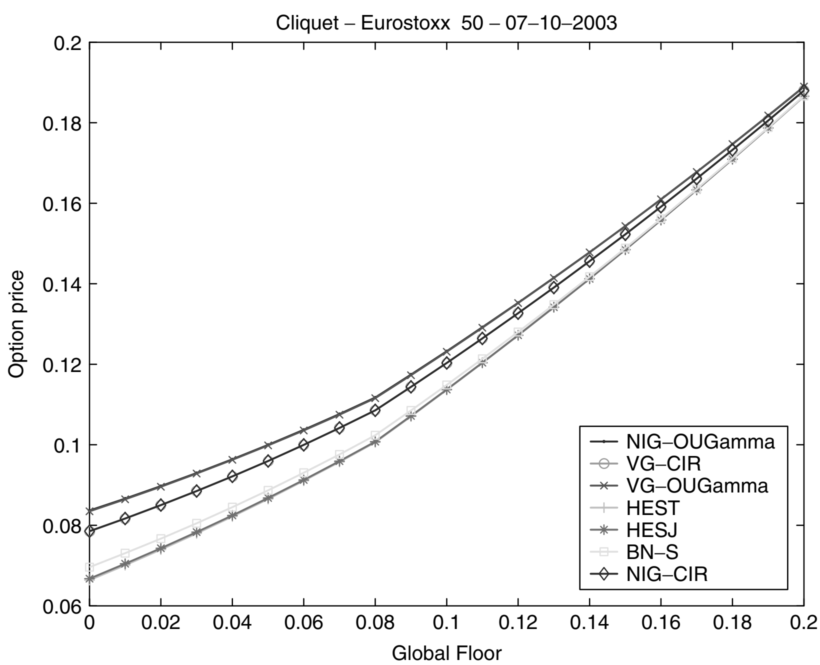


Figure 8: Cliquet prices: $cap_{loc} = 0.08$, $flo_{loc} = -0.08$, $cap_{glo} = +\infty$, $N = 3$, $t_1 = 1$, $t_2 = 2$, $t_3 = 3$

at all time-points $0 \leq t \leq T$ marginal distributions among different models coincide, this does not imply that exotic prices should also be the same. This can be seen from the following discrete-time example. Let $n \geq 2$ and $X = \{X_i, i = 1, \ldots, n\}$ be an iid sequence and let $\{u_i, i = 1, \ldots, n\}$ be a independent sequence which randomly varies between $u_i = 0$ and 1. We propose two discrete

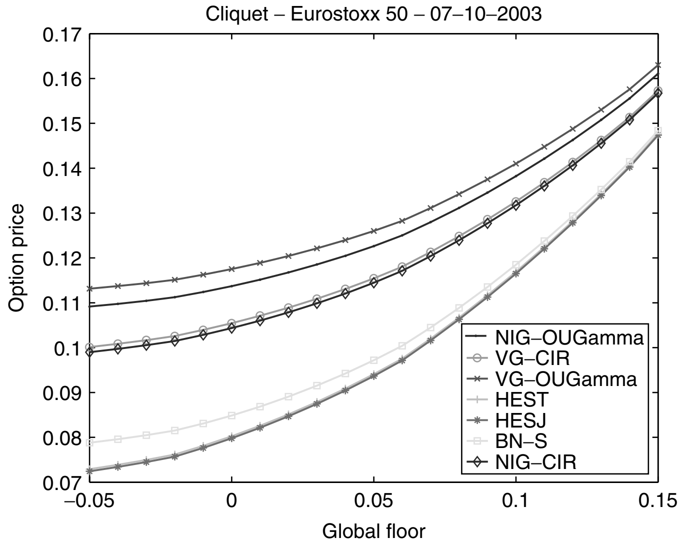


Figure 9: Cliquet prices: $flo_{loc} = -0.03$, $cap_{loc} = 0.05$, $cap_{glo} = +\infty$, $T = 3$, $N = 6$, $t_i = i/2$

TABLE 6: CLIQUET PRICES: $CAP_{LOC} = 0.08$, $FLO_{-LOC} = 0.08$, $CAP_{GLO} = +\infty$, $FLO_{GLO} \in [0, 0.20]$, $N = 3$, $T_1 = 1$, $T_2 = 2$, $T_3 = 3$

flo_{glo}	NIG–CIR	NIG–OUΓ	VG–OUΓ	VG–CIR	HESJ	HEST	BN–S
0.00	0.0785	0.0837	0.0835	0.0785	0.0667	0.0683	0.0696
0.01	0.0817	0.0866	0.0865	0.0817	0.0704	0.0719	0.0731
0.02	0.0850	0.0897	0.0896	0.0850	0.0743	0.0757	0.0767
0.03	0.0885	0.0930	0.0928	0.0885	0.0783	0.0796	0.0805
0.04	0.0922	0.0964	0.0963	0.0921	0.0825	0.0837	0.0845
0.05	0.0960	0.1000	0.0998	0.0960	0.0868	0.0879	0.0887
0.06	0.1000	0.1037	0.1036	0.1000	0.0913	0.0923	0.0930
0.07	0.1042	0.1076	0.1075	0.1042	0.0959	0.0969	0.0976
0.08	0.1086	0.1117	0.1116	0.1085	0.1008	0.1017	0.1024
0.09	0.1144	0.1174	0.1173	0.1144	0.1072	0.1080	0.1085
0.10	0.1203	0.1232	0.1231	0.1203	0.1137	0.1145	0.1149
0.11	0.1264	0.1292	0.1291	0.1264	0.1204	0.1211	0.1214
0.12	0.1327	0.1353	0.1352	0.1327	0.1272	0.1279	0.1280
0.13	0.1391	0.1415	0.1414	0.1391	0.1342	0.1348	0.1348
0.14	0.1456	0.1478	0.1478	0.1456	0.1412	0.1418	0.1418
0.15	0.1523	0.1543	0.1543	0.1523	0.1485	0.1490	0.1489
0.16	0.1591	0.1610	0.1610	0.1591	0.1558	0.1562	0.1561
0.17	0.1661	0.1677	0.1678	0.1661	0.1633	0.1637	0.1635
0.18	0.1732	0.1747	0.1747	0.1733	0.1709	0.1712	0.1711
0.19	0.1805	0.1817	0.1818	0.1806	0.1787	0.1789	0.1788
0.20	0.1880	0.1889	0.1890	0.1880	0.1866	0.1868	0.1867

TABLE 7: CLIQUET PRICES: $FLO_{LOC} = -0.03$, $CAP_{LOC} = 0.05$, $CAP_{GLO} = +\infty$, $T = 3$, $N = 6$, $T_I = I/2$

flo_{glo}	NIG–CIR	NIG–OUΓ	VG–OUΓ	VG–CIR	HESJ	HEST	BN–S
−0.05	0.0990	0.1092	0.1131	0.1001	0.0724	0.0762	0.0788
−0.04	0.0997	0.1098	0.1137	0.1008	0.0734	0.0771	0.0796
−0.03	0.1005	0.1104	0.1144	0.1017	0.0745	0.0781	0.0805
−0.02	0.1015	0.1112	0.1151	0.1026	0.0757	0.0792	0.0815
−0.01	0.1028	0.1124	0.1162	0.1039	0.0776	0.0811	0.0831
0.00	0.1044	0.1137	0.1175	0.1054	0.0798	0.0831	0.0849
0.01	0.1060	0.1152	0.1189	0.1071	0.0821	0.0853	0.0869
0.02	0.1079	0.1168	0.1204	0.1089	0.0847	0.0877	0.0891
0.03	0.1099	0.1185	0.1221	0.1109	0.0874	0.0904	0.0915
0.04	0.1121	0.1205	0.1240	0.1131	0.0904	0.0932	0.0942
0.05	0.1145	0.1226	0.1260	0.1154	0.0937	0.0963	0.0972
0.06	0.1171	0.1250	0.1283	0.1180	0.0971	0.0996	0.1004
0.07	0.1204	0.1280	0.1311	0.1213	0.1016	0.1039	0.1045
0.08	0.1239	0.1312	0.1342	0.1248	0.1063	0.1084	0.1088
0.09	0.1277	0.1346	0.1375	0.1286	0.1113	0.1132	0.1135
0.10	0.1317	0.1382	0.1410	0.1326	0.1165	0.1183	0.1185
0.11	0.1361	0.1421	0.1448	0.1368	0.1220	0.1237	0.1238
0.12	0.1406	0.1463	0.1488	0.1414	0.1278	0.1293	0.1294
0.13	0.1456	0.1508	0.1531	0.1462	0.1339	0.1352	0.1353
0.14	0.1508	0.1556	0.1576	0.1514	0.1403	0.1414	0.1415
0.15	0.1567	0.1611	0.1630	0.1573	0.1474	0.1484	0.1484

(be it unrealistic) stock price models, $S^{(1)}$ and $S^{(2)}$, with the same marginal distributions:

$$S_i^{(1)} = u_i X_1 + (1 - u_i) X_2 \text{ and } S_i^{(2)} = X_i.$$

The first process flips randomly between two states X_1 and X_2, both of which follow the distribution of the iid sequence, and so do all the marginals at the time-points $i = 1, \ldots, n$. The second process changes value in all time-points. The values are independent of each other and all follow again the same distribution of the iid sequence. In both cases all the marginal distributions (at every $i = 1, \ldots, n$) are the same (as the distribution underlying the sequence X). It is clear, however, that the maximum and minimum of both processes behave completely differently. For the first process, the maximal $\max_{j \le i} S_i^{(1)} = \max(X_1, X_2)$ and minimal process $\min_{j \le i} S_i^{(1)} = \min(X_1, X_2)$ for i are large enough, whereas for the second process there is much more variation possible and it clearly leads to other distributions. In summary, it should be clear that equal marginal distributions of a process do not at all imply equal marginal distributions of the associated minimal or maximal process. This explains why matching European call prices do not lead necessarily to matching exotic prices. It is the underlying fine-grain structure of the process that will have an important impact on the path-dependent option prices.

We have illustrated this by pricing exotics by Monte Carlo simulation, showing that price differences of over 200% are no exception. For lookback call options a price range of more than 15% amongst the models was observed. A similar conclusion was valid for the digital barrier premiums. Even for cliquet options, which only depend on the stock realizations over a limited amount of time-points, prices vary substantially among the models. At the same time the presented details of the Monte Carlo implementation should allow the reader to embark on his/her own pricing experiments.

The conclusion is that great care should be taken when employing attractive fancy-dancy models to price (or even more importantly, to evaluate hedge parameters for) exotics. As far as we know no detailed study about the underlying path structure of assets has been done yet. Our study motivates such a deeper study.

Acknowledgments

The first author is a Postdoctoral Fellow of the Fund for Scientific Research–Flanders (Belgium) (F.W.O.–Vlaanderen). We thank Marc Jeannin for his devoted programming work.

REFERENCES

■ Bakshi, G., Cao, C. and Chen, Z. (1997) Empirical performance of alternative option pricing models. *The Journal of Finance*, Vol. LII, No. 5, 2003–2049.

■ Barndorff-Nielsen, O. E. and Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B*, 63, 167–241.

■ Barndorff-Nielsen, O. E., Nicolata, E. and Shephard, N. (2002) Some recent developments in stochastic volatility modelling. *Quantitative Finance*, 2, 11–23.

■ Bertoin, J. (1996) Lévy processes. *Cambridge Tracts in Mathematics*, 121, Cambridge University Press, Cambridge.

■ Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–654.

■ Carr, P. and Madan, D. (1998) Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2, 61–73.

■ Carr, P., Geman, H., Madan, D. H. and Yor, M. (2001) Stochastic volatility for Lévy processes. *Prépublications du Laboratoire de Probabilités et Modèles Aléatoires*, 645, Universités de Paris 6 & Paris 7, Paris.

■ Clark, P. (1973) A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41, 135–156.

■ Cox, J., Ingersoll, J. and Ross, S. (1985) A theory of the term structure of interest rates, *Econometrica*, 53, 385–408.

■ Devroye, L. (1986) *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

■ Heston, S. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6, 327–343.

■ Jäckel, P. (2002) *Monte Carlo Methods in Finance*. John Wiley & Sons.

■ Knudsen, Th. and Nguyen-Ngoc, L. (2003) Pricing European options in a stochastic volatility-jump-diffusion model, Working paper.
■ Madan, D. B., Carr, P. and Chang, E. C. (1998) The variance gamma process and option pricing, *European Finance Review*, 2, 79–105.
■ Schoutens, W. (2003) *Lévy Processes in Finance: Pricing Financial Derivatives*. John Wiley & Sons.
■ Wilmott, P. (2003) Cliquet options and volatility models. *Wilmott Magazine*, Dec.

24 Timing the Smile

Jean-Pierre Fouque,[*] George Papanicolaou,[**] Ronnie Sircar[†] and Knut Sølna[‡]

Within the general framework of stochastic volatility, the authors propose a method, which is consistent with no-arbitrage, to price complicated path-dependent derivatives using only the information contained in the implied volatility skew. This method exploits the time scale content of volatility to bridge the gap between skews and derivatives prices. Here they present their pricing formulas in terms of Greeks free from the details of the underlying models and mathematical techniques.

1 Underlying or smile?

Our goal is to address the following fundamental question in pricing and hedging derivatives. How traded call options, quoted in terms of implied volatilities, can be used to price and hedge more complicated contracts. One can approach this difficult problem in two different ways: modeling the evolution of the underlying or modeling the evolution of the implied volatility surface. In both cases one requires that the model is free of arbitrage.

Modeling the underlying usually involves the specification of a multi-factor Markovian model under the risk-neutral pricing measure (see Duffie *et al*. 2000, for instance). The calibration to the observed implied volatilities of the parameters of that model, including the market prices of risk, is a challenging task because of the complex relation between call option prices and model parameters (through a pricing partial differential equation, for instance). A major problem with this approach is to find the 'right model' which will produce a stable parameter estimation. We like to think of this problem as the '(t, T, K)' problem: for a given present time t and a fixed maturity T, it is usually easy with low dimensional models to fit the skew with respect to strikes K. Getting a good fit of the term structure of implied volatility, that is when a range of observed maturities are taken into account, is a much harder problem which can be handled with a sufficient number of parameters and eventually including jumps in the model (see Duffie *et al*. 2000, Carr *et al*. 2000 for instance). The main problem remains: the stability with respect to t of these calibrated parameters.

Contact addresses: [*]Department of Mathematics, NC State University, Raleigh NC 27695-8205, [**]Department of Mathematics, Stanford University, Stanford CA 94305, [†]Department of Operations Research & Financial Engineering, Princeton University, E-Quad, Princeton, NJ 08544 and [‡]Department of Mathematics, University of California, Irvine CA 92697.
E-mail: fouque@math.ncsu.edu, papanico@math.stanford.edu, sircar@princeton.edu and ksolna@math.uci.edu

However, this is a highly desirable quality if one wants to use the model to compute no-arbitrage prices of more complex path-dependent derivatives, since in this case the distribution over time of the underlying is crucial.

Modeling directly the evolution of the implied volatility surface is a promising approach but involves some complicated issues. One has to make sure that the model is free of arbitrage or, in other words, that the surface is produced by some underlying under a risk-neutral measure. This is not an obvious task (see Cont and da Fonseca 2002 and references therein). The choice of a model and its calibration is also an important issue in this approach. But most importantly, in order to use this modeling to price other path-dependent contracts, one has to identify a corresponding underlying which typically does not lead to a low dimensional Markovian evolution.

Wouldn't it be nice to have a direct and simple connection between the observed implied volatilities and prices of more complex path-dependent contracts! Our objective is to provide such a bridge. This is done by using a combination of singular and regular perturbations techniques corresponding respectively to fast and slow time scales in volatility. We obtain a parametrization of the implied volatility surface in terms of Greeks, which involves four parameters at the first order of approximation. This procedure leads to parameters which are exactly those needed to price other contracts at this level of approximation. In our previous work presented in Fouque *et al*. (2000) we used only the fast volatility time scale combined with a statistical estimation of an effective constant volatility from historical data. The introduction of the slow volatility time scale enables us to capture more accurately the behavior of the term structure of implied volatility at long maturities. Moreover in the framework presented here, statistics of historical data are not needed. Thus, in summary, we directly link the implied volatilities to prices of path-dependent contracts by exploiting volatility time scales. We refer to Fouque *et al*. (2003a) for a detailed presentation of volatility time scales in the S&P500 index. The mathematical derivation of the combined regular and singular perturbations can be found in Fouque *et al*. (2004b).

2 Volatility time scales

Stochastic volatility models can be seen as continuous time versions of ARCH-type models which have been introduced by R. Engle. The importance of volatility modeling is reflected in the fact that R. Engle was awarded the 2003 Nobel Prize for Economics, shared with C. Granger whose work also deals with time scale modeling. Our modeling point of view is that volatility is driven by several stochastic factors running on different time scales. The presence of these volatility factors is well documented in the literature using underlying returns data (see for instance Alizadeh *et al*. (2002), Anderson and Bollerslev (1997), Chernov *et al*. (2003), Engle and Patton (2001), Fouque *et al*. (2003a), Hillebrand (2003), LeBaron (2001) Melino and Turnbull (1990), Muller *et al*. (1997)). In fact these factors play a central role in derivatives pricing and generate in a complex way the term structure of implied volatility. Our perturbative approach vastly simplifies this complex relation and leads to simple formulas which reflect the main features of the implied volatilities that follow from the effects of these various volatility time scales.

Before going into formulas, we describe in simple words what these time scales represent and their effects on derivatives pricing.

A stochastic volatility factor running on a **slow scale** means that it takes a long time (compared with typical maturities) for this factor to change appreciably and decorrelate. In the slow scale limit this would then become a constant volatility factor frozen at the present level. In this limit, derivatives prices would be obtained by the usual Black–Scholes pricing theory at this constant volatility level. Our regular perturbation analysis gives corrections to this limit which affect long dated options and therefore are reflected in the behavior of the skew at large maturities. Slow scales, or small perturbations, have been considered in Fournie *et al*. (1997), Lee (1999), Sircar and Papanicolaou (1999).

A stochastic volatility factor running on a **fast scale** means that it takes a short time (compared with typical maturities) for this factor to come back to its mean level and decorrelate. In the fast scale limit this would then also become a constant volatility factor at an effective level $\overline{\sigma}$ determined by the averaged square volatility

$$\overline{\sigma}^2 \approx \frac{1}{T-t}\int_t^T \sigma^2(s)ds, \tag{1}$$

the slow volatility factor being frozen, and where we assume that the fast volatility factor is mean-reverting with rapid mixing properties. Our singular perturbation analysis gives corrections to this Black–Scholes limit which affect options over various maturities and therefore are reflected in the behavior of the skew.

The formulas presented below are obtained by considering that volatility is driven by both slow and fast scale factors. Our analysis, which combines regular and singular perturbations, leads to a parametrization of the term structure of implied volatility which is valid over a wide range of maturities. In that sense, to the leading order, we solve the '(T, K) problem'. In fact it turns out that the calibration of our parameters is stable in time and therefore, to the leading order, we provide a solution to the full (t, T, K) problem, and we demonstrate that modeling volatility with at least two factors (a slow and a fast) is consistent with the behavior of derivative markets.

3 Volatility skew formulas

3.1 Vanilla prices

Our asymptotic analysis performed on European vanilla options leads to an explicit formula for the approximated price when the underlying model has a volatility driven by a slow and a fast factor. The leading order term, $P_{BS}(\sigma^\star)$, is the classical Black–Scholes price of the contract evaluated at the constant volatility $\sigma^\star$ which will be calibrated from the observed implied volatilities in Section 3.2. The correction is a combination of three terms expressed in terms of the Greeks of the Black–Scholes price at the volatility level $\sigma^\star$:

$$P \approx P_{BS}(\sigma^\star) + (T-t)\left\{v_0\mathcal{V} + v_1 S\Delta(\mathcal{V}) + v_3 S\Delta(S^2\Gamma)\right\}, \tag{2}$$

where S denotes the present value at time t of the underlying, T denotes the maturity, and the Greeks are given by

$$\mathcal{V} = \frac{\partial P_{BS}}{\partial \sigma}(\sigma^\star) \qquad \text{(Vega)}$$

$$S\Delta(\mathcal{V}) = S\frac{\partial^2 P_{BS}}{\partial S \partial \sigma}(\sigma^\star) \qquad (S\text{Delta(Vega)})$$

$$S\Delta(S^2\Gamma) = S\frac{\partial}{\partial S}\left(S^2\frac{\partial^2 P_{BS}}{\partial S^2}\right)(\sigma^\star) \qquad (S\text{Delta}(S^2\text{Gamma})).$$

An extensive discussion of the role of the Greeks can be found in Haug (2003).

The small parameters (v_0, v_1, v_3) will also be calibrated from the observed implied volatilities as we will explain in Section 3.2. The terms involving v_0 and v_1 are price corrections that come from the effect of the slow factor. The term involving v_3 is caused by the fast factor in the volatility and its leverage effect. We remark that the **effective volatility** $\sigma^\star$ includes a correction that comes from the market price of *fast* volatility risk; this volatility level correction could alternatively have been incorporated as a price correction term proportional to S^2Gamma (the apparently missing v_2 term). In that sense $\sigma^\star$ is a corrected value of the average volatility $\overline{\sigma}$ introduced in (1). The main advantage of introducing $\sigma^\star$ is that it can be estimated from the smile as explained below in Section 3.2. In contrast, $\overline{\sigma}$ can only be estimated from long records of historical returns data.

Observe that for European vanilla options we have the explicit relation:

$$\mathcal{V} = (T-t)\sigma S^2\Gamma,$$

and therefore the price approximation can be written in the form

$$P \approx P_{BS}(\sigma^\star) + (T-t)v_0\mathcal{V} + \left\{(T-t)v_1 + (v_3/\sigma^\star)\right\} S\Delta(\mathcal{V}). \tag{3}$$

It is crucial to observe that we can implement this level of price approximation knowing only the present value, S, and the four parameters $\sigma^\star$, v_0, v_1 and v_3. We next show that these parameters in fact can be estimated from the implied volatilities.

3.2 Calibrating the smile

The price approximation given above in the case with European call options leads to the following approximation of the implied volatility skew:

$$I(t, S; T, K) \approx b_0 + b_1(T-t) + \{m_0 + m_1(T-t)\}\,\mathrm{LMMR}, \tag{4}$$

where as in Fouque *et al.* (2000) the Log-Moneyness-to-Maturity Ratio is defined by

$$\mathrm{LMMR} = \frac{\log(K/S)}{T-t}.$$

In fact the coefficients m_0 and b_0 are due to the fast volatility factor while the coefficients m_1 and b_1 are due to the slow volatility factor which becomes important for large maturities.

Our method now consists of the following steps:

(I) Given a discrete set of implied volatilities $I(t, S; K_i, T_j)$, we carry out the linear least squares fits, $b + m$ LMMR, with respect to LMMR for each time to maturity $\tau_j = T_j - t$ on a given day t.This is illustrated in Figure 1 for six different maturities and for strikes not far out-of-the-money.

We will see in section 4 that higher order corrections are needed to capture the turn of the skew as illustrated in Figure 3.

On a given day the above regression gives a pair of estimates of m and b for each maturity $T-t$ that is available on that day. Next we estimate the parameters (m_0, b_0), respectively (m_1, b_1), by linear regression with respect to $(T-t)$ of m, respectively b.

In Figure 2 we show the results of these linear regressions on a given day (June 5, 2003) for the S&P500 implied volatilities.

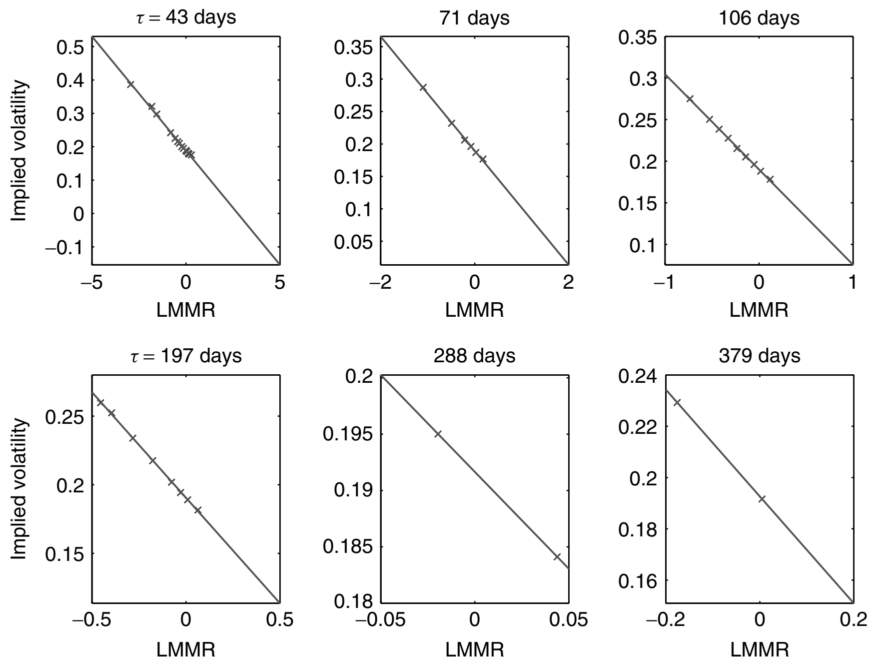


Figure 1: S&P500 implied volatility data on June 5, 2003 and fits to the affine LMMR approximation (4) for six different maturities

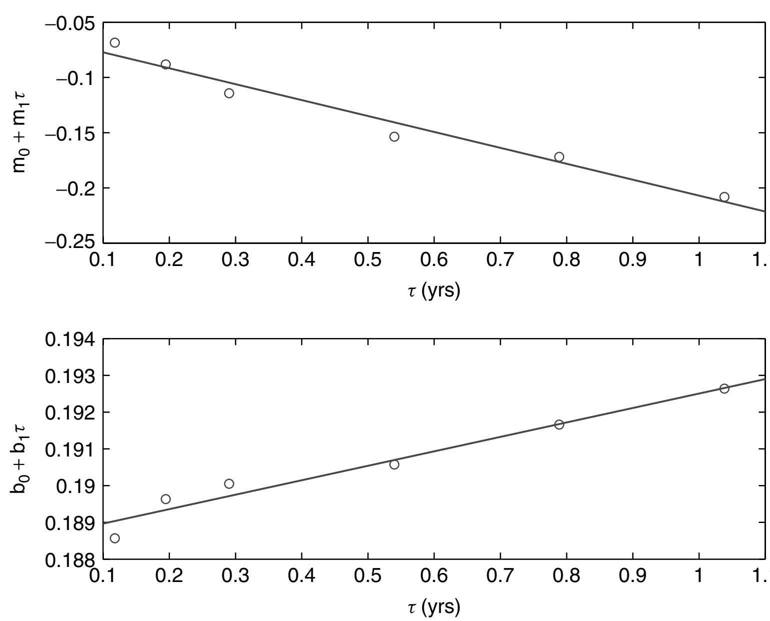


Figure 2: S&P500 implied volatility data on June 5, 2003 and fits to the two-scales asymptotic theory. The bottom (resp. top) figure shows the linear regression of b (resp. a) with respect to time to maturity $\tau = T - t$

(II) The parameters $\sigma^\star, v_0, v_1$ and v_3 that are needed for pricing are given *explicitly* by the following formulas:

$$\begin{aligned}
\sigma^\star &= b_0 + m_0\left(r - \frac{b_0^2}{2}\right)\\
v_0 &= b_1 + m_1\left(r - \frac{b_0^2}{2}\right)\\
v_1 &= m_1 b_0^2\\
v_3 &= m_0 b_0^3
\end{aligned} \tag{5}$$

Observe that in the regime that our approximation is valid the parameters v_0, v_1 and v_3 are expected to be small, while $\sigma^\star$ is the leading order magnitude of volatility. This is what we see on Figure 2 for S&P500 on June 5, 2003. Here, r is the short rate which we assume to be known and constant.

3.3 Pricing equations

We explain some of the background for the above results and relate this to deriving pricing equations for rather general contracts. The price approximation given by the right-hand side of (3) can be written

$$P_{BS}(\sigma^\star) + P_1(\sigma^\star)$$

where the correction $P_1(\sigma^\star)$ is given by:

$$P_1(\sigma^\star) = (T-t)v_0\mathcal{V} + \left\{(T-t)v_1 + (v_3/\sigma^\star t)\right\} S\Delta(\mathcal{V}).$$

The leading order term $P_{BS}(\sigma^\star)$ is the classical Black–Scholes price at the constant volatility level $\sigma^\star$. It is the solution of the PDE problem

$$\mathcal{L}_{BS}(\sigma^\star)P_{BS} = 0$$

with the terminal condition $P_{BS}(T,S) = h(S)$ where h is the payoff function for the European vanilla option that we consider. Recall that the Black–Scholes operator is given by

$$\mathcal{L}_{BS}(\sigma^\star) = \frac{\partial}{\partial t} + \frac{1}{2}(\sigma^\star)^2 S^2 \frac{\partial^2}{\partial S^2} + r\left(S\frac{\partial}{\partial S} - \cdot\right).$$

The price correction $P_1(\sigma^\star)$ solves the following partial differential equation

$$\mathcal{L}_{BS}(\sigma^\star)P_1(\sigma^\star) = -\left(2v_0\frac{\partial P_{BS}}{\partial\sigma} + 2v_1 S\frac{\partial^2 P_{BS}}{\partial S\partial\sigma} + v_3 S\frac{\partial}{\partial S}\left(S^2\frac{\partial^2 P_{BS}}{\partial S^2}\right)\right)(\sigma^\star), \tag{6}$$

with a zero terminal condition $P_1(\sigma^\star)(T, S) = 0$. In terms of the Greeks introduced in (2) this equation reads

$$\mathcal{L}_{BS}(\sigma^\star) P_1(\sigma^\star) = -\left(2v_0\mathcal{V} + 2v_1 S\Delta(\mathcal{V}) + v_3 S\Delta(S^2\Gamma)\right) \tag{7}$$

where again the Greeks are evaluated at the effective volatility $\sigma^\star$.

3.4 Pricing exotic contracts

We are now in a position to carry out our main task, that is, with the parameters calibrated from the smile we will price more general contracts than just the vanilla cases considered above. The pricing procedure is simply:

1. Compute the **leading order** (Black–Scholes) price $P_0(\sigma^\star)$ which is the price of the contract at the **constant** volatility level $\sigma^\star$ defined in (5). This involves solving partial differential equations with appropriate boundary and terminal conditions.
2. Compute the **Greeks** $\mathcal{V}$, $S\Delta(\mathcal{V})$, $S\Delta(S^2\Gamma)$ of the price $P_0(\sigma^\star)$ of the exotic contract.
3. Compute the price correction $P_1(\sigma^\star)$ by solving the same pricing problem as in Step 1 for $P_0(\sigma^\star)$ with the constant volatility $\sigma^\star$, but with a **zero payoff** and with a **source**, as in (7), defined in terms of the computed Greeks and the three parameters v_0, v_1 and v_3 that are calibrated from the skew as explained in Section 3.2.
4. The price is now given by correcting the leading order price:

   $$P \approx P_0(\sigma^\star) + P_1(\sigma^\star).$$

We present next some remarks regarding the above procedure.

- For complicated contracts, computing the price $P_0(\sigma^\star)$ along with the Greeks usually requires numerical methods (finite differences, Monte Carlo, etc.) depending on the nature of the contract. We do not comment on the details of these numerical methods. These methods are well documented elsewhere (see for instance Wilmott 2000), what is important to note is that in this framework they only need to be applied in a setting with a **constant** volatility.
- Solving the problem for the correction P_1 requires generalizations of these methods to the case with a source term. The authors have explicitly considered some of these problems (Asian, Barriers, American, etc.) in Fouque *et al*. (2000) with only the fast scale, in Fouque *et al*. (2004b) and Fouque *et al*. (2004c) with both fast and slow scales. Note that for American options the free boundary is determined by solving the problem for $P_0(\sigma^\star)$ and it is then used as a fixed boundary in the problem with a source that determines $P_1(\sigma^\star)$.

4 Further corrections

Observe that above we used a leading order expansion of the price in the context of a multi-factor stochastic volatility to obtain a connection between the implied volatility skew and pricing formulas. The mathematical tools underlying the approximation (6) consists of writing first a class of stochastic volatility models containing fast and slow volatility factors. We then expand the corresponding pricing equations with respect to the small parameters defining these two time scales: one parameter being the time scale of the fast factor and the other being the reciprocal of the time scale of the slow factor. The formulas above constitute the first-order approximation with respect to these parameters.

A natural extension of this approach is to include further terms of the asymptotic expansion. In particular, as the first-order terms describe affine skews (as a function of log-moneyness), but often we observe slight turns (or wings) at extreme strikes, we consider the next set of terms, which turn out to allow for skews that are quartic polynomials in log-moneyness. By including these terms we improve the quality of the fit to the skew and the accuracy of the pricing formulas. Indeed the number of parameters increases (from four to eleven), higher order Greeks are involved (up to sixth-order derivatives) and consequently the computational cost also increases.

The upshot of a long calculation that includes the next three (second-order) terms in the combined fast and slow scales expansion, is that, outside of a small terminal layer (very close to expiration), implied volatilities are approximated by

$$I \approx \sum_{j=0}^{4} a_j(\tau)\,(LM)^j + \frac{1}{\tau}\Phi_t, \tag{8}$$

where τ denotes the time-to-maturity $T - t$, LM denotes the moneyness $\log(K/S)$, and Φ_t is a rapidly changing component that varies with the fast volatility factor. In (8) we choose to separate the log-moneyness and the maturity dependence. Alternatively we could have written the implied volatility as a polynomial in LMMR as we did in (4) for the first-order approximation.

Again, this calibration formula is employed in a two-stage fitting procedure that recognizes the thinness of data in the maturity dimension, relative to the many available strikes. On each day, the skew for each available maturity is fit to a quartic polynomial in log-moneyness to obtain estimates of $a_1(\tau)$, $a_2(\tau)$, $a_3(\tau)$ and $a_4(\tau)$ for those τ that are observed on that day. The a_0 estimates include the small component Φ_t, and we discuss only the $a_1, \cdots, a_4$ fits here.

Figure 3 shows some typical quartic fits of S&P500 implied volatilities for a few maturities. Here we use a wider range of strikes than in the linear fit shown in Figure 1, in particular in the out-of-the-money direction. We see from these plots that the quartic produced by the second-order approximation becomes important in capturing the turn of the skew. In these fits, it is important to fit the main body of the skew to an affine function of log-moneyness first (corresponding to the first-order approximation presented in section 3.2), and then fit the remainder

$$\frac{I - (a_0 + a_1(LM))}{(LM)^2}$$

to a quadratic in moneyness LM (in practice, LM is shifted to $LM + 1$ to avoid divide-by-zero issues). This split procedure is necessary because a free one-stage fit often uses the freedom of the quartic to catch stray data points, leading to large estimates of a_3 and a_4. By viewing the wings as small corrections to the linear skew, we avoid ‘tail wagging the dog’ phenomena.

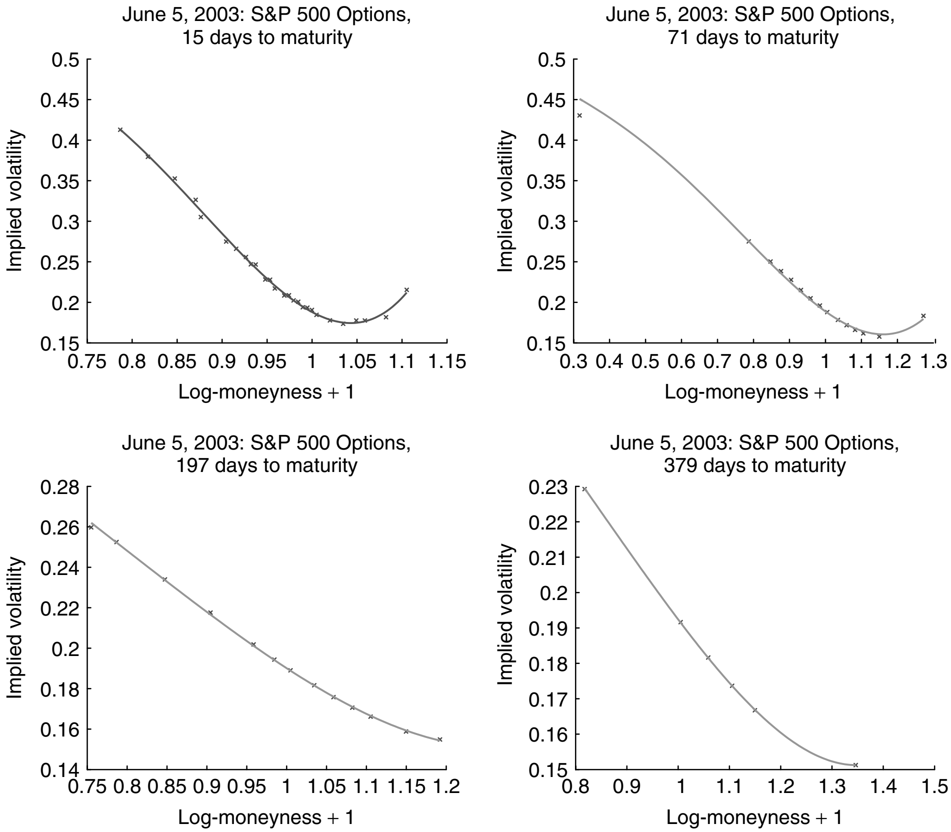


Figure 3: S&P500 implied volatility data on June 5, 2003 and quartic fits to the asymptotic theory for four maturities

Then, we fit the quartic coefficients to the following term-structure formulas coming from the asymptotics:

$$
\begin{aligned}
a_1(\tau) &= \sum_{k=-1}^{2} a_{1,k}\tau^k \\
a_2(\tau) &= \sum_{k=-2}^{1} a_{2,k}\tau^k \\
a_3(\tau) &= \sum_{k=-1}^{0} a_{3,k}\tau^k \\
a_4(\tau) &= \sum_{k=-2}^{-1} a_{4,k}\tau^k.
\end{aligned} \tag{9}
$$

The calibrated parameters $\{a_{j,k}\}$ play the role played by (b_0, b_1, m_0, m_1) in the first-order theory.

Figure 4 shows the fits of the $a(\tau)$'s to their term-structure formulas for S&P500 data on June 5, 2003.

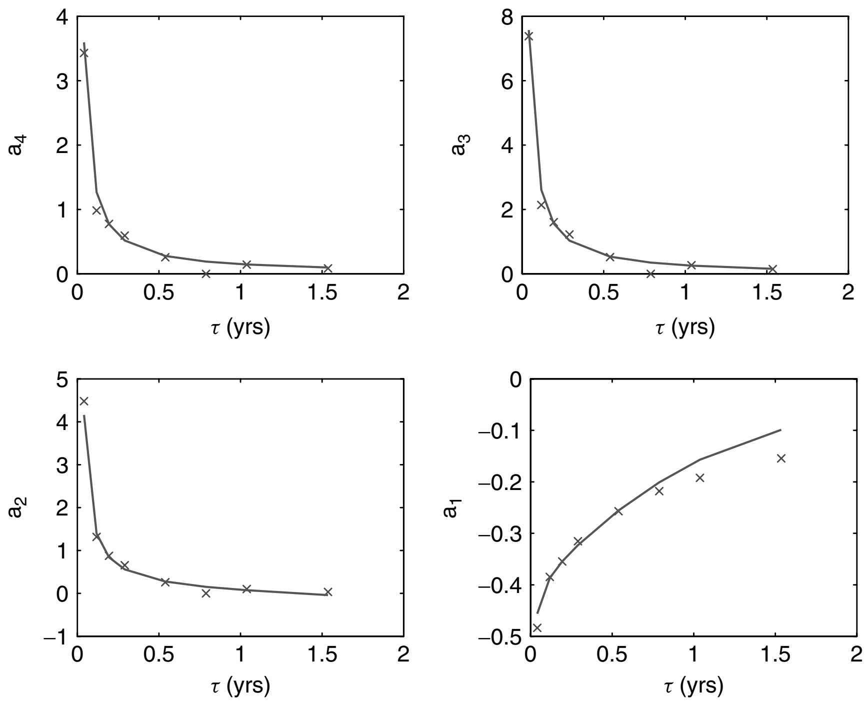


Figure 4: S&P500 term-structure fit using second-order approximation. Data from June 5, 2003

As discussed in the introduction, one of the main issues in volatility calibration is the stability with respect to t of the parameter estimates. To illustrate this point we carried out the quartic fits on S&P500 implied volatilities collected over the course of a month, We obtain estimates of $a_1(\tau)$, $a_2(\tau)$, $a_3(\tau)$ and $a_4(\tau)$ for those τ that are observed over this period. Figure 5 shows the fits of $a_1, \cdots, a_4$ to their corresponding term-structure formulas given in (9). The reasonable fits shown in Figure 5, using a month's data, demonstrate the stability of the approximation over some time. We remark that the a_1 estimates become less structured at small maturities because of a periodic maturity cycle component due to the option expiration ('witching') dates the third Friday of each month. This is studied in detail in Fouque *et al*. (2004a).

The final step is to recover the parameters needed for pricing from the estimates of $\{a_{j,k}\}$, the analog of (5) in the first-order theory. However, these relations are no longer linear in the second-order theory, and a non-linear inversion algorithm is required. This aspect has to be treated case by case in order to take advantage of the particular features of the market under study. For instance in FX markets, the correlation between the underlying and its volatility tends to be zero which reduces the complexity of the implementation of the second-order theory.

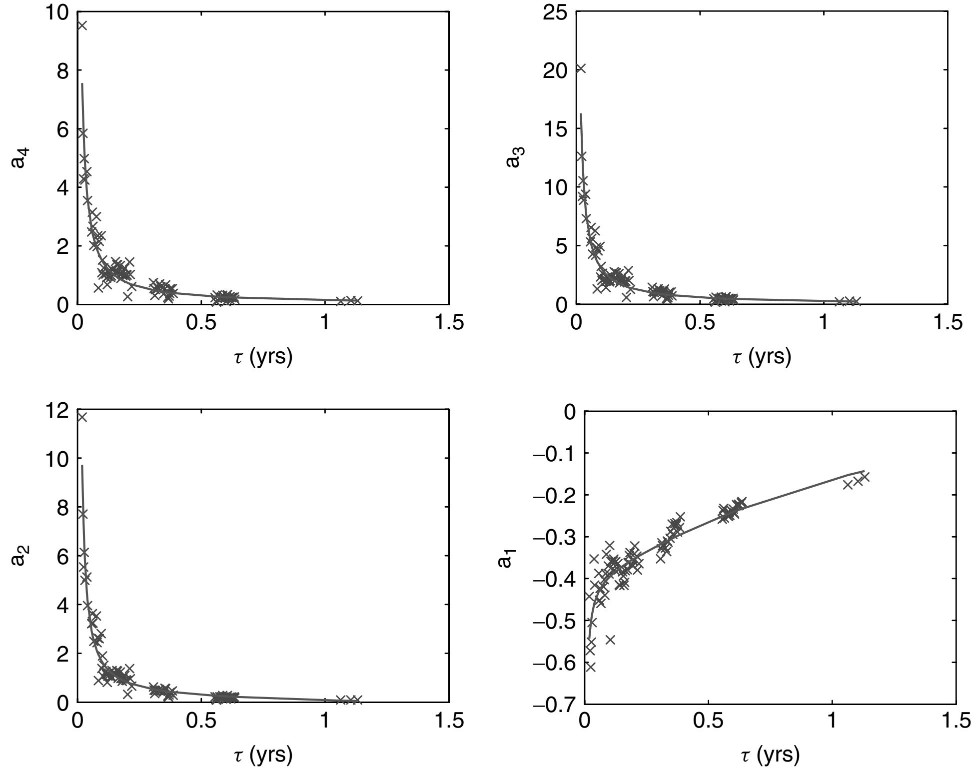


Figure 5: S&P500 term-structure fit. Data from every trading day in May 2003

REFERENCES

■ Alizadeh, S., Brandt, M. and Diebold, F. (2002) Range-based estimation of stochastic volatility models. *Journal of Finance*, 57(3): 1047–91.

■ Andersen, T. and Bollerslev, T. (1997) Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4, 115–158.

■ Bakshi, G., Cao, C. and Chen, Z. (1997) Empirical performance of alternative option pricing models. *Journal of Finance*, 52(5): 2003–2049.

■ Carr, P., Geman, H., Madan, D. and Yor, M. (2000) The fine structure of asset returns: an empirical investigation. Working paper.

■ Chernov, M., Gallant, R., Ghysels, E. and Tauchen, G. (2003) Alternative models for stock price dynamics. *Journal of Econometrics*, 116: 225–257.

■ Cont, R. and da Fonseca, J. (2002) Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1): 45–60.

■ Duffie, D., Pan, J. and Singleton, K. (2000) Transform analysis and option pricing for affine jump-diffusions. *Econometrica*, 68, 1343–1376.

■ Engle, R. and Patton, A. (2001) What good is a volatility model? *Quantitative Finance*, 1, 237–245, March.

■ Fouque, J. P., Papanicolaou, G. and Sircar. K. R. (2000) *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press.
■ Fouque, J. P., Papanicolaou, G., Sircar, K. R. and Solna, K. (2004a) Maturity cycles in implied volatility. *Finance & Stochastics*, 8(4): 451–77.
■ Fouque, J.-P., Papanicolaou, G., Sircar K. R. and Solna. K. (2003a) Short time-scale in S&P 500 volatility. *Journal of Computational Finance*, 6(4): 1–23.
■ Fouque, J. P., Papanicolaou, G., Sircar, K. R. and Solna, K. (2003b) Singular perturbations in option pricing. *SIAM J. Applied Math.*, 63(5): 1648–1665.
■ Fouque, J. P., Papanicolaou, G., Sircar, K. R. and Solna, K. (2004b) Multiscale stochastic volatility asymptotics. *SIAM Journal Multiscale Modeling and Simulation*, 2(1): 22–42.
■ Fouque, J. P., Papanicolaou, G., Sircar, K. R. and Solna, K. (2004c) Volatility perturbations in financial markets. In preparation.
■ Fournie, E., Lebuchoux, J. and Touzi, N. (1997) Small noise expansion and importance sampling. *Asymptotic Analysis*, 14(4): 361–376.
■ Haug, E. G. (2003) Know your weapon, Parts 1 and 2. *Wilmott*, May and August.
■ Hillebrand, E. (2003) Overlaying time scales and persistence estimation in GARCH(1,1) models. Preprint.
■ LeBaron, B. (2001) Stochastic volatility as a simple generator of apparent financial power laws and long memory. *Quantitative Finance*, 1(6): 621–631, November.
■ Lee, R. (1999) Local volatilities under stochastic volatility. *International Journal of Theoretical and Applied Finance*, 4(1): 45–89.
■ Lewis, A. (2000) *Option Valuation under Stochastic Volatility*. Finance Press, Newport Beach, CA.
■ Melino, A. and Turnbull, S. (1990) Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45, 239–265.
■ Muller, U., Dacorogna, M., Dave, R., Olsen, R., Pictet, O. and von Weizsacker, J. (1997) Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2–3): 213–239, June.
■ Sircar, K. R. and Papanicolaou, G. C. (1999) Stochastic volatility, smile and asymptotics. *Applied Mathematical Finance*, 6(2): 107–145.
■ Wilmott, P. (2000) *Paul Wilmott on Quantitative Finance*. Volume 2, John Wiley & Sons, Ltd.

25 Inference and Stochastic Volatility

Alireza Javaheri*

1 Introduction

Consider a stochastic volatility mode such as the square-root (Lewis Alan 2000) model

$$dS_t/S_t = \mu_S dt + \sqrt{v_t}dB_t$$
$$dv_t = (\omega - \theta v_t)dt + \xi\sqrt{v_t}dZ_t$$

with Brownian motions $< dB_t, dZ_t >= \rho$ as usual.

The Euler log-normal equations corresponding to discrete observations would be

$$LnS_{k+1} = LnS_k + (\mu_S - \tfrac{1}{2}v_k)\Delta t + \sqrt{v_k}\sqrt{\Delta t}B_k$$
$$v_{k+1} = v_k + (\omega - \theta v_k)\Delta t + \xi\sqrt{v_k}\sqrt{\Delta t}Z_k$$

with (B_k) and (Z_k) temporally uncorrelated Gaussian random variables with a mutual correlation ρ.

Considering μ_S known, one could attempt to infer the parameter set $\Psi = (\omega, \theta, \xi, \rho)$ from a given time series of N asset prices $(S_k)_{1\le k\le N}$. This could be accomplished via various methods such as maximization of likelihood as suggested by Fridman and Harris (1998) and Javaheri *et al*. (2003), or via Markov chain Monte Carlo algorithms as done by Kim *et al*. (1998) and Jacquier *et al*. (1994).

Much of the recent financial econometrics literature (Bakshi *et al*. 1997, Bates 2000) uses these inference methodologies to estimate the embedded stochastic volatility parameters from the time series under the statistical measure and then compares them to those obtained from options markets

---

*Based on a dissertation supervised by Prof. Alain Galli, *Ecole des Mines de Paris*.

Alireza Javaheri is a Quantitative Analyst at Citigroup in the Fixed Income Derivatives Research area. The opinions expressed in this article are solely those of the author and do not reflect any views by Citigroup.

Alireza Javaheri would like to thank the participants at the WILMOTT Technical Forum and in particular Vladimir Piterbarg for their helpful comments.

under the risk-neutral measure. According to the Girsanov theorem there should be a consistency between the two measures in that the parameters (ξ,ρ) should be identical for the two markets.

In practice, however, researchers observe a much higher estimated value for ξ and $|\rho|$ from the options markets. They then conclude this could mean there is a model misspecification or a trade opportunity.

The object of this chapter is to see how reliable the estimations from the time series actually are. Indeed even if the Maximum Likelihood Estimators (MLE) are *asymptotically* unbiased and efficient, are we sure that we have enough data-points to make strong inference-based conclusions?

In our following study we use the filtered MLE method as described in Javaheri *et al*. (2003).

2 The inference tests

2.1 Single parameter estimation

A known weakness of optimization algorithms is the following. The higher the number of parameters, the worse the performance of the algorithm. This means that a one-parameter optimization should perform best. To test this, we simulate 5000 points via the Heston model with a parameter set Ψ^* as shown in Table 1 (see also Figure 1).

We use a drift of $\mu_S = 0.025$ and a time step $\Delta t = 1/252$ as before.

In order to get the best performance we fix all parameters except one. For instance to obtain $\hat{\omega}$ we fix $\theta = 10.0, \xi = 0.03, \rho = -0.50, \mu_S = 0.025$, we choose a reasonable initial point ω_0 and then optimize upon ω only. We choose an initial parameter-set Ψ_0 as shown in Table 2. The results are displayed in the following tables. See Javaheri *et al*. (2003) for an explanation on EKF, EPF, UKF and UPF.

TABLE 1: THE TRUE PARAMETER-SET Ψ^* USED FOR DATA SIMULATION

Ψ^*	$\omega^* = 0.10$	$\theta^* = 10.0$	$\xi^* = 0.03$	$\rho^* = -0.50$

TABLE 2: THE INITIAL PARAMETER-SET Ψ_0 USED FOR THE OPTIMIZATION PROCESS

Ψ_0	$\omega_0 = 0.15$	$\theta_0 = 15.0$	$\xi_0 = 0.02$	$\rho_0 = -0.50$

TABLE 3: THE OPTIMAL PARAMETER-SET $\hat{\Psi}$. THE ESTIMATION IS PERFORMED INDIVIDUALLY FOR EACH PARAMETER ON THE ARTIFICIALLY GENERATED TIME SERIES. PARTICLE FILTERS USE 1000 SIMULATIONS

Filter	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
EKF	0.098212	10.188843	0.052324	−0.873571
UKF	0.107281	10.089381	0.000001	+0.598434
EPF	0.098287	10.130531	0.044437	−0.827729
UPF	0.100581	10.221816	0.051902	−0.487695

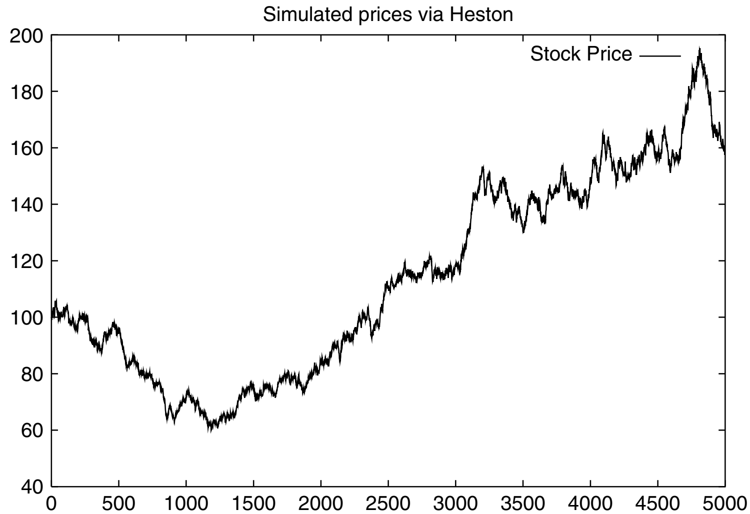


Figure 1: Simulated stock-price path via Heston using Ψ^*

It is interesting to note that the estimation of the volatility-drift parameters (ω, θ) could be done fairly well via EKF. This makes sense since the dependence on these parameters is linear.

The estimation of volatility and correlation parameters (ξ, ρ) is not as straightforward. This could be seen by plotting the likelihood $L(\Psi)$ as a function of ω, θ, ξ and ρ separately. We fix three parameters to their optimal values and plot $L(\Psi)$ as a function of the last one. We observe in Figures 2 to 5 that the likelihood function is fairly easy to optimize for (ω, θ). However, the function is very *flat* around the optimal ξ and ρ. Hence the difficulty of finding the optimums!

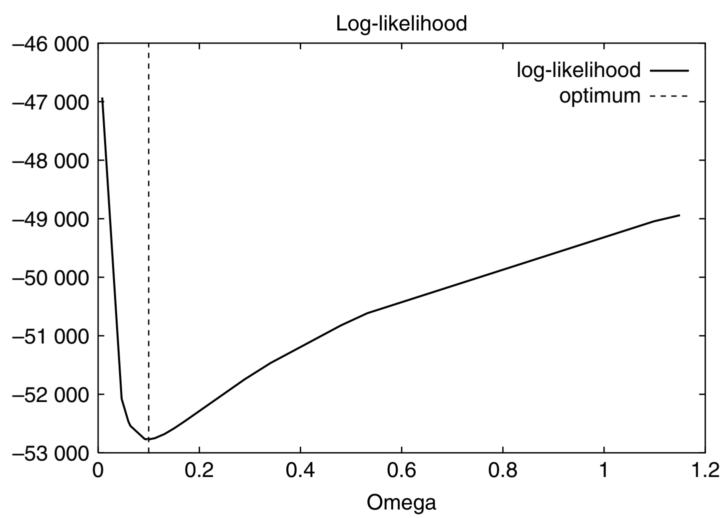


Figure 2: $f(\omega) = L(\omega, \hat{\theta}, \hat{\xi}, \hat{\rho})$ has a good slope around $\hat{\omega} = 0.10$

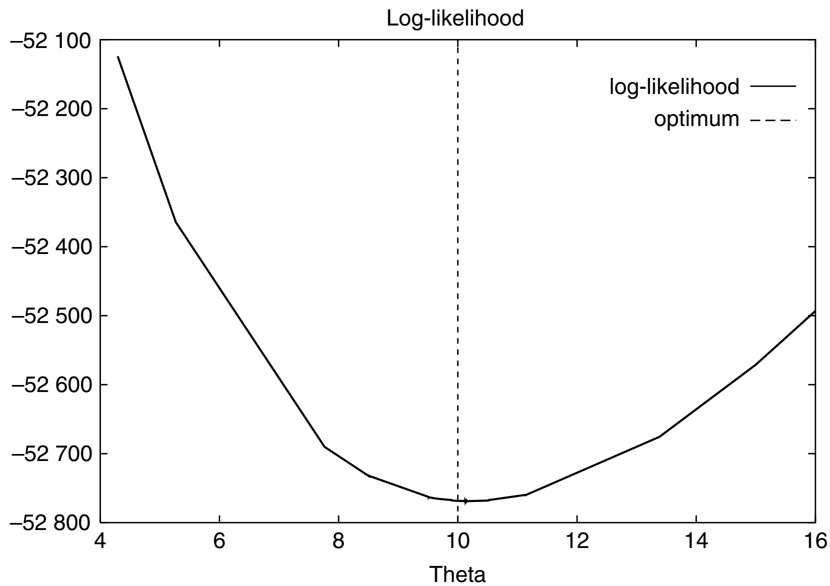


Figure 3: $f(\theta) = L(\hat{\omega}, \theta, \hat{\xi}, \hat{\rho})$ **has a good slope around** $\hat{\theta} = 10.0$

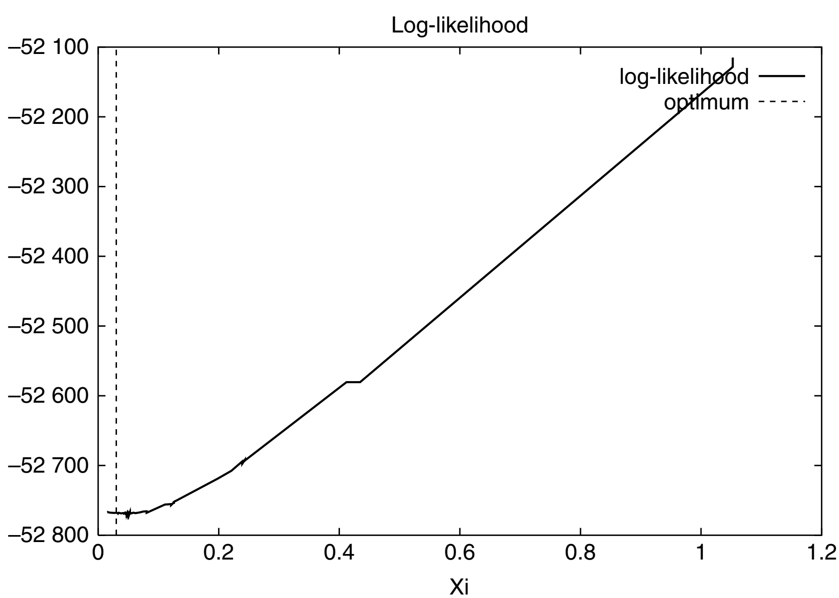


Figure 4: $f(\xi) = L(\hat{\omega}, \hat{\theta}, \xi, \hat{\rho})$ **is flat around** $\hat{\xi}$**=0.03**

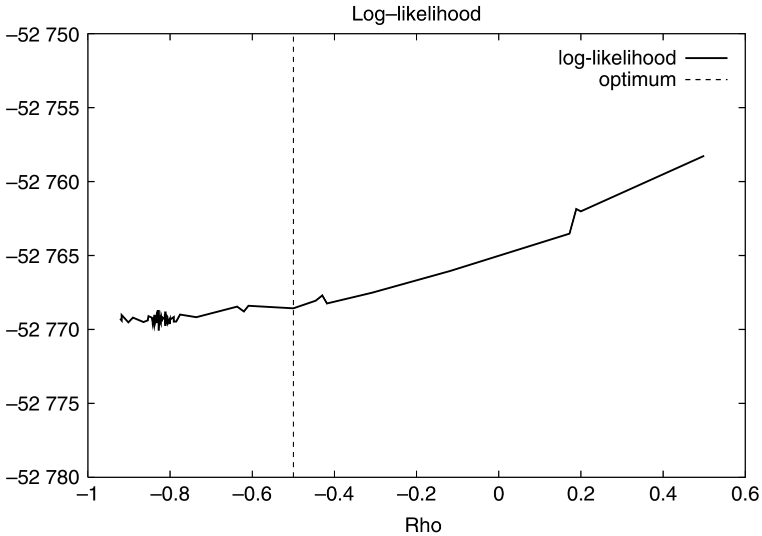


Figure 5: $f(\rho) = L(\hat{\omega}, \hat{\theta}, \hat{\xi}, \rho)$ **is flat and irregular around** $\hat{\rho} = -0.50$

2.2 Sample size

It seems therefore that the estimation is inefficient for the parameter ξ no matter which filter we use. The issue is that of inefficiency (large error variance) for this given sample size. This is indeed one of the shortcomings of the Maximum Likelihood Estimators (MLE). For a given sample size they can very well be inefficient and even have a bias. The choice of the filter will not solve this issue. However (under minimal regularity conditions) MLEs are consistent and therefore asymptotically converge to the correct optimum. This means that the sample size is key.

To test this we can choose larger samples of $N = 50\,000$, $N = 100\,000$ and $N = 500\,000$ points and rerun the simplest filter, namely the EKF. As expected the optimum of the likelihood function becomes closer and closer to ξ^*. This can be seen in Figures 6 to 9 as well as in Table 4.

TABLE 4: THE OPTIMAL EKF PARAMETERS $\hat{\xi}$ AND $\hat{\rho}$ GIVEN A SAMPLE SIZE N. THE TRUE PARAMETERS ARE $\xi^* = 0.03$ AND $\rho^* = -0.50$. THE INITIAL VALUES WERE $\xi_0 = 0.03$ AND $\rho_0 = -0.040$

N	$\hat{\xi}$	$\hat{\rho}$
5000	0.052324	−0.873571
50 000	0.036463	−0.608088
100 000	0.033400	−0.556868
500 000	0.031922	−0.532142

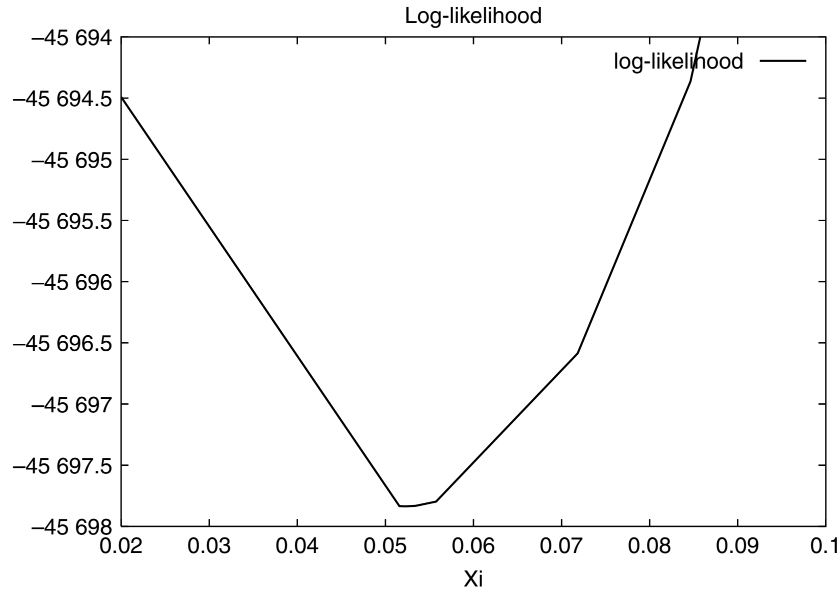


Figure 6: $f(\xi) = L(\hat{\omega}, \hat{\theta}, \xi, \hat{\rho})$ **via EKF for** $N = 5000$ **points. The true value is** $\xi^* = 0.03$

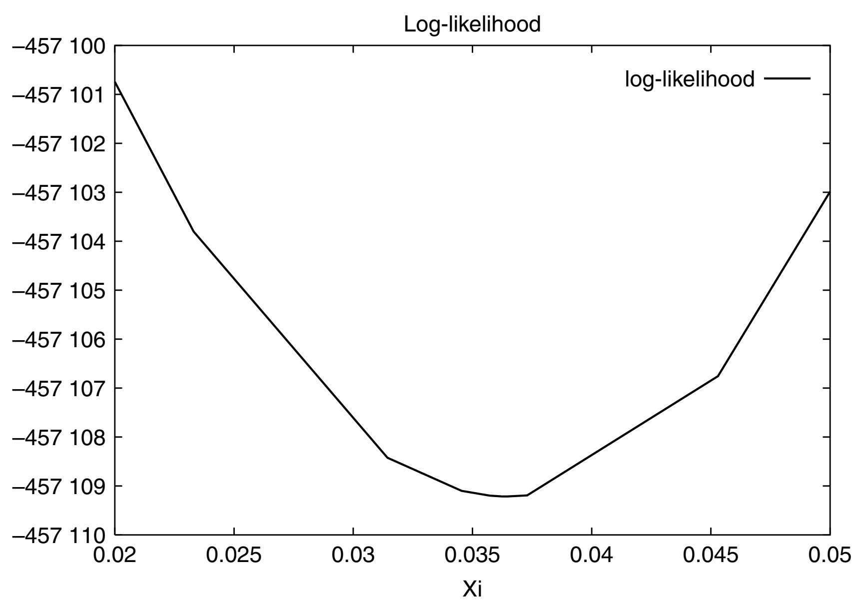


Figure 7: $f(\xi) = L(\hat{\omega}, \hat{\theta}, \xi, \hat{\rho})$ **via EKF for** $N = 50\,000$ **points. The true value is** $\xi^* = 0.03$

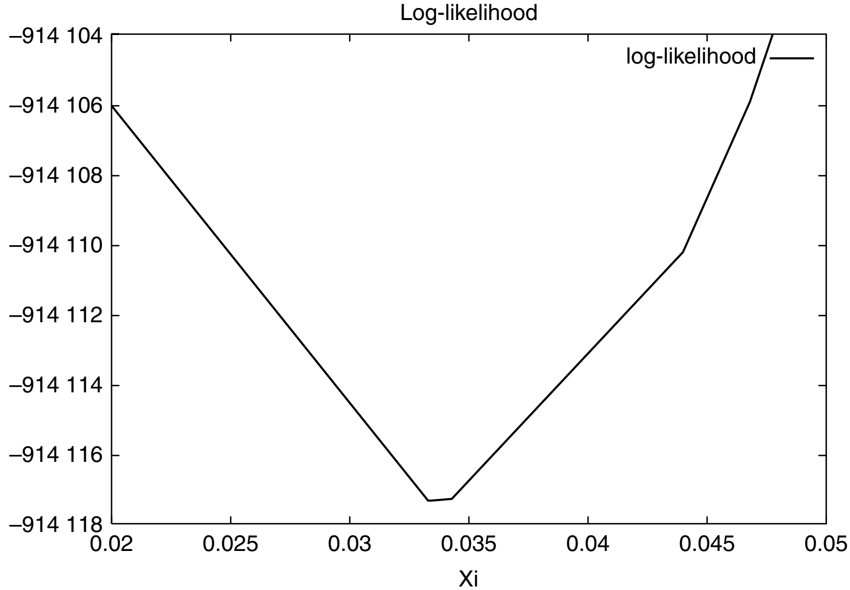


Figure 8: $f(\xi) = L(\hat{\omega}, \hat{\theta}, \xi, \hat{\rho})$ **via EKF for** $N = 100\,000$ **points. The true value is** $\xi^* = 0.03$

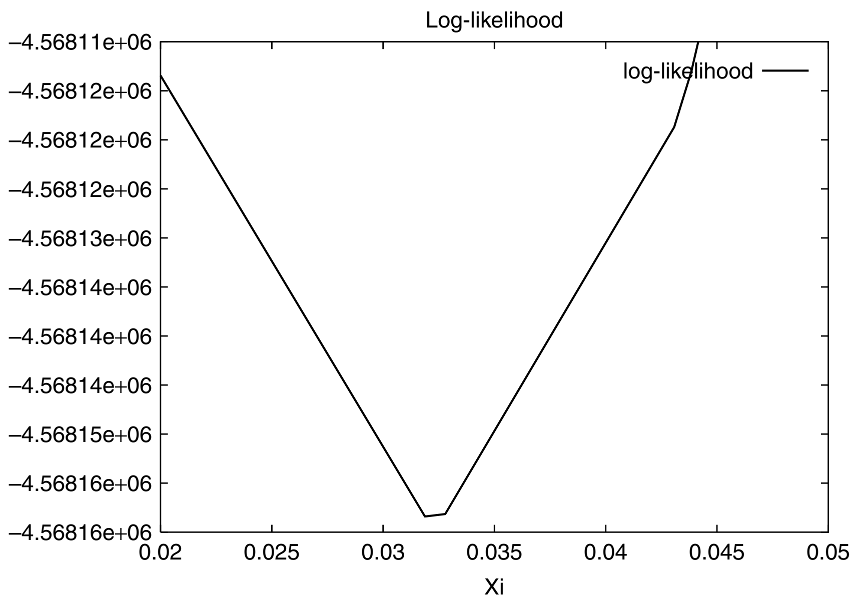


Figure 9: $f(\xi) = L(\hat{\omega}, \hat{\theta}, \xi, \hat{\rho})$ **via EKF for** $N = 500\,000$ **points. The true value is** $\xi^* = 0.03$

The same exact observations could be made for the correlation parameter ρ and the results are displayed in the same Table 4. The likelihood graphs are omitted in the interest of brevity.

As for the drift parameters ω and θ, the convergence was good even for $N = 5000$ as previously observed.

Unfortunately in reality we have limited historic data. Even at a daily frequency 50 000 points would correspond to 200 years!

One possibility would be to use intra-day data; however, that assumes that the behavior of the stock price is the same intra-day (which is reasonable considering we started with a *continuous* SDE). Moreover, clean intra-day data is usually not readily available and needs preprocessing. In any case, ultra-high frequency data has its known problems such as market micro-structures.

Therefore, having p parameters in the optimal parameter-set $\hat{\Psi}_N = \left(\hat{\Psi}_N[j]\right)_{1 \le j \le p}$ for a sample size N, we have for each parameter $\Psi[j]$

$$\lim_{N \to +\infty} \hat{\Psi}_N[j] \mid \left\{\Psi[k] = \Psi^*[k];\ 1 \le k \le p;\ k \ne j\right\} = \Psi^*[j] \tag{1}$$

What is more this is true for any valid initial value $\Psi_0[j]$, which means the MLE is robust.

2.3 Joint estimation of the parameters

Let us now assume that, as in reality, we do not know any of the parameters, choose an initial set Ψ_0 and test the consistency of the MLE. We shall apply the EKF to the data and take the same true parameter set Ψ^* as in the previous section (see Tables 5 and 6). We assume that $\mu_S = 0.025$ is known, otherwise it could be estimated together with the model parameters. The results are displayed in the following tables.

As previously mentioned, the likelihood function becomes flat and therefore harder to maximize under a higher number of parameters. The convergence of the estimator will therefore be slower.

Despite this, we can observe in Table 7 the asymptotic convergence of the estimator even under the joint estimation of all parameters.

Indeed we have now

$$\lim_{N \to +\infty} \hat{\Psi}_N = \Psi^* \tag{2}$$

which corresponds to the generalization of (1) in the previous section.

TABLE 5: THE TRUE PARAMETER-SET Ψ^* USED FOR DATA GENERATION

Ψ^*	$\omega^* = 0.10$	$\theta^* = 10.0$	$\xi^* = 0.03$	$\rho^* = -0.50$

TABLE 6: THE INITIAL PARAMETER-SET Ψ_0 USED FOR THE OPTIMIZATION PROCESS

Ψ_0	$\omega_0 = 0.15$	$\theta_0 = 15.0$	$\xi_0 = 0.02$	$\rho_0 = -0.40$

TABLE 7: THE OPTIMAL EKF PARAMETER-SET $\hat{\Psi}$ FOR A GIVEN SAMPLE SIZE N. THE FOUR PARAMETERS ARE ESTIMATED JOINTLY

N	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
5000	0.150854	15.294576	0.266175	−0.128835
50 000	0.126387	12.748852	0.020521	−1.000000
100 000	0.136023	13.700906	0.044353	−0.439961
500 000	0.100097	10.030336	0.061688	−0.257305
1 000 000	0.105264	10.548642	0.043818	−0.356234
2 000 000	0.103183	10.334876	0.039767	−0.374677
4 000 000	0.105292	10.538019	0.043288	−0.347562
5 000 000	0.101097	10.118951	0.028588	−0.514346

We ran other filters (UKF, EPF, UPF) on the same data set and observed only marginal improvement. The results are omitted for brevity. It therefore seems that the fundamental issue is related to the slow convergence of the MLEs *regardless of the filtering method*.

2.4 Error size

A related issue is the size of the observation error $u_k \propto \sqrt{\Delta t}$ which is large compared to the observation function $H_k \propto \Delta t$ for daily observations.

This underlines the more fundamental problem for the SV estimation: by definition, volatility represents the noise of the stock process. Indeed if we had taken the spot price S_k as the observation and the variance v_k as the state, we would have

$$S_{k+1} = S_k + S_k \mu_S \Delta t + S_k \sqrt{v_k} \sqrt{\Delta t} B_k$$

we would then have an observation function gradient $H = 0$ and the system would be *unobservable!*

It is precisely because we use a Taylor second-order expansion

$$\ln(1 + x) \approx x - \frac{1}{2} x^2$$

that we obtain access to v_k through the observation function. However, the error remains *dominant* as the first order of the expansion.[1]

Harvey *et al*. (1994) use the approximation $\Delta t = o(\sqrt{\Delta t})$ and take

$$z_k = \ln\left(\ln^2\left(\frac{S_{k+1}}{S_k}\right)\right) \approx \ln(v_k) + \ln(\Delta t) + \ln(B_k^2)$$

Note that under this form EKF would blow up since $z_k^- = h(v_k, 0) = -\infty$.

They therefore use the fact that $E[\ln(B_k^2)] = -1.27$ and $\text{stdev}[\ln(B_k^2)] = \pi/\sqrt{2}$ and consider the Gaussian approximation

$$\ln(B_k^2) \sim -1.27 + \frac{\pi}{\sqrt{2}}\mathcal{N}(0, 1)$$

which may or may not be valid. We call this approximation Harvey–Ruiz–Shephard (HRS) and apply it to the same case as in the previous paragraphs. As can be seen in Table 8 the approximation seems to be valid for our example. Note that UKF would not have this issue since we would work with the real non-linear function $z = h(x, u)$ above. However, we would still deal with logs of very small quantities which could be numerically unstable.

TABLE 8: THE OPTIMAL EKF PARAMETER-SET $\hat{\Psi}$ VIA THE HRS APPROXIMATION FOR A GIVEN SAMPLE SIZE N. THE FOUR PARAMETERS ARE ESTIMATED JOINTLY

N	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
5000	0.722746	71.753861	0.044602	−1.000000
50 000	0.234110	23.575193	0.028056	−1.000000
100 000	0.150512	15.186113	0.017748	−1.000000
500 000	0.109738	11.020391	0.027140	−0.531481

Another way of tackling the same equation would be via a particle filter where

$$z_k = \ln\left(\left|\ln\left(\frac{S_{k+1}}{S_k}\right)\right|\right) \approx \frac{1}{2}\ln(v_k) + \frac{1}{2}\ln(\Delta t) + \ln(|B_k|)$$

and as stated in Alizadeh *et al*. (2002) the density of $\ln(|B_k|)$ is

$$f(x) = 2e^x n(e^x)$$

with $n()$ the normal density.[2]

Testing the same data set provides Table 9 which does not seem to improve upon the KF.

TABLE 9: THE OPTIMAL PF PARAMETER-SET $\hat{\Psi}$ FOR A GIVEN SAMPLE SIZE N. THE FOUR PARAMETERS ARE ESTIMATED JOINTLY

N	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
5000	0.147212	14.999999	0.070407	−0.555263

2.5 High frequency data

Given that the results seem to converge for a large number of data points, one idea would be to use a higher sampling frequency. Indeed if instead of using daily data we sample every five seconds, on a ten year range we will have $10 \times 252 \times 6.5 \times 60 \times 60 \div 5 = 11\,793\,600$ data points which is very sufficient for our MLEs.

For testing the use of high frequency data, we can generate via Monte Carlo 5 000 000 points with a $\Delta t = 1/252\,000$ which corresponds to 20 years. We obtain the results in Table 10 below. Both rows have reasonable results. It is, however, notable that the EKF/HRS method seems to perform better than the plain EKF.

TABLE 10: THE OPTIMAL PARAMETER-SET $\hat{\Psi}$ FOR 5 000 000 DATA POINTS. THE SAMPLING IS PERFORMED 1000 TIMES A DAY AND THEREFORE THE DATA-SET CORRESPONDS TO 5000 BUSINESS DAYS. THE FOUR PARAMETERS ARE ESTIMATED JOINTLY

	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
EKF	0.090280	9.019962	0.042984	-0.283236
EKF/HRS	0.092372	9.224421	0.030951	-0.507763

2.6 Sampling distribution

Even if in practice we deal with one historic path, we should determine the distribution of the optimal parameter-set as follows.

We simulate $P = 500$ paths of length $N = 5000$ and estimate for each path j the optimal set $\hat{\Psi}^{(j)}$. We can then estimate

$$\overline{\hat{\Psi}} = \frac{1}{P} \sum_{j=0}^{P-1} \hat{\Psi}^{(j)}$$

as well as the variance

$$V(\hat{\Psi}) = \frac{1}{P} \sum_{j=0}^{P-1} (\hat{\Psi}^{(j)} - \overline{\hat{\Psi}})^2$$

This way we will know how the estimator performs on average and how far we could be from this average. The distribution of the parameter-set around its mean is referred to as the *sampling distribution*.

As we can see in Table 11 the average-estimated parameter-set is closer to the true-set than the one-path-estimated-set we were considering in the previous section. However, the corresponding standard deviation is quite high and we could very well get poor results as previously seen.

TABLE 11: MEAN (AND STANDARD DEVIATION) FOR THE ESTIMATION OF EACH PARAMETER VIA EKF OVER $P = 500$ PATHS OF LENGTHS $N = 5000$ AND $N = 50\,000$. THE TRUE VALUES ARE ($\omega^* = 0.10$, $\theta^* = 10$, $\xi^* = 0.03$, $\rho^* = -0.50$)

	$\hat{\omega}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\rho}$
$N = 5000$	0.11933899	11.92271488	0.056092146	−0.34321724
	(0.098995729)	(9.673829518)	(0.049741887)	(0.297433861)
$N = 50\,000$	0.102554592	10.26233092	0.04383931	−0.351998284
	(0.027020734)	(2.706564396)	(0.013004526)	(0.074998408)

From Figures 10 to 13 we can see that for this data length N and this sample size P the parameters ω and θ are determined via EKF in a fairly unbiased way. However, the estimator is not efficient and has a large standard deviation. As for ξ and ρ we have both bias and inefficiency.

This is not surprising given the results of the previous paragraphs. We obtained good results for (ω, θ) when estimated alone, and not so good results for (ξ, ρ). Classical filtering and estimation theories work well when the parameters affect the drift of the observation and not the noise. This causes a slow convergence issue for all our parameters. But this is doubly true for (ξ, ρ) since they affect the 'noise of the noise'.

As previously observed the bias and inefficiency will disappear as $N \to +\infty$ as is the case for any MLE estimator. Indeed the biases and the standard deviations are smaller for $N = 50\,000$ than for $N = 5000$ as we can see in Table 11.

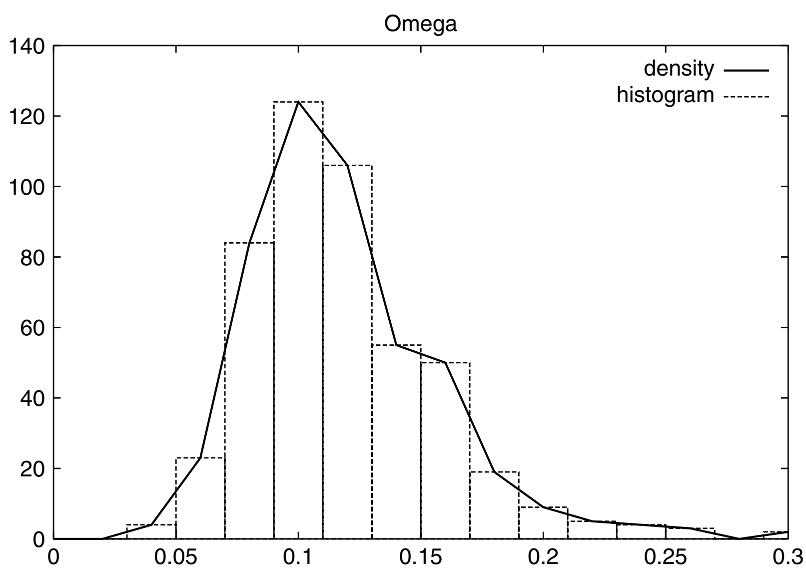


Figure 10: Density for $\hat{\omega}$ estimated from 500 paths of length 5000 via EKF. The true value is $\omega^* = 0.10$

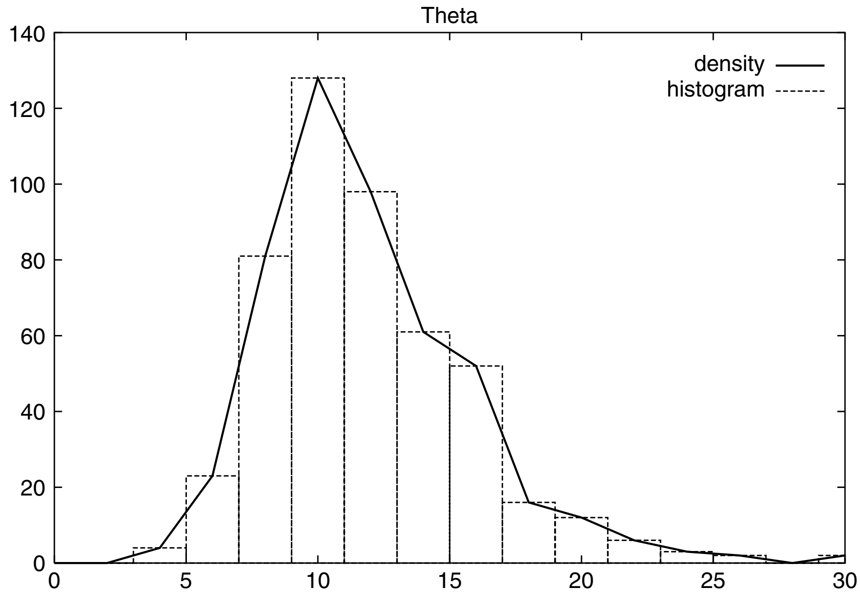


Figure 11: Density for $\hat{\theta}$ estimated from 500 paths of length 5000 via EKF. The true value is $\theta^* = 10$

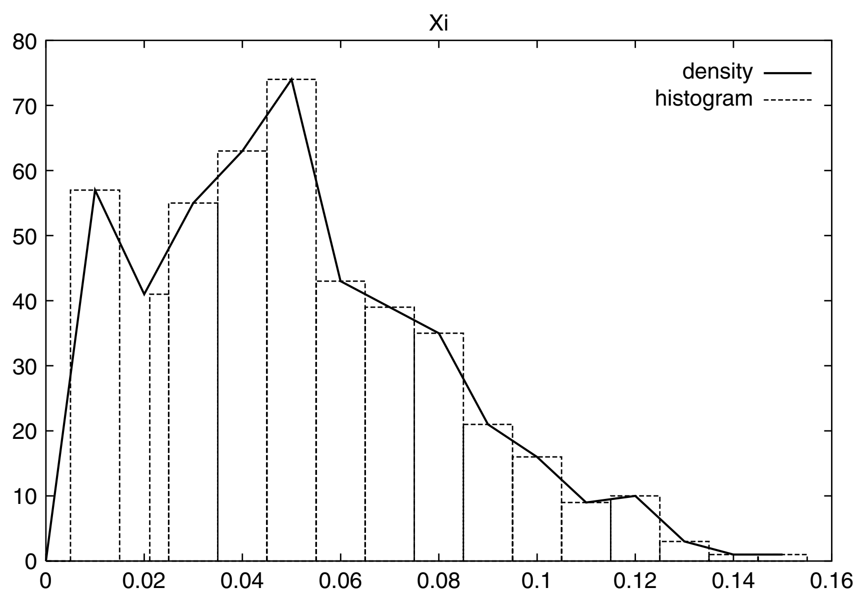


Figure 12: Density for $\hat{\xi}$ estimated from 500 paths of length 5000 via EKF. The true value is $\xi^* = 0.03$

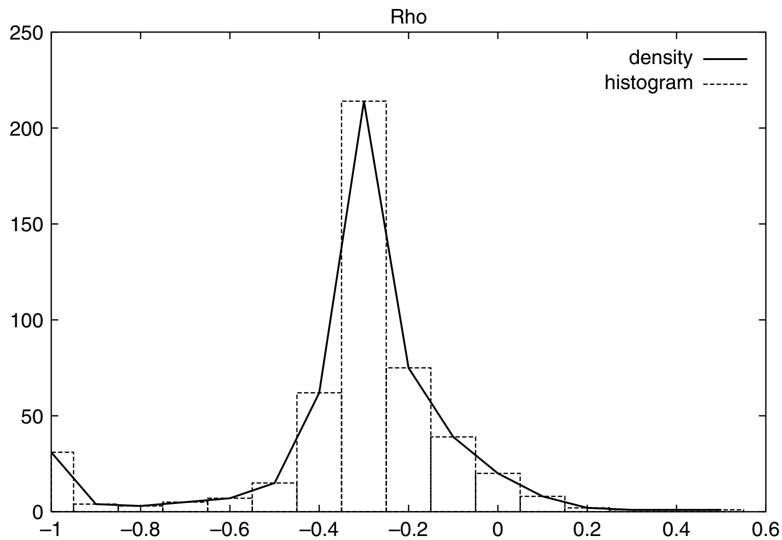


Figure 13: Density for $\hat{\rho}$ estimated from 500 paths of length 5000 via EKF. The true value is $\rho^* = -0.50$

3 Conclusion

As we can see inferring parameters from a time series of a limited length could be very dangerous. This does not mean that the estimations are always wrong but rather that they *could* very well be wrong given the size of our estimation error.

In order to check whether there actually is inconsistency between the assets and the options market, an interesting test was suggested in Ait-Sahalia *et al*. (2001). One could use the profits generated from a *skewness trade* (buying out-of-the-money calls and selling out-of-the-money puts) as an empirical and model-free measure of the consistency between the two markets. If there is no conclusive and clear profit generated, this means that the discrepancy could be artificial and due to the inaccuracy of the time-series estimators.

FOOTNOTES & REFERENCES

1. Note that this is different from a variance swap where we work with the *expected values*. Indeed the approximation is perfectly valid if for the return $R = \Delta S/S$ we write

$$E[\ln(1+R) - R] \approx -\frac{1}{2}v$$

but again, the approximation breaks if we work for one sample path.
2. It is easy to see that if X is a standard Normal variable, then the CDF of $\ln(|X|)$ is

$$F(x) = P(\ln(|X|) \leq x) = P(|X| \leq e^x) = P(-e^x \leq X \leq e^x)$$

therefore

$$F(x) = N(e^x) - N(-e^x) = 2N(e^x) - 1$$

and the density is determined by taking the derivative with respect to x as usual.

■ Aït-Sahalia, Y., Wang, Y. and Yared, F. (2001) Do option markets correctly price the probabilities of movement of the underlying asset? *Journal of Econometrics*, 101.
■ Alizadeh, S., Brandt, M. W. and Diebold, F. X. (2002) Range-based estimation of stochastic volatility models. *Journal of Finance*, Vol. 57, No. 3.
■ Bakshi, G., Cao, C. and Chen, Z. (1997) Empirical performance of alternative option pricing models. *Journal of Finance*, Vol. 52, Issue 5.
■ Bates, D. S. (2000) Post-87 crash fears in the S&P500 futures option market. *Journal of Econometrics*, 94.
■ Fridman, M. and Harris, L. (1998) A maximum likelihood approach for non-Gaussian stochastic volatility models. *Journal of Business and Economic Statistics* 16:3, 284–91.
■ Harvey, A. C., Ruiz, E. and Shephard, N. (1994) Multivariate stochastic variance models. *Review of Economic Studies*, Vol. 61, Issue 2.
■ Jacquier, E., Polson, N. G. and Rossi, P. E. (1994) Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, Vol. 12, No. 4.
■ Javaheri, A., Lautier, D. and Galli, A. (2003) Filtering in finance. *Wilmott*, Issue 5.
■ Sangjoon, K., Shephard, N. and Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, Vol. 65.
■ Lewis, A. L. (2000) *Option Valuation under Stochastic Volatility*. Finance Press.

26
A Critique of the Crank Nicolson Scheme Strengths and Weaknesses for Financial Instrument Pricing

Daniel J. Duffy

In this chapter we apply the Finite Difference Method (FDM) to the Black–Scholes equation. In particular, we analyse the famous Crank Nicolson method that is very popular in financial engineering. Unfortunately, the method does not always produce accurate results and it is the objective of this chapter to enumerate the problems and then to propose more robust finite difference schemes. More detailed accounts of the current problem can be found in Duffy (2001, 2004).

1 A short history of Crank Nicolson in financial engineering

The Crank Nicolson finite difference scheme was invented by John Crank and Phyllis Nicolson. They originally applied it to the heat equation and they approximated the solution of the heat equation on some finite grid by approximating the derivatives in space x and time t by finite differences. Much earlier, Richardson devised a finite difference scheme that was easy to compute but was numerically unstable and thus useless. The instability was not recognized until Crank,

Contact address: Datasim Component Technology, B.V., Schipluidenlaan 4, 1062 HE Amsterdam, The Netherlands.
E-mail: dduffy@datasim.nl

Nicolson and others carried out lengthy numerical calculations. In short, the Crank Nicolson method is numerically stable and it only requires the solution of a very simple system of linear equations (namely, a tridiagonal system) at every time level.

The Crank Nicolson method has become one of the most popular finite difference schemes for approximating the solution of the Black–Scholes equation and its generalizations (see, for example, Tavella 2000, Bhansali 1998). The method is essentially a second-order approximation to the time derivative that appears in the Black–Scholes equation and this property, plus the fact that the method is stable and is easy to program, makes it very appealing in practical applications. Numerous articles and publications in the financial engineering literature use Crank Nicolson as the de-facto scheme for time discretization. Unfortunately, the method breaks down in certain situations and there are better and more robust alternatives that have been documented in the numerical analysis and computational fluid dynamics literature. To this end, we wish to discuss the shortcomings of the method and how they can be resolved.

2 What is Crank Nicolson, really?

The one-factor Black–Scholes equation for a derivative quantity V depending on an underlying S is given by

$$-\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0. \tag{1}$$

In general, this equation must be augmented by other boundary and initial conditions in order to ensure a unique solution. In some cases it may be possible to come up with a exact solution to this problem but in the most general cases we must resort to some kind of approximate method. In this chapter we discuss the Finite Difference Method and it is based on the tactic of replacing the continuous derivatives in (1) by *divided differences* defined on a discrete mesh (see Richtmyer 1967).

In order to motivate the Crank Nicolson scheme let us first consider the following *fully implicit scheme* that we define by replacing derivatives with respect to S by three-point divided differences and the derivative with respect to t by one-sided differences. The scheme is given by

$$\begin{aligned}
&-\frac{V_j^{n+1} - V_j^n}{k} + rj\Delta S\left(\frac{V_{j+1}^{n+1} - V_{j-1}^{n+1}}{2\Delta S}\right)\\
&+\frac{1}{2}\sigma^2 j^2 \Delta S^2\left(\frac{V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}}{\Delta S^2}\right)\\
&= rV_j^{n+1}.
\end{aligned} \tag{2}$$

In general, the values of V at time level n are known and the values at time level $n+1$ need to be calculated. Rewriting (2) gives the new form

$$\begin{cases} a_j^{n+1}\, V_{j-1}^{n+1} + b_j^{n+1}\, V_j^{n+1} + c_j^{n+1}\, V_{j+1}^{n+1} = F_j^{n+1} \\ \text{where} \\ a_j^{n+1} = \left(\dfrac{1}{2}\sigma^2 j^2 k - \dfrac{krj}{2}\right) \\ b_j^{n+1} = -\left(1 + \sigma^2 j^2 k + r\right) \\ c_j^{n+1} = \left(\dfrac{1}{2}\sigma^2 j^2 k + \dfrac{krj}{2}\right) \\ F_j^{n+1} = -\, V_j^n. \end{cases} \tag{3}$$

This is a tridiagonal scheme that we solve at each time level using standard matrix solvers, for example LU decomposition (see Isaacson 1966, Duffy 2004). The fully implicit scheme has a number of desirable features. First, it is stable and there is no restriction on the relative sizes of the time mesh size k and the space mesh size ΔS. Furthermore, no spurious oscillations are to be seen in the solution or its Δ (as is the case with some other methods). A disadvantage is that it is only first-order accurate in k. On the other hand, this can be rectified by using extrapolation and this results in a second-order scheme.

Crank Nicolson is a variation of (2) but in this case we take averages of V at levels n and $n+1$ when approximating the derivative with respect to t. We define the quantity

$$V_j^{n+\frac{1}{2}} \equiv \frac{1}{2}\left(V_j^{n+1} + V_j^n\right). \tag{4}$$

Then the Crank Nicolson method is defined as follows:

$$\begin{aligned} &-\frac{V_j^{n+1} - V_j^n}{k} + rj\Delta S\left(\frac{V_{j+1}^{n+\frac{1}{2}} - V_{j-1}^{n+\frac{1}{2}}}{2\Delta S}\right) \\ &+\frac{1}{2}\sigma^2 j^2 \Delta S^2\left(\frac{V_{j+1}^{n+\frac{1}{2}} - 2V_j^{n+\frac{1}{2}} + V_{j-1}^{n+\frac{1}{2}}}{\Delta S^2}\right) \\ &= \tau V_j^{n+\frac{1}{2}}. \end{aligned} \tag{5}$$

Again, this is a system that can be posed in the form (3) and hence can be solved by standard matrix solver techniques at each time level.

The Crank Nicolson method has gained wide acceptance in the financial literature and it seems to be the de-facto finite difference scheme for one-factor and two-factor Black–Scholes equations. It has second-order accuracy in the parameter k and is stable. Unfortunately, it has been known for some considerable time (Il'in 1969) that centred differencing schemes in space combined with averaging in time (what essentially CN is in this context) lead to spurious oscillations in the approximate solution. These oscillations have nothing to do with the physical or financial problem that the scheme is approximating.

3 The problems with Crank Nicolson: the details

We now give a detailed discussion of Crank Nicolson and when it breaks down or fails to live up to its perceived expectations.

3.1 A critique of Crank Nicolson

The Crank Nicolson method has become a very popular finite difference scheme for approximating the Black–Scholes equation.

This equation is an example of a *convection–diffusion* equation and it has been known for some time that centred-difference schemes are inappropriate for approximating it (Il'in 1969, Duffy 1980). In fact, many independent discoveries of novel methods have been made in order to solve difficult convection–diffusion problems in fluid dynamics, atmospheric pollution modelling, semiconductor equations, the Fokker–Planck equation and groundwater transport (Morton 1996).

The main problem is that traditional finite difference schemes start to oscillate when the coefficient of the second derivative (the *diffusion* term) is very small or when the coefficient of the first derivative (the *convection* term) is large (or both). In this case, the mesh size h in the space direction must be smaller than a certain critical value if we wish to avoid these oscillations. This problem has been known since the 1950s (see de Allen 1955).

We now discuss Crank Nicolson from a number of viewpoints. For convenience and generality reasons, we cast the Black–Scholes equation as a generic parabolic initial boundary value problem in the domain $D = (A, B)X(0, T)$ where $A < B$:

$$\begin{aligned}
&Lu \equiv -\frac{\partial u}{\partial t} + \sigma(x,t)\frac{\partial^2 u}{\partial x^2} + \mu(x,t)\frac{\partial u}{\partial x} + b(x,t)u = f(x,t) \text{ in } D\\
&u(x,0) = \varphi(x),\ x \in (A,B)\\
&u(A,t) = g_0(t),\ u(B,t) = g_1(t),\ t \in (0,T).
\end{aligned} \tag{6}$$

In this case the time variable t corresponds to increasing time while the space variable x corresponds to the underlying asset price S. We specify Dirichlet boundary conditions on a finite space interval and this is a common situation for several kinds of exotic options, for example barrier options. Actually, the system (6) is more general than the original Black–Scholes equation.

3.2 How are derivatives approximated?

There are two kinds of independent variables associated with the one-factor Black–Scholes as can be seen in (6). These correspond to the x and t variables. We concentrate on the x direction for the moment. We discretize in this direction using centred differences at the point (jh, nk):

$$\begin{cases}
\dfrac{\partial^2 u}{\partial x^2} \sim \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}\\[2ex]
\dfrac{\partial u}{\partial x} \sim \dfrac{u_{j+1}^n - u_{j-1}^n}{2h}.
\end{cases}$$

Using this knowledge we can apply the Crank Nicolson method to (6), namely:

$$\left\{\begin{aligned}
&-\frac{u_j^{n+1}-u_j^n}{k}+\sigma_j^{n+\frac{1}{2}}\frac{u_{j+1}^{n+\frac{1}{2}}-2u_j^{n+\frac{1}{2}}+u_{j-1}^{n+\frac{1}{2}}}{h^2}\\
&+\mu_j^{n+\frac{1}{2}}\frac{u_{j+1}^{n+\frac{1}{2}}-u_{j-1}^{n+\frac{1}{2}}}{2h}\\
&+b_j^{n+\frac{1}{2}}u_j^{n+\frac{1}{2}}=f_j^{n+\frac{1}{2}}.
\end{aligned}\right. \tag{7}$$

A bit of simple arithmetic allows us to rewrite (7) in the standard form:

$$\left\{\begin{aligned}
&a_j^n u_{j-1}^{n+1}+b_j^n u_j^{n+1}+c_j^n u_{j+1}^{n+1}=F_j^n\\
&F_j^n \text{ known quantity.}
\end{aligned}\right. \tag{8}$$

Of course, this system of equations can be posed in the form of a matrix system. A number of researchers have examined such systems in conjunction with convection–diffusion equations (for example, Farrell 2000, Morton 1996). A critical observation is that if the coefficient a_j^n is not positive then the resulting solution will show oscillatory behaviour at best or produce non-physical solutions at worst.

This will give problems in general for Black–Scholes applications where the volatility is a decaying function of time (see van Deventer 1997), for example:

$$\sigma(t)=\sigma_0 e^{-\alpha(T-t)}$$
where σ_0 and α are given constants.

We speak of a *singular perturbation* problem associated with problem (6) when the coefficient of the second derivative is small (see Duffy 1980). In this case traditional finite difference schemes perform badly at the *boundary layer* situated at $x=0$. In fact, if we formally set volatility to zero in equation (7) we get a so-called *weakly stable* difference scheme (see Peaceman 1977) that approximates the *first-order hyperbolic equation*

$$-\frac{\partial u}{\partial t}+\mu\frac{\partial u}{\partial x}+bu=f.$$

This has the consequence that the initial errors in the scheme are not dissipated and hence we can expect oscillations especially in the presence of rounding errors. We need other *one-sided schemes* in this degenerate case (Peaceman 1977, Duffy 1977).

3.3 Boundary conditions

In general, we distinguish three kinds of boundary conditions:

- Dirichlet (as seen in the system (6))
- Neumann conditions
- Robin conditions

The last two boundary conditions involve the first derivative of the unknown u at the boundaries. We must then decide on how we are going to approximate this derivative. We can choose between first-order accurate one-sided schemes and *ghost points* (Thomas 1998) that produce a second-order approximation to the first derivative. We must thus be aware of the fact that the low-order accuracy at the boundary will adversely impact the second-order accuracy in the interior of the region of interest. To complicate matters, some models have a boundary condition involving the second derivative of u or even a 'linearity' boundary condition (see Tavella 2000).

Finally, the boundary conditions may be discontinuous. We may resort to non-uniform meshes to accommodate the discontinuities. This strategy will also destroy the second-order accuracy of the Crank Nicolson method. The conclusion is that the wrong discrete boundary conditions adversely affect the accuracy of the finite difference scheme.

3.4 Initial conditions

It is well known that discontinuous initial conditions adversely impact the accuracy of finite difference schemes (see Smith 1978). In particular, the solution of the difference schemes exhibits oscillations just after $t = 0$ but the solution becomes more smooth as time goes on. This has consequences for options pricing applications because in general the initial condition (this is in fact a payoff function) is not always smooth. For example, the payoff function for a European call option is:

$$C = \max(S - K, 0)$$

where K is the strike price and S is the stock price. Its derivative is given by the jump function:

$$\frac{\partial C}{\partial S} = \begin{cases} 0, & S \leq K \\ 1, & S > K. \end{cases}$$

This derivative is discontinuous and in general we can expect to get bad accuracy at the points of discontinuity (in this case, at the strike price where at-the-money issues play an important role). It is possible to determine mathematically what the accuracy is in some special cases (Smith 1978) but numerical experiments show us that things are going wrong as well. Of course, if the option price is badly approximated there is not much hope of getting good approximations to the delta and gamma. This statement is borne out in practice. Another source of annoyance is that the boundary and initial conditions may not be *compatible* with each other. By compatibility, we mean that the solution is smooth at the corners $(A, 0)$ and $(B, 0)$ of the region of interest and we thus demand that the solution is the same irrespective of the direction from which we approach the corners. If we assume that $u(x, t)$ is continuous as we approach the boundaries, then we must satisfy the *compatibility conditions*:

$$\begin{cases} \varphi(A) \equiv u(A, 0) = g_0(0) \\ \varphi(B) \equiv u(B, 0) = g_1(0). \end{cases}$$

Failure to take these conditions into account in a finite difference scheme will lead to inaccuracies at the corner points of the region of interest. On the upside, the discontinuities are quickly damped out.

3.5 Proving stability

Much of the literature uses the von Neumann theory to prove stability of finite difference schemes (Tavella 2000). This theory was developed by John von Neumann, a Hungarian–American mathematician, the father of the modern computer and probably one of the greatest brains of the twentieth century. Strictly speaking, the von Neumann approach is only valid for constant coefficient, linear initial value problems. The Black–Scholes equation does not fall under this category. Furthermore, much work has been done in the engineering field to prove stability in other ways, for example using the maximum principle and matrix theory (Morton 1996, Duffy 1980). A discussion of von Neumann stability for the constant coefficient, linear convection–diffusion equation can be found in Thomas (1998).

4 An introduction to exponentially fitted finite difference schemes

4.1 A new class of robust difference schemes

Exponentially fitted schemes are stable, have good convergence properties and do not produce spurious oscillations. In order to motivate what an exponentially fitted difference scheme is, let us look at the simple boundary value problem:

$$\begin{aligned}
&\sigma \frac{d^2u}{dx^2} + \mu \frac{du}{dx} = 0 \qquad \text{in} \qquad (A, B) \\
&u(A) = \beta_0, \quad u(B) = \beta_1.
\end{aligned} \tag{9}$$

Here we assume that σ and μ are positive constants. We now approximate (9) by the difference scheme defined as follows:

$$\begin{aligned}
&\sigma\rho D_+ D_- U_j + \mu D_0 U_j = 0, \quad j = 1, \ldots, J-1 \\
&U_0 = \beta_0, \quad U_J = \beta_1
\end{aligned} \tag{10}$$

where ρ is a so-called fitting factor (this factor is identically equal to 1 in the case of the centred difference scheme). We now choose ρ so that the solutions of (9) and (10) are identical at the mesh-points. Some easy arithmetic shows that

$$\rho = \frac{\mu h}{2\sigma} \coth \frac{\mu h}{2\sigma}$$

where $\coth x$ is the hyperbolic cotangent function defined by

$$\coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{e^{2x} + 1}{e^{2x} - 1}.$$

The fitting factor ρ will be used when developing fitted difference schemes for variable coefficient problems. In particular, we discuss the following problem:

$$\begin{aligned}&\sigma(x)\frac{d^2u}{dx^2}+\mu(x)\frac{du}{dx}+b(x)u=f(x)\\&u(A)=\beta_0,\quad u(B)=\beta_1\end{aligned}\tag{11}$$

where σ, μ and b are given continuous functions, and

$$\sigma(x)\geq 0,\quad \mu(x)\geq\alpha>0,\quad b(x)\leq 0\ \text{ for } x\in(A,B).$$

The fitted difference scheme that approximates (11) is defined by:

$$\begin{aligned}&\rho_j^h D_+D_-U_j+\mu_j D_0U_j+b_jU_j=f_j,\quad j=1,\ldots,J-1\\&U_0=\beta_0,\quad U_J=\beta_1\end{aligned}\tag{12}$$

where

$$\begin{aligned}&\rho_j^h=\frac{\mu_j h}{2}\coth\frac{\mu_j h}{2\sigma_j}\\&\sigma_j=\sigma(x_j),\quad \mu_j=\mu(x_j),\quad b_j=b(x_j),\ f_j=f(x_j)\end{aligned}\tag{13}$$

We now state the following fundamental results (see Il'in 1969, Duffy 1980).

The solution of scheme (12) is *uniformly stable*, that is

$$|U_j|\leq|\beta_0|+|\beta_1|+\frac{1}{\alpha}\max_{k=1,\ldots,J}|f_k|,\quad j=1,\ldots,J-1$$

Furthermore, scheme (12) is monotone in the sense that the matrix representation of (12)

$$AU=F$$

where $U={}^t(U_1,\ldots,U_{J-1})$, $F={}^t(f_1,\ldots,f_{J-1})$ and

$$\mathbf{A}=\begin{pmatrix}\ddots & & \ddots & & 0\\ & \ddots & & a_{j,j+1} & \\ \ddots & & a_{j,j} & & \ddots\\ & a_{j,j-1} & & \ddots & \\ 0 & & \ddots & & \ddots\end{pmatrix}\tag{14}$$

$$\begin{aligned}a_{j,j-1}&=\frac{\rho_j^h}{h^2}-\frac{\mu_j}{2h}>0 &&\text{always}\\a_{j,j}&=-\frac{2\rho_j^h}{h^2}+b_j<0 &&\text{always}\\a_{j,j+1}&=\frac{\rho_{jh}}{h^2}+\frac{\mu_j}{2h}>0 &&\text{always}\end{aligned}$$

produces positive solutions from positive input.

Sufficient conditions for a difference scheme to be monotone have been investigated by many authors in the last 30 years; we mention the work of Samarski (1976) and Stoyan (1979).

Stoyan also produced stable and convergent difference schemes for the convection–diffusion equation producing results and conclusions that are similar to the author's work (see Duffy 1980).

Let u and U be the solutions of (11) and (12), respectively. Then

$$|u(x_j) - U_j| \leq Mh$$

where M is a positive constant that is independent of h and σ (Il'in 1969).

The conclusion is that the fitted scheme (12) is stable, convergent and produces no oscillations. In particular, the scheme 'degrades gracefully' to a well-known stable scheme when σ tends to zero.

5 Exponentially fitted schemes for the Black–Scholes equation

We discretize the rectangle $[A, B] \times [0, T]$ as follows:

$$A = x_0 < x_1 < \ldots < x_J = B \quad (h = x_j - x_{j-1}), h \text{ constant}$$

$$0 = t_0 < t_1 < \ldots < t_N = T \quad (k = T/N), k \text{ constant.}$$

Consider again the operator L in equation (6) defined by

$$Lu \equiv -\frac{\partial u}{\partial t} + \sigma(x,t)\frac{\partial^2 u}{\partial x^2} + \mu(x,t)\frac{\partial u}{\partial x} + b(x,t)u.$$

We replace the derivatives in this operator by their corresponding divided differences and we define the fitted operator L_k^h by

$$L_k^h U_j^n \equiv -\frac{U_j^{n+1} - U_j^n}{k} + \rho_j^{n+1} D_+ D_- U_j^{n+1} + \mu_j^{n+1} D_0 U_j^{n+1} + b_j^{n+1} U_j^{n+1}. \tag{15}$$

Here we use the notation

$$\varphi_j^{n+1} = \varphi(x_j, t_{n+1}) \text{ in general}$$

and

$$\rho_j^{n+1} \equiv \frac{\mu_j^{n+1} h}{2} \coth \frac{\mu_j^{n+1} h}{2\sigma_j^{n+1}}.$$

We now formulate the fully discrete scheme that approximates the initial boundary value problem (6).

Find a discrete function $\{U_j^n\}$ such that

$$L_k^h U_j^n = f_j^{n+1}, \; j = 1, \ldots, J-1, \; n = 0, \ldots, N-1$$

$$U_0^n = g_0(t_n),\ U_J^n = g_1(t_n),\ n = 0, \ldots, N \tag{16}$$
$$U_j^0 = \varphi(x_j),\ j = 1, \ldots, J-1.$$

This is a two-level implicit scheme. We wish to prove that scheme (16) is stable and is consistent with the initial boundary value problem (6). We prove stability of (16) by the so-called discrete maximum principle instead of the von Neumann stability analysis. The von Neumann approach is well known but the discrete maximum principle is more general and easier to understand and apply in practice. It is also the de-facto standard technique for proving stability of finite difference and finite element schemes (see Morton 1996, Farrell 2000).

Lemma 1 *Let the discrete function* w_j^n *satisfy* $L_k^h w_j^n \leq 0$ *in the interior of the mesh with* $w_j^n \geq 0$ *on the boundary* Γ.
Then $w_j^n \geq 0, \forall j = 0, \ldots, J;\ n = 0, \ldots, N$.

Proof We transform the inequality $L_k^h w_j^n \leq 0$ into an equivalent vector inequality. To this end, define the vector $W^n = {}^t(w_1^n, \ldots, w_{J-1}^n)$. Then the inequality $L_k^h w_j^n \leq 0$ is equivalent to the vector inequality

$$A^n W^{n+1} \geq W^n \tag{17}$$

where

$$A^n = \begin{pmatrix} \ddots & & \ddots & & 0 \\ & \ddots & & t_j^n & \\ \ddots & & s_j^n & & \ddots \\ & r_j^n & & \ddots & \\ 0 & & \ddots & & \ddots \end{pmatrix}$$

$$r_j^n = \left(-\frac{\rho_j^n}{h^2} + \frac{\mu_j^n}{2h}\right)k$$

$$s_j^n = \left(\frac{2\rho_j^n}{h^2} - b_j^n + k^{-1}\right)k$$

$$t_j^n = \left(-\left(\frac{\rho_j^n}{h^2} + \frac{\mu_j^n}{2h}\right)\right)k.$$

It is easy to show that the matrix A^n has non-positive off-diagonal elements, has strictly positive diagonal elements and is irreducibly diagonally dominant. Hence (see Varga 1962, pages 84–85) A^n is non-singular and its inverse is positive:

$$(A^n)^{-1} \geq 0$$

Using this result in (17) gives the desired result.

Lemma 2 *Let $\{U_j^n\}$ be the solution of scheme (16) and suppose that*

$$max|U_j^n| \leq m \ on \ \Gamma \ \ for \ all \ j \ and \ n$$
$$max|f_j^n| \leq N \ in \ D \ \ for \ all \ j \ and \ n$$

Then

$$max_j|U_j^n| \leq -\frac{N}{\beta} + m \ \ in \ \overline{D}$$

Proof Define the discrete barrier function

$$w_j^n = -\frac{N}{\beta} + m \pm U_j^n$$

Then $w_j^n \geq 0$ on Γ. Furthermore,

$$L_k^h w_j^n \leq 0$$

Hence $w_j^n \geq 0$ in $\overline{Q}$ which proves the result.

Let $u(x, t)$ and $\{U_j^n\}$ be the solutions of (6) and (16), respectively.
Then

$$|u(x_j, t_n) - U_j^n| \leq M(h + k) \tag{18}$$

where M is a constant that is independent of h, k and σ.

This result shows that convergence is assured regardless of the size of σ. No classical scheme (for example, centred differencing in x and Crank Nicolson in time) has error bounds of the form (18) where M is independent of h, k and σ.

Summarizing, the advantages of the fitted scheme are:

- It is uniformly stable for all values of h, k and σ.
- It is oscillation-free. Its solution converges to the exact solution of (6). In particular, it is a powerful scheme for the Black–Scholes equation and its generalizations.
- It is easily programmed, especially if we use object-oriented design and implementation techniques.

6 Problems with small volatility

We now examine some 'extreme' cases in system (16). In particular, we examine the cases

(pure convection/drift) $\sigma \to 0$
(pure diffusion/volatility) $\mu \to 0$

We shall see that the 'limiting' difference schemes are well-known schemes and this is reassuring. To examine the first extreme case we must know what the limiting properties of the hyperbolic cotangent function are:

$$\lim_{\sigma\to 0} \rho_j^n = \lim_{\sigma\to 0} \frac{\mu_j^n h}{2} \coth \frac{\mu_j^n h}{2\sigma_j^n}.$$

We use the formula

$$\lim_{\sigma\to 0} \frac{\mu h}{2} \coth \frac{\mu h}{2\sigma} = \begin{cases} +\dfrac{\mu h}{2} \; if \; \mu > 0 \\ -\dfrac{\mu h}{2} \; if \; \mu < 0. \end{cases}$$

Inserting this result into the first equation in (16) gives us the first-order scheme

$$\mu > 0, \quad -\frac{U_j^{n+1} - U_j^n}{k} + \mu_j^{n+1} \frac{(U_{j+1}^{n+1} - U_j^{n+1})}{h} + b_j^{n+1} U_j^{n+1} = f_j^{n+1}$$
$$\mu < 0, \quad -\frac{U_j^{n+1} - U_j^n}{k} + \mu_j^{n+1} \frac{(U_j^{n+1} - U_{j-1}^{n+1})}{h} + b_j^{n+1} U_j^{n+1} = f_j^{n+1}.$$

These are so-called implicit upwind schemes and are stable and convergent (Duffy 1977, Dautray 1993). We thus conclude that the fitted scheme degrades to an acceptable scheme in the limit. The case $\mu \to 0$ uses the formula

$$\lim_{x\to 0} x \coth x = 1.$$

Then the first equation in system (16) reduces to the equation

$$-\frac{U_j^{n+1} - U_j^n}{k} + \sigma_j^{n+1} D_+ D_- U_j^{n+1} + b_j^{n+1} U_j^{n+1} = f_j^{n+1}.$$

This is a standard approximation to pure diffusion problems and such schemes can be found in standard numerical analysis textbooks.

These limiting cases reassure us that the fitted method behaves well for 'extreme' parameter values.

7 Exponential fitting and exotic options

We have applied the method to a range of plain and exotic European and American type options. In particular, we have applied it to various kinds of barrier options (see Topper 1998, Haug 1998), for example:

- Double barrier call options
- Single barrier call options
- Equations with time-dependent volatilities (for example, a linear function of time)

- Asymmetric plain vanilla power call options
- Asymmetric capped power call options

We have compared our results with those in Haug (1998) and Topper (1998) and they compare favourably (Mirani 2002). The main difference between these types lies in the specific payoff functions (initial conditions) and boundary conditions. Since we are working with a specific kind of parabolic problem these functions must be specified by us. For example, for a double barrier option we must give the value of the option at these barriers while for a single barrier option we define the 'down' barrier at $S = 0$. Summarizing, the exponentially fitted finite difference scheme gives good approximations to the option price and delta of the above exotic option types. We have compared the results with Monte Carlo, Haug (1998) and Topper (1998).

8 Uniform approximation of the Greeks

It is well known by now that CN produces bad approximation to option delta and gamma (see, for example, Zvan 1997, Cooney 1999). Thus, we need to devise schemes that do give uniform approximation to option sensitivities, especially in the vicinity of the strike price K. The exponentially fitted scheme (16) is a good candidate and more information can be found in Duffy (2001) and Cooney (1999).

8.1 Is there more hope? The Keller scheme

In this section, however, we give a short overview of the box scheme (Keller 1971) that resolves many of the problems associated with Crank Nicolson. In short, we reduce the second-order Black–Scholes equation to a system of first-order equations containing at most first-order derivatives. We then approximate the first derivatives in x and t by averaging in a box. We motivate the box scheme by examining the generic parabolic initial boundary value problem in the space interval (0, 1):

$$\begin{aligned}
&\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right) + cu + S, 0 < x < 1,\ t > 0\\
&u(x,0) = g(x),\ 0 < x < 1\\
&\alpha_0 u(0,t) + \alpha_1 a(0,t)u_x(0,t) = g_0(t)\\
&\beta_0 u(1,t) + \beta_1 a(1,t)u_x(1,t) = g_1(t).
\end{aligned} \tag{19}$$

Here u is the (unknown) solution to the problem that satisfies the *self-adjoint equation* in (19) and it must also satisfy the initial and boundary conditions (note the latter contain derivatives of the unknown at the boundaries of the interval). In general, the coefficients in (19) are functions of both x and t.

We now transform (19) to a first-order system by defining a new variable v. The new transformed set of equations is given by:

$$\begin{aligned}
&a\frac{\partial u}{\partial x} = v\\
&\frac{\partial v}{\partial x} = \frac{\partial u}{\partial t} - cu - S
\end{aligned}$$

$$\begin{aligned} &u(x,0) = g(x) \\ &\alpha_0 u(0,t) + \alpha_1 v(0,t) = g_0(t) \\ &\beta_0 u(1,t) + \beta_1 v(1,t) = g_1(t). \end{aligned} \tag{20}$$

We now see that we have to deal with a first-order system of equations with no derivatives on the boundaries!

We now need to introduce some notation. First, we define average values for x and t coordinates as follows:

$$\begin{aligned} x_{j\pm 1/2} &= \frac{1}{2}(x_j + x_{j\pm 1}) \\ t_{n\pm 1/2} &= \frac{1}{2}(t_n + t_{n\pm 1}) \end{aligned}$$

and for general nets (in principle the approximations to u and v) by

$$\begin{aligned} \phi^n_{j\pm 1/2} &= \frac{1}{2}(\phi^n_j + \phi^n_{j\pm 1}) \\ \phi^{n\pm 1/2}_j &= \frac{1}{2}(\phi^n_j + \phi^{n\pm 1}_j). \end{aligned}$$

Finally, we define notation for divided differences in the x and t directions as follows:

$$\begin{aligned} D_x^- \phi^n_j &= h_j^{-1}(\phi^n_j - \phi^n_{j-1}) \\ D_t^- \phi^n_j &= k_n^{-1}(\phi^n_j - \phi^{n-1}_j). \end{aligned}$$

We are now ready for the new scheme. To this end, we use one-sided difference schemes in both directions while taking averages and we thus solve for both u and v simultaneously at each time level:

$$\begin{aligned} &a^n_{j-1/2}\, D_x^- u^n_j = v^n_{j-1/2} \\ &D_x^- v^{n-1/2}_j = D_t^- u^n_{j-1/2} - c^{n-1/2}_{j-1/2} u^{n-1/2}_{j-1/2} - S^{n-1/2}_{j-1/2} \\ &1 \le j \le J,\ 1 \le n \le N. \end{aligned} \tag{21}$$

The corresponding boundary and initial conditions are:

$$\left.\begin{aligned} \alpha_0 u^n_0 + \alpha_1 v^n_0 &= g^n_0 \\ \beta_0 u^n_J + \beta_1 v^n_J &= g^n_1 \end{aligned}\right\} 1 \le n \le N. \tag{22}$$

The box scheme has a number of very desirable properties, namely: (a) It is simple, efficient and easy to program, (b) It is unconditionally stable, (c) It approximates u and its partial derivative in x with second-order accuracy. For the Black–Scholes equation this means that we can approximate both option price and the option delta without trace of spurious oscillation as is experienced with Crank Nicolson, (d) Richardson extrapolation is applicable and yields two orders of accuracy improvement per extrapolation (with non-uniform nets!), (e) It supports data, coefficients and solutions that are only piecewise smooth. In a financial setting it is able to model piecewise smooth payoff functions. We then define the approximate initial condition as follows:

$$v^0_{j-\frac{1}{2}} = a^0_{j-\frac{1}{2}} \frac{dg\left(x_{j-1/2}\right)}{dx}, 1 \le j \le J. \tag{23}$$

For piecewise smooth boundary conditions we use the following tactic:

$$\begin{aligned} &\alpha_0 u_0^{n-\frac{1}{2}} + \alpha_1 v_0^{n-\frac{1}{2}} = g_0^{n-\frac{1}{2}} \\ &\beta_0 u_J^{n-\frac{1}{2}} + \beta_1 v_J^{n-\frac{1}{2}} = g_1^{n-\frac{1}{2}} \\ &1 \le n \le N \\ &\text{Discontinuities at } t = t_n! \end{aligned} \tag{24}$$

Of course we are assuming that the mesh points are sitting on the discontinuities!

9 Conclusions

We have discussed the popular Crank Nicolson method from a number of viewpoints. In particular, we have made an inventory of the situations where it breaks down or where it deviates from our expectations:

- The standard von Neumann stability analysis fails to predict the infamous spurious oscillation problem. Hedging applications that use CN will run the risk of inaccuracy at values in the payoff function where this function is not smooth (for example, the strike price).
- Second-order accuracy is lost when using non-uniform meshes. Sometimes uniform meshes are not sufficient to approximate the exact solution in a boundary layer (small volatility) or with nasty payoff functions (for example, binary options or barrier options with discrete and intermittent barriers). A good discussion of how Crank Nicolson breaks down for barrier options is given in Tavella (2000).
- There are finite difference schemes that are just as good as, or even better than, Crank Nicolson, for example fully implicit schemes with extrapolation or Runge–Kutta (Crouzeix 1975).
- For two-factor and multi-factor problems, we use predictor–corrector, Alternating Direction Implicit (ADI) and Operator Splitting methods (see Peaceman 1977, Janenko 1971, Sun 1999). In these cases we see that Crank Nicolson is just one possibility for time discretization.

A modest proposal would be to investigate robust and effective alternatives to the Crank Nicolson schemes. This will hopefully improve the *FDM gene pool* as it were.

Acknowledgements

Permission to use some of the text in this chapter has been given by John Wiley publishers and is based on Daniel Duffy's book *Designing and Implementing Financial Instrument Pricing using C++* (ISBN: 0-470-85509-6).

The author

Daniel Duffy works for Datasim (www.datasim.nl), an Amsterdam-based trainer and software developer. He has been working in IT since 1979 and with object-oriented technology since 1987. He received his M.Sc. and Ph.D. theses (in numerical analysis) from Trinity College, Dublin. At present he is working on Finite Differences and C++ for instrument pricing problems.

REFERENCES

■ Aho, A., Kernighan, B. and Weinberger, P. (1988) *The AWK Programming Language*. Addison-Wesley.

■ de Allen, D. and Southwell R. (1955) Relaxation methods applied to determining the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. Appl. Math.*, 129–145.

■ Bhansali, V. (1998) *Pricing and Managing Exotic and Hybrid Options*. McGraw-Hill Irwin Library Series, New York.

■ Cooney, M. (1999) Benchmarking the Black Scholes Equation by Finite Differences. MSc thesis, Trinity College Dublin.

■ Crank, J. and Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Cambridge Philos. Soc.*, 43, 50–67, re-published in: John Crank 80th birthday special issue *Adv. Comput. Math*. 6 (1997) 207-226.

■ Crouzeix, M. (1975) On the approximation of linear operational differential equations by the Runge–Kutta method. Doctorate Thesis, Paris VI University, France.

■ Dautray, R. and Lions, J. L. (1993) *Mathematical Analysis and Numerical Methods for Science and Technology*, Volume 6. Springer-Verlag, Berlin.

■ van Deventer, D. R. and Imai, K. (1997) *Financial Risk Analytics*. McGraw-Hill, Chicago.

■ Doolan, E. P., Miller, J. J. H. and Schilders, W. H. A. (1980) *Uniform Numerical Methods for Problems with Initial and Boundary Layers*. Boole Press, Dublin, Ireland.

■ Douglas, J. and Rachford, H. H. (1955) On the numerical solution of heat conduction equations in two and three dimensions. *Trans. Am. Math.*, 82, 421–439.

■ Duffy, D. (1977) Finite elements for mixed initial boundary value problems for hyperbolic systems of equations. M.Sc. Thesis, Trinity College Dublin, Ireland.

■ Duffy, D. (1980) *Uniformly convergent difference schemes for problems with a small parameter in the leading derivative*. Ph.D. Thesis, Trinity College, Dublin.

■ Duffy, D. (2001) Robust and accurate finite difference methods in option pricing: one factor models. Working report, Datasim www.datasim-component.com

■ Duffy, D. (2004) *Designing and Implementing Financial Instrument Pricing using C++*. John Wiley & Sons, Chichester.

■ Farrell, P., Hegarty, A. F., Miller, J. J. H., O'Riordan, E. and Shishkin, G. I. (2000) *Robust Computational Techniques for Boundary Layers*. Chapman and Hall/CC, Boca Raton.
■ Godunov, S. and Riabenki, V. S. (1987) *Difference Schemes, An Introduction to the Underlying Theory*. North-Holland, Amsterdam.
■ Haug, E. (1998) *The Complete Guide to Option Pricing Formulas*. McGraw-Hill, New York.
■ Il'in, A. M. (1969) Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Mat. Zametki*, 6, 237–248.
■ Isaacson, E. and Keller, H. (1966) *Analysis of Numerical Methods*. John Wiley & Sons, New York.
■ Janenko, N. N. (1971) *The Method of Fractional Steps*. Springer-Verlag, Berlin.
■ Keller, H. (1968) *Numerical Methods for Boundary-Value Problems*. Blaisdell Publishing Company, Waltham, MA.
■ Keller, H. (1971) A new difference scheme for parabolic problems. In B. Hubbard (ed.); *Numerical Solution of Partial Differential Equations-II*, Synspade.
■ Lam, D. C. L. and Simpson, R. B. (1976) Centered differencing and the box scheme for diffusion convection problems. *Journal of Computational Physics*, 22, 486–500.
■ Levin, A. (2000) Two-factor Gaussian term structure: analytics, historical fit and stable finite-difference pricing schemes. Lecture held at Courant Institute, New York University, May.
■ Mirani, R. (2002) Application of Duffy's finite difference method to barrier options. Working Paper, Datasim Education BV, Amsterdam.
■ Morton, K. (1996) *Numerical Solution of Convection-Diffusion Equations*. Chapman and Hall, London, UK.
■ Peaceman, D. (1977) *Numerical Reservoir Simulation*. Elsevier.
■ Richtmyer, R. D. and Morton, K. W. (1967) *Difference Methods for Initial-Value Problems*. Interscience Publishers (John Wiley), New York. Reprint edition (1994) Krieger Publishing Company, Malabar.
■ Press, W., Flannery, B., Teukolsky, S. and Vetterling, W. (1989) *Numerical Recipes*. Cambridge University Press.
■ Roscoe, D. F. (1975) New methods for the derivation of stable difference representations. *J. Inst. Math. Appl.*, 16, 291–301.
■ Samarski, A. A. (1976) Some questions from the general theory of difference schemes. *Amer. Math. Soc. Trans.*, 2, Vol. 105.
■ Seydel, R. (2003) *Tools for Computational Finance*. Springer, Berlin.
■ Smith, G. D. (1978) *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Oxford University Press.
■ Stoyan, G. (1979) Monotone difference schemes for diffusion-convection problems. *ZAMM*, 59, 361–372.
■ Strang, G. and Fix, G. (1973) *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, NJ.
■ Sun, Y. (1999) High order methods for evaluating convertible bonds. Ph.D. thesis, University of North Carolina.
■ Tavella, D. and Randall, C. (2000) *Pricing Financial Instruments: The Finite Difference Method*. John Wiley & Sons, New York.
■ Thomas, J. W. (1998) *Numerical Partial Differential Equations, Volume I Finite Difference Methods*. Springer, New York.

■ Topper, J. (1998) Finite element modeling of exotic options. Internal Report, University of Hannover, ISSN 0949-9962.
■ van Deventer, D. R. and Imai K. (1997) *Financial Risk Analytics*. McGraw-Hill, Chicago.
■ Varga, R. (1962) *Matrix Iterative Analysis*. Prentice Hall Inc., Englewood Cliffs, NJ.
■ Wilmott, P., Dewynne, J. and Howison, S. (1993) *Option Pricing*. Oxford Financial Press, Oxford, UK.
■ Wilmott, P. (1998) *Derivatives*. John Wiley & Sons, Chichester.
■ Zvan, R., Forsyth, P. A. and Vetzal, K. R. (1997) Robust numerical methods for PDE models of Asian options. *J. Comp. Finance*, Vol. 1/Nu. 2, Winter 1997/1998.

27
Finite Elements and Streamline Diffusion for the Pricing of Structured Financial Instruments

Andreas Binder and Andrea Schatz

The numerical treatment of partial differential equations in computational finance started with binomial and trinomial trees, with all the drawbacks related to these approaches. In the meanwhile (see, e.g., Duffy 2004, in the July issue of *Wilmott*), finite differences are widely used in modern derivatives pricing. We present how pricing software can be developed on the basis of finite element techniques, which allow more flexibility than finite differences.

Mean reverting models for interest rates tend to become numerically difficult in regions sufficiently far away from the mean-reverting level. The reason is that the convection dominates the diffusion in these regions, and therefore techniques for convection-dominated flows should be applied. We present how streamline diffusion is applied to obtain stable numerical schemes.

We implemented these approaches in a strictly object-oriented software framework. Some software engineering aspects are also highlighted.

Introduction

We consider models for financial instruments which can, after some manipulation, be written in the form of parabolic partial differential equations backwards in time. The manipulation typically requires some Itô calculus, the creation of a risk-free portfolio and self-financing hedging strategies and some assumptions (like zero transaction costs), which are certainly wrong but a

E-mail: binder@mathconsult.co.at, schatz@mathconsult.co.at

good starting point. LIBOR market models typically do not fall into this category, but short rate models do.

For example, let us start with a two-factor Hull–White interest rate model (see Hull and White 1994)

$$dr = [\theta(t) + u(t) - a(t)r(t)]dt + \sigma_1(t)dX_1$$
$$du = -b(t)u(t)dt + \sigma_2(t)dX_2.$$

The first factor r denotes the spot rate, the second factor u some kind of long-term development of the interest rates. a is the mean reversion speed of the spot rate r, $(\theta + u)/a$ its reversion level. The stochastic variable u itself reverts to a level of zero at rate b. dX_1 and dX_2 are increments of Wiener processes with instantaneous correlation $\rho(t)$. σ_1 and σ_2 are the volatilities.

No-arbitrage arguments then lead to the fundamental Hull–White equation

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma_1(t)^2\frac{\partial^2 V}{\partial r^2} + \rho(t)\sigma_1(t)\sigma_2(t)\frac{\partial^2 V}{\partial r\partial u} + \frac{1}{2}\sigma_2(t)^2\frac{\partial^2 V}{\partial u^2}$$
$$+ (\theta(t) + u - a(t)r)\frac{\partial V}{\partial r} - b(t)u\frac{\partial V}{\partial u} - rV = 0,$$

which needs additional end and transition conditions. The calculation domain is, in principle, unbounded. We will discuss the problem of boundary conditions, when restricting ourselves to a bounded calculation domain, below.

The end and transition conditions describe the special shape of a financial contract, like coupons, callabilities and so on.

The given partial differential equation can be interpreted as a diffusion–convection–reaction equation. This type of equation is typically found in applications in continuum mechanics, especially in fluid mechanics. The dissolving of sugar in a cup of coffee, for example, could be described by this type of equation. The dispersion of the sugar due to concentration differences is a diffusion process, described by the second-order terms in the equation. The spreading of the sugar driven by a stirring spoon is the convective part, given by the first-order terms, and strongly dominated by the velocity of the coffee. The dissolving of the sugar itself is described by the reactive part, which is the last term at the left-hand side of the equation. Figure 1, which shows

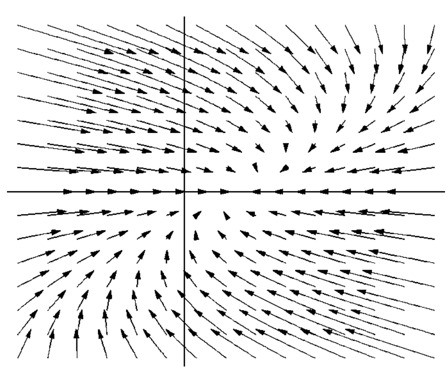


Figure 1: Velocity field in a Hull-White model

velocity vectors in the *ru*-plane, could give the motion of the coffee forced by the stirring spoon, but, in fact, it gives the deterministic movement of the interest rates in a two-factor Hull–White interest rate model.

The figure demonstrates that in the two-factor Hull–White model the convective part become more and more important the larger the considered domain.

It is obvious now that numerical methods used in computational fluid dynamics to solve equations of this type will work well also for our pricing problems. In computational fluid dynamics it is well known that in the cases of comparatively large or dominating convection standard numerical discretisation techniques lead to instabilities in the numerical solution. These instabilities result in high oscillations. We have to use so-called upwind-strategies, which take into account that in the case of dominating convection the solution in each point is strongly determined by the information transported with the velocity.

Since in the considered pricing problems end conditions for the quantity V are prescribed, we have to solve the equation backwards in time. So the information transport due to convection starts in the centre and goes to the boundary.

Numerical schemes and finite elements

Finite volume method

The basic idea of the finite difference method is to approximate the derivatives in the partial differential equation by finite differences. In the case of higher dimensions, especially including mixed derivatives, a more general formulation is preferred and known under the name finite volume method. The essential idea is to use an integral formulation, integrating the equation over a mesh region and applying the divergence theorem before carrying out the discretisation.

We start already from the time discretised equation, either fully implicit

$$\begin{aligned}&\frac{V^{n+1}-V^n}{\Delta t}+\frac{1}{2}\sigma_1^{n+1^2}\frac{\partial^2 V^{n+1}}{\partial r^2}+\rho^{n+1}\sigma_1^{n+1}\sigma_2^{n+1}\frac{\partial^2 V^{n+1}}{\partial r\partial u}+\frac{1}{2}\sigma_1^{n+1^2}\frac{\partial^2 V^{n+1}}{\partial u^2}\\&\quad+(\theta^{n+1}+u-a^{n+1}r)\frac{\partial V^{n+1}}{\partial r}-b^{n+1}u\frac{\partial V^{n+1}}{\partial u}-rV^{n+1}=0,\end{aligned}$$

or, e.g., of Crank Nicolson type ($\alpha = 0.5$)

$$\begin{aligned}&\frac{V^{n+1}-V^n}{\Delta t}+\left(\frac{1}{2}\sigma_1^{n+1^2}\frac{\partial^2 V^{n+1}}{\partial r^2}+\rho^{n+1}\sigma_1^{n+1}\sigma_2^{n+1}\frac{\partial^2 V^{n+1}}{\partial r\partial u}+\frac{1}{2}\sigma_1^{n+1^2}\frac{\partial^2 V^{n+1}}{\partial u^2}\right.\\&\quad\left.+(\theta^{n+1}+u-a^{n+1}r)\frac{\partial V^{n+1}}{\partial r}-b^{n+1}u\frac{\partial V^{n+1}}{\partial u}-rV^{n+1}\right)\alpha\\&\quad+\left(\frac{1}{2}\sigma_1^{n^2}\frac{\partial^2 V^{n}}{\partial r^2}+\rho^{n}\sigma_1^{n}\sigma_2^{n}\frac{\partial^2 V^{n}}{\partial r\partial u}+\frac{1}{2}\sigma_1^{n^2}\frac{\partial^2 V^{n}}{\partial u^2}\right.\\&\quad\left.+(\theta^{n}+u-a^{n}r)\frac{\partial V^{n}}{\partial r}-b^{n}u\frac{\partial V^{n}}{\partial u}-rV^{n}\right)(1-\alpha)=0.\end{aligned}$$

The top indices n and $n+1$ are used for the values at different time levels, where the values at time level n are known and the values at time level $n+1$ are unknown. For ease of readability we will use the fully implicit time discretisation in the following.

The computational domain Ω is discretised into finite volumes $\Omega_i, i = 1, \ldots, N$. The next step is to integrate the equation over these finite subdomains:

$$\begin{aligned}
&\int_{\Omega_i} \frac{V^{n+1} - V^n}{\Delta t} d(r,u) + \int_{\Omega_i} \frac{1}{2}\sigma_1^2 \frac{\partial^2 V^{n+1}}{\partial r^2} + \rho\sigma_1\sigma_2 \frac{\partial^2 V^{n+1}}{\partial r \partial u} \\
&\quad + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2} d(r,u) + \int_{\Omega_i} (\theta + u - ar)\frac{\partial V^{n+1}}{\partial r} \\
&\quad - bu\frac{\partial V^{n+1}}{\partial u} d(r,u) - \int_{\Omega_i} rV^{n+1} d(r,u) = 0 \quad \forall i = 1, \ldots, N.
\end{aligned}$$

This is equivalent to

$$\begin{aligned}
&\int_{\Omega_i} \frac{V^{n+1} - V^n}{\Delta t} d(r,u) + \int_{\Omega_i} \frac{\partial}{\partial r}\left(\frac{1}{2}\sigma_1^2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial u}\right) \\
&\quad + \frac{\partial}{\partial u}\left(\frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2}\right) d(r,u) \\
&\quad + \int_{\Omega_i} \frac{\partial}{\partial r}((\theta + u - ar)V^{n+1}) - \frac{\partial}{\partial u}(buV^{n+1}) + aV^{n+1} + bV^{n+1} d(r,u) \\
&\quad - \int_{\Omega_i} rV^{n+1} d(r,u) = 0 \quad \forall i = 1, \ldots, N.
\end{aligned}$$

Those volume integrals which contain a divergence term are converted into surface integrals by the divergence theorem and are evaluated as fluxes across the boundaries Γ_i of each finite volume. (n_r, n_u) denotes the outer unit normal vector at the boundaries.

$$\begin{aligned}
&\int_{\Omega_i} \frac{V^{n+1} - V^n}{\Delta t} d(r,u) + \int_{\Gamma_i} \left(\frac{1}{2}\sigma_1^2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial u}\right) n_r \\
&\quad + \left(\frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2}\right) n_u ds + \int_{\Gamma_i} ((\theta + u - ar)V^{n+1}) n_r \\
&\quad - (buV^{n+1}) n_u ds + \int_{\Omega_i} (a + b - r)V^{n+1} d(r,u) = 0 \quad \forall i = 1, \ldots, N.
\end{aligned}$$

The finite dimensional equation is then obtained by the use of quadrature rules for the given integrals. As outlined in the introduction the discretisation of the convection term requires special attention. The flux across the boundaries due to convection has to be treated with special upwind techniques, like Lax–Wendroff or QUICK schemes (see Morton 1996). Detailed analysis of the obtained numerical schemes leads to the conclusion that the introduction of upwind schemes is equivalent to the addition of artificial numerical diffusion.

In the early references the finite volumes are usually rectangular and occasionally quadrilateral, extending to hexahedral volumes in three dimensions. In the case of rectangular finite volumes the obtained discretisation schemes are equivalent to the one obtained using the finite difference approach.

Finite element method

The finite volume method itself can be treated as a variant of the finite element method. The starting point for the finite element method is the weak formulation of the given equation. Under the assumption that we are looking for a function V in a function space U we can write the weak form of the already implicitly time discretised problem as:

Find $V^{n+1} \in U$ such that, for all $w \in U$,

$$\begin{aligned}
&\int_\Omega \frac{V^{n+1}-V^n}{\Delta t} w\,d(r,u) + \int_\Omega \left(\frac{\partial}{\partial r}\left(\frac{1}{2}\sigma_1^2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial u}\right)\right.\\
&\quad \left.+\frac{\partial}{\partial u}\left(\frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2\frac{\partial^2 V^{n+1}}{\partial u^2}\right)\right) w\,d(r,u)\\
&\quad + \int_\Omega \left((\theta + u - ar)\frac{\partial V^{n+1}}{\partial r} - bu\frac{\partial V^{n+1}}{\partial u}\right) w\,d(r,u)\\
&\quad - \int_\Omega (rV^{n+1}) w\,d(r,u) = 0.
\end{aligned}$$

Applying Gauss' theorem in the second-order terms leads us to

Find $V^{n+1} \in U$ such that, for all $w \in U$,

$$\begin{aligned}
&\int_\Omega \frac{V^{n+1}-V^n}{\Delta t} w\,d(r,u) - \int_\Omega \left(\frac{1}{2}\sigma_1^2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial u}\right)\frac{\partial w}{\partial r}\\
&\quad + \left(\frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2\frac{\partial^2 V^{n+1}}{\partial u^2}\right)\frac{\partial w}{\partial u} d(r,u)\\
&\quad + \int_\Gamma \left(\frac{1}{2}\sigma_1^2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial u}\right) w n_r\\
&\quad + \left(\frac{1}{2}\rho\sigma_1\sigma_2\frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2\frac{\partial^2 V^{n+1}}{\partial u^2}\right) w n_u\,ds\\
&\quad + \int_\Omega \left((\theta + u - ar)\frac{\partial V^{n+1}}{\partial r} - bu\frac{\partial V^{n+1}}{\partial u}\right) w\,d(r,u)\\
&\quad - \int_\Omega (rV^{n+1}) w\,d(r,u) = 0.
\end{aligned}$$

Consider a discretisation Ω_i, $i = 1, \ldots, N$ of the domain Ω and $U_h \subset U$ a finite element space that consists of piecewise polynomials. Replacing the trial and test space U by this finite dimensional space U_h and approximating the function V by a linear combination of basis functions of the trial space lead to the finite dimensional problem. This would be the standard finite element approach disregarding possible difficulties caused by dominating convection. Up to now the special type of the equation is not taken into account. A rather elegant way to introduce upwind techniques to this scheme is used in the method of streamline diffusion (see also Roos *et al*. 1996).

Streamline diffusion—going with the flow

The fundamental idea of this method is to add extra diffusion in the direction of the streamline—hence the name streamline diffusion. From the technical point of view this is realised by replacing the test function w with a test function of the form $w + \delta_i v \cdot \nabla w$, where $v(= (\theta + u - ar, -bu)^T)$ denotes the velocity, and δ_i is called the SD-parameter.

The weak formulation then reads as:

Find $V^{n+1} \in U$ such that, for all $w \in U$,

$$
\begin{aligned}
&\int_\Omega \frac{V^{n+1} - V^n}{\Delta t} w d(r,u) - \int_\Omega \left(\frac{1}{2}\sigma_1^2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial u} \right) \frac{\partial w}{\partial r} \\
&+ \left(\frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2} \right) \frac{\partial w}{\partial u} d(r,u) \\
&+ \int_\Gamma \left(\frac{1}{2}\sigma_1^2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial u} \right) w n_r \\
&+ \left(\frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2} \right) w n_u ds \\
&+ \int_\Omega \left((\theta + u - ar) \frac{\partial V^{n+1}}{\partial r} - bu \frac{\partial V^{n+1}}{\partial u} \right) w d(r,u) \\
&- \int_\Omega (r V^{n+1}) w d(r,u) + \sum_{i=1}^N \int_{\Omega_i} \delta_i \frac{V^{n+1} - V^n}{\Delta t} v \cdot \nabla w d(r,u) \\
&+ \sum_{i=1}^N \int_{\Omega_i} \delta_i \left(\frac{\partial}{\partial r} \left(\frac{1}{2}\sigma_1^2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial u} \right) \right. \\
&\left. + \frac{\partial}{\partial u} \left(\frac{1}{2}\rho\sigma_1\sigma_2 \frac{\partial V^{n+1}}{\partial r} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 V^{n+1}}{\partial u^2} \right) \right) v \cdot \nabla w d(r,u) \\
&+ \sum_{i=1}^N \int_{\Omega_i} \delta_i \left((\theta + u - ar) \frac{\partial V^{n+1}}{\partial r} - bu \frac{\partial V^{n+1}}{\partial u} \right) v \cdot \nabla w d(r,u) \\
&+ \sum_{i=1}^N \int_{\Omega_i} \delta_i (r V^{n+1}) v \cdot \nabla w d(r,u) = 0.
\end{aligned}
$$

The additional term in the convective part is:

$$\sum_{i=1}^{N} \int_{\Omega_i} \delta_i \left((\theta + u - ar)^2 \frac{\partial V^{n+1}}{\partial r} \frac{\partial w}{\partial r} - (bu)^2 \frac{\partial V^{n+1}}{\partial u} \frac{\partial w}{\partial u} \right) \, d(r, u),$$

which has the typical form of a diffusion term. The SD-parameter δ_i depends on the size of the finite elements and on the convection–diffusion ratio, so artificial diffusion is chosen higher in convection-dominated regions and smaller in regions where diffusion dominates.

Although the size of the computational domain is, in principle, unbounded, we have to do our calculations on a bounded domain. It is always difficult to find appropriate and realistic boundary conditions for each structured financial instrument considered. We choose the size of the computational domain in a way such that the information of the prescribed boundary condition does not get through to the centre, during the considered time interval. The centre of the domain is determined by the current short rates. So the choice of boundary conditions, which have to be set for solving the partial differential equation, has no influence on the solution. This may be interpreted in such a way that the probability of very high or low, maybe even negative, interest rates is very small.

Therefore it is clear that the size of the computational domain depends on the lifetime of the considered instrument and on the parameter which forms the coefficient functions of the partial differential equation: volatility, drift and mean reversion.

In the method of finite elements we are very flexible concerning the discretisation of Ω. Structured as well as unstructured grids with adaptive refinement in regions where it is necessary can be chosen. The standard setting in our calculation is a structured, two-dimensional, quadrilateral grid with graded higher resolution in both directions near the values of interest of the factors r and u.

Discretisations in time and space (r-u-plane) can be chosen independently in the case that we use implicit time discretisation, either fully implicit or some kind of Crank-Nicolson (see e.g. Duffy 2004).

Solution of the linear equations

The discretisation leads then to sparse linear systems with typically thousands of variables for each time step. These are then solved iteratively by Krylov subspace techniques, which typically show very fast convergence.

Comparison to analytic solution

In Table 1 we compare the numerical results for the pricing of zero coupon bonds with face amount 1 and different lifetimes obtained by the use of standard finite elements, finite elements with streamline diffusion, and the analytical solution under the two factor Hull–White interest rate model with constant model parameters. The used parameter settings are:

$$a = 1.2, \quad b = 0.03, \quad \theta = 0.05, \quad \sigma_1 = 0.02, \quad \sigma_2 = 0.01, \quad \rho = 0.5$$

TABLE 1: VALUES OF ZEROBONDS FOR A 30 × 30 GRID

Lifetimes	Analytical solution	Standard finite elements	Finite elements with streamline diffusion
1 year	0.954581	0.95458	0.95458
10 years	0.661886	0.661777	0.661855
20 years	0.461421	0.460902	0.461405
40 years	0.268027	0.265955	0.268311

These results confirm, even for this simple example, that the longer the lifetime of an instrument, the more important the usage of upwind techniques. We used a discretisation with a time step of 20 days and a space discretisation of 30×30 points (which are fairly few points for instruments with such long lifetimes).

If we use a space discretisation which is even coarser, namely 10×10 (obviously too coarse), we still obtain realistic results in the streamline diffusion case, but unacceptable results for long life times in the standard finite element case (see Table 2).

TABLE 2: VALUES OF ZEROBONDS FOR A 10 × 10 GRID

Lifetimes	Analytical solution	Standard finite elements	Finite elements with streamline diffusion
1 year	0.954581	0.954592	0.954582
10 years	0.661886	0.663975	0.661414
20 years	0.461421	0.472521	0.459271
40 years	0.268027	0.464601	0.262467

Software architecture

We laid special emphasis on implementing these numerical techniques in a strictly object-oriented framework. We used C++ as a programming language utilising the concepts of objects, class hierarchies and polymorphism.

- An *object* has state (data) and behaviour (functions). Each object is created from a class which is a specification of the data and functions. All objects of a class have common behaviour but generally different state.
- Using *class hierarchies*, classes with common components and operations need not be recoded. This mechanism is called inheritance.
- And last but not least, *polymorphism* allows different kinds of objects that have common behaviour to be used in code that only uses this common behaviour.

A detailed introduction to the object-oriented programming style with special emphasis on scientific and engineering programs can be found in Barton–Nackman (1994), for a general description and as a reference manual see Stroustrup (2000).

How are these concepts realised in our code?

Each instrument which can be priced by our finite element code consists of the base class *BasisInstrument* and different *AttributeManagers*. These *AttributeManagers* handle different possible attributes of a structured financial instrument, like callability, coupon payments, or discrete dividends. So, for example, a callable, convertible, fixed rate bond inherits the same class *Callable* as a callable constant maturity floater. So the implementation of a new structured instrument having already existing attributes is rather easy. All attributes which exist already can be combined with new ones to add new instruments.

The core of the two-factor pricing is built by the class FEPricer (FiniteElementPricer). This class knows everything needed about finite elements with streamline diffusion. With the aid of pointers to an object of the class BasisModel and to an object of the class BasisInstrument the information which two-factor model should be used and which instrument should be priced is obtained. In this part of the program, polymorphism is strongly applied.

Going further

In the previous sections, we have derived the numerical schemes for the solution of the two-factor Hull–White differential equation. These methods (finite elements and streamline diffusion) can of course also be applied to different problems like: different interest rate models which can be written as PDEs, quanto swap problems being built from two one-factor interest rate models, callable convertible bonds and many more. The techniques can also be applied for models in more than two space dimensions. Realistically, there will be a performance problem in problems with 4+ space dimensions which would lead to equations with millions of unknowns.

Until now, we have not said too much about end and interface conditions. Consider, e.g., a callable reverse CMS, i.e. a bond, which pays annual coupons of, say, 10% minus the 5 year swap rate, capped at 7% and floored at 2%. These coupons should be set at the beginning of each coupon period. The lifetime of these instruments is typically quite long (10 to 30 years). To make it more complex, the instrument is equipped with a Bermudan callability at each coupon date.

How do we obtain the swap rates at the coupon set dates? The Hull–White equation has a Green's function (the calculations may become quite tedious if the parameters in the model are not constant but, say, piecewise constant). The value $V(r, u, t)$ of a zero coupon bond maturing at a time T requires the calculation of some integrals only. Swap rates can then be obtained by reverse bootstrapping and taking into account the appropriate day count conventions for the swaps.

At maturity, the bond pays the redemption plus the coupon which was set at the beginning of the last coupon period and which we therefore do not know when propagating backwards from maturity. What we can do is calculate the different discount factors from maturity to the coupon set date in the different states of r and u at the coupon set date and then multiply them by the coupon rate at the set date.

If the instrument is callable, we have to compare the staying alive value and the call price and take the minimum at the Bermudan call dates. Continuing this propagating backwards, we finally reach the valuation date.

More software architecture

Our UnRisk library is not linked to some external C++-code, but is installed within Mathematica as an application package. Therefore we have the following architecture.

UnRisk is called by the Mathematica Kernel, which itself is called either by the Mathematica front end or (via Mathematica Link for Excel) by the Excel front end (Figure 2). Using the Excel front end, the user typically obtains market information like interest rates or volatilities from information providers like Reuters or Bloomberg.

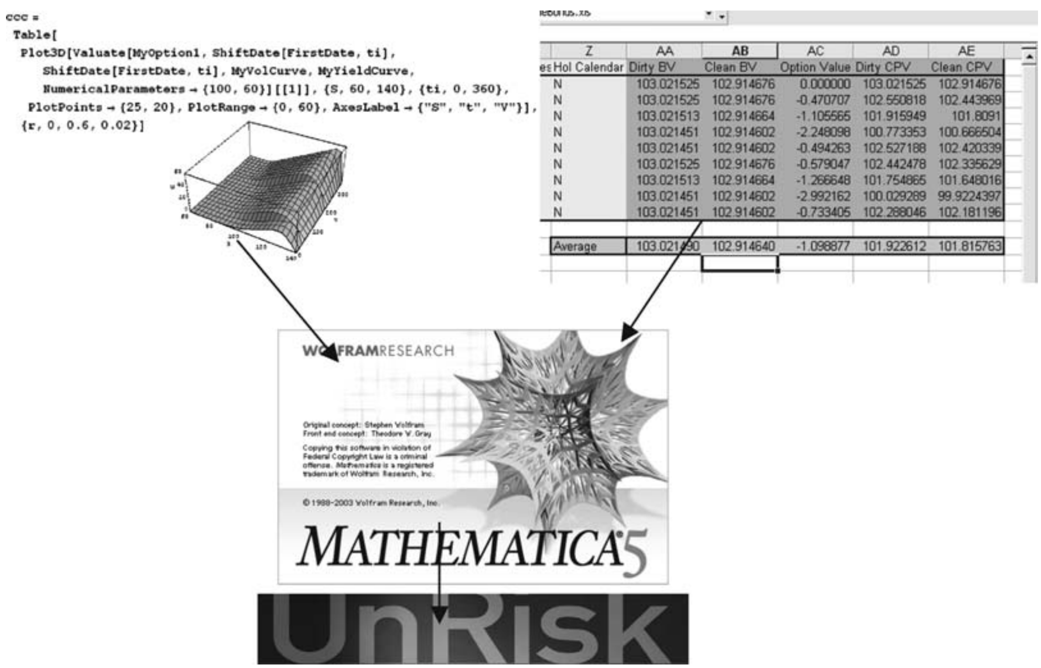


Figure 2: UnRisk software architecture

The Mathematica front end, on the other hand, may be used to write additional code, to produce interactive documents or to generate graphics and animations.

The valuation of a callable reverse floater in the Mathematica front end might look like this:

Load the package

```
Needs["UnRisk`UnRiskFrontEnd`"]
```

Construct a reverse floater (maturity 2024) which pays annual coupons of 12% ("Margin") minus ("Reference Weight") the 5 years (= 60 months) swap rate set in advance ("RefixAttributes") with caps and floors at 8 and 2%, respectively.

```
MyGeneralCMF=MakeGeneralCMFloater[{0.05}, {2024, 10, 10},
{2004, 10, 10}, {2005, 10, 10}, 60, FaceAmount → 100,
CouponFrequency → "Annual", CouponBasis->"30/360",
```

```
RateFrequency->"Annual", RateBasis->"30/360", Margin → 0.12,
ReferenceRate->"Swap", RefixAttributes → {1, 0, 12},
ReferenceWeight→1, Cap →0.08, Floor→0.02];
```

The bond should be callable annually, starting in 2009

```
MyCallSchedule=MakeCallPutSchedule[Table[{2008+i, 10, 10}, 1.},
{i, 1, 15}]];
MyCPGeneralCMF=MakeCPGeneralCMFloater[MyGeneralCMF,
CallSchedule→ MyCallSchedule, CallExercise->"Bermudan",
CallAccrued → True];
```

Construct the two-factor Hull–White model from interest rate curves, cap volatilities and at-the-money swaption volatilities.

```
MyToday= {2004, 10, 26;}
MySwapCurve=MakeSwapCurve[MyToday, {{7, .03331}, {31, .03162},
{62, .03125}, {92, .03043}, {123, .03011}, {153, .02989}, {184,
.0297}, {274, .02959}, {365, .02973}, {730, .0324}, {1095,
.03525}, {1461, .0378}, {1826, .03995}, {2191, .04185}, {2556,
.0435}, {2921, .0448}, {3286, .0459}, {3651, .0468}, {4380,
.04825}, {5475, .0497}, {7300, .051}, {9125, .05145}, {10950,
.0513}, MoneyMarketBasis->"ACT/360", SwapBasis->"30/360",
SwapFrequency->"Annual"];
MyYieldCurve=MakeYieldCurve[MySwapCurve];
MyCapStrikes={0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.055,
0.06, 0.07, 0.08, 0.09, 0.1};
MyCapMaturities={{2, "30/360", "Quarter-Annual"}, {3, "30/360",
"Semi-Annual"}, {4, "30/360", "Semi-Annual"}, {5, "30/360",
"Semi-Annual"}, {6, "30/360", "Semi-Annual"}, {7, "30/360",
"Semi-Annual"}, {8, "30/360", "Semi-Annual"}, {9, "30/360",
"Semi-Annual"}, {10, "30/360", "Semi-Annual"}};
MyCapVolas ={{.288, .262, .249, .253, .26, .268, .276, .283,
.296, .31, .324, .339 {.289, .257, .236, .221, .218, .219,
.222, .229, .239, .252, .267, .284}, {.281, .249, .224, .207,
.199, .196, .196, .199, .206, .218, .233, .247}, {.274, .242,
.216, .198, .187, .18, .178, .178, .183, .193, .205, .219},
{.267, .236, .21, .191, .178, .168, .165, .164, .166, .174,
.184, .197}, {.261, .231, .205, .185, .171, .161, .156, .154,
.154, .16, .169, .18}, {.256, .226, .201, .181, .166, .155,
.149, .146, .146, .15, .158, .167}, {.251, .222, .198, .177,
.162, .151, .144, .14, .139, .142, .149, .157}, {.246, .219,
.195, .174, .159, .148, .139, .136, .135, .137, .143, .15}};

MySwaptionExpiries={2, 5, 10;}
MySwaptionEnds={3, 5, 10, 20;}
MySwaptionVolas ={{.179, .156, .131, .118 {.129, .121, .112,
.105}, {.105, .104, .101, .096}};}
MySwapFrequency="Annual";
MySwapBasis="ACT/360";

MyModel=Make2DModel[MyYieldCurve, MyCapMaturities,
MyCapStrikes, MyCapVolas, MySwaptionExpiries, MySwaptionEnds,
MySwaptionVolas, SwapFrequency->MySwapFrequency, SwapBasis-
>MySwapBasis];
```

The calibration problem is an ill-posed problem meaning that small perturbations in the data can lead to arbitrarily large perturbations in the resulting interest rate model parameters if no special stabilising techniques, so-called regularisation methods, are applied. We will discuss this aspect in a forthcoming paper.

Our experience shows that one should use as many swaption data as available especially on the long end of lifetimes to obtain good pricing results for bonds with long lifetimes.

Valuate the bond

```
SettlementDay=ShiftByBusinessDays[MyToday, 3];
Valuate[MyCPGeneralCMF, MyToday, SettlementDay, MyModel]
{113.676,113.426,-11.3674,102.309,102.059}
```

The returned vector contains dirty and clean value of the pure reverse CM floater (without callability), the option value of the callability, and dirty and clean value of the callable reverse CMF.

Conclusions

We have presented how finite element techniques can successfully be applied to the pricing of complex structured instruments. Streamline diffusion turns out to be a method which is capable of stabilising problems with large or dominating convection.

Authors

Andreas Binder got his Ph.D. in Applied Mathematics (University of Linz) in 1991 for the numerical treatment of some problems in continuous casting of steel. He is CEO of MathConsult since 1996. Thereafter MathConsult has worked on numerical schemes in engineering applications and computational finance. In 2001, they released the first version of the UnRisk PRICING ENGINE.

Andrea Schatz worked on the mathematical modelling and numerical treatment of the COREX process for iron production to obtain her Ph.D. in Industrial Mathematics (University of Linz). She is responsible for the finite element development in the UnRisk PRICING ENGINE.

UnRisk is a registered trademark of MathConsult.

REFERENCES

■ Barton, J. J. and Nackman, L. R. (1994) *Scientific and Engineering C++, An Introduction with Advanced Techniques and Examples*. Addison-Wesley, New York.

■ Duffy, D. J. (2004) A critique of the Crank Nicolson scheme strengths and weaknesses for financial instrument pricing. *Wilmott magazine*, July, 68–76.

■ Hull, J. and White, A. (1994) Numerical procedures for implementing term structure models II: two-factor models. *Journal of Derivatives*, 37–48.

■ Morton, K. W. (1996) *Numerical Solution of Convection-Diffusion Problems*. Chapman & Hall, London, UK.

■ Roos, H.-G., Stynes, M. and Tobiska, L. (1996) *Numerical Methods for Singularly Perturbed Differential Equations—Convection-Diffusion and Flow Problems*. Springer Verlag, Berlin.
■ Stroustrup, B. (2000) *The C++ Programming Language*. Addison-Wesley, Reading, Mass., USA.
■ UnRisk Manual: Available from www.unriskderivatives.com

Acknowledgement

The work of Andrea Schatz has been supported by the Austrian Science Foundation (FWF, www.fwf.ac.at) in the project E67, 'Fast numerical methods in computational finance'.

28
No Fear of Jumps

Y. d'Halluin, D. M. Pooley** and P. A. Forsyth***

Jump diffusion-based models have recently increased in popularity. In this chapter, we develop robust and efficient techniques for the numerical solution of option pricing. models where the underlying process is a jump diffusion process. The numerical techniques can be applied to a variety of contingent claim valuations. Numerical examples for European, American and Parisian options are provided.

1 Introduction

In 1973, the Black–Scholes model revolutionized derivative pricing (Black and Scholes 1973). Using only a volatility and an interest rate, Robert Black and Myron Scholes developed an arbitrage-free pricing formula that did not require knowledge of investor beliefs about the underlying stock's expected return. However, over the years practitioners have recognized the limitations of the Black–Scholes model. In particular, the constant volatility assumption is insufficient to capture the smile or skew that is exhibited by the implied volatilities of traded financial options.

To better capture these volatility profiles, numerous avenues of research have been explored which either extend the Black–Scholes model or explore completely new approaches. Among these extensive works, the jump diffusion model (Merton 1976) and the stochastic volatility model (which could include jumps as well) (Bates 1996, Scott 1997, Bakshi *et al*. 1997) appear to be the most popular among practitioners. Unfortunately, a large portion of the literature devoted to these approaches is limited to analytical or quasi-analytical solutions for vanilla options. Very few of these methods can be extended to price exotic or path-dependent options. For these more complicated scenarios, numerical partial differential equation techniques must be used.

The objective of this chapter is to present a robust and efficient numerical method for solving the partial integro differential equation (PIDE) which arises from the jump diffusion model. We limit ourselves to pricing options under the jump diffusion model, but this framework is also applicable to credit risk models or more complex valuation models such as stochastic volatility with jumps. In the latter case, one simply has to solve a two-dimensional PIDE problem, and apply the techniques presented below for the jump diffusion part in the stock direction. A major

Contact addresses: *School of Computer Science, University of Waterloo, Waterloo ON, Canada, **ITO 33 SA, 36 rue Lacépède, 75005 Paris, France
E-mail: yann@ito33.com, david@ito33.com, paforsyt@uwaterloo.ca

advantage of the methods introduced here is that they are easily added to existing numerical option pricing software. In particular, software that uses an implicit approach for valuing American options can be easily modified to price American options with jump diffusion.

The title of this chapter is obviously based on the very readable article ‘Fear of Jumps’ by Lewis (2002). This article was mostly analytical in nature, and relied on an equilibrium-based approach to option pricing. In contrast, this chapter has a numerical focus for pricing options under jump diffusion. Further, we attempt to convince the reader that adding a jump component to pricing software can be approached with ‘no fear’. Alternatively, this chapter could have been entitled ‘Fear of No Jumps’, as our examples are intended to show that a jump component adds essential features to a pricing model. Without these features, one should be concerned about the accuracy and stability of the pricing framework.

Our technique is similar in some respects to Zhang (1997), though less constrained in terms of stability restrictions. Our method also offers a higher rate of convergence than Zhang’s. Similar comments apply if we compare our approach to that of Andersen and Andreasen (2000), at least in the case of American options.

In this chapter, the PIDE presented by Merton (1976) and Andersen and Andreasen (2000) is studied exclusively. While it is true that Merton’s assumption about jump risk being diversifiable does not hold for index-based options, and in this case one must use an equilibrium-based method (Lewis 2002) or a mean variance hedging approach (Ayache *et al*. 2004), the PIDEs resulting in either case are essentially identical. Consequently, the numerical techniques presented here can be applied.

This chapter is organized as follows. In section 2, the numerical method for solving the option pricing PIDE which results from a jump diffusion model is presented. In section 3, a wide variety of numerical examples of exotic, path-dependent contracts are presented. In particular, we include numerical examples for American and Parisian options. Finally, section 4 contains concluding remarks.

2 Mathematical model

This section provides an overview of the mathematical modeling issues that arise in a jump diffusion framework. The presentation and notation closely follows that of d’Halluin *et al*. (2003). However, particular attention is paid here to the practical issues that arise in a numerical implementation. Further, since the goal of this chapter is somewhat illustrative, several proofs and technical details have been omitted. The reader is referred to d’Halluin *et al*. (2005) and the references therein for a complete treatment of the theory of option pricing in a jump diffusion framework.

In the usual (no jumps) Black–Scholes model for option pricing (Black and Scholes 1973, Merton 1976), the underlying asset price S evolves according to

$$\frac{dS}{S} = \mu\, dt + \sigma\, dZ, \tag{2.1}$$

where μ is the (real) drift rate, σ is the volatility, and dZ is the increment of a Gauss–Wiener process. Let $V(S,t)$ be the value of a contingent claim that depends on the underlying asset S and time t. By appealing to the principle of no-arbitrage, a partial differential equation (PDE) for the value of V can be derived:

$$V_\tau = \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV, \tag{2.2}$$

where $\tau = T - t$ is the time remaining until expiry T, and r is the continuously compounded risk-free interest rate. Equation (2.2) is simply a second order parabolic PDE of one space dimension and one time dimension. This equation has been the subject of countless studies, and is well understood from a variety of viewpoints (financial, mathematical, numerical). Letting

$$\mathcal{L}V = \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV \tag{2.3}$$

equation (2.2) can be written in the simple form

$$V_\tau = \mathcal{L}V. \tag{2.4}$$

It is assumed that the reader is familiar with the numerical solution of PDEs of the form (2.4). Software for this problem is easily written, and off-the-shelf implementations are readily available.

Nevertheless, the process specified by equation (2.1) is not sufficient to explain observed market behavior (Bakshi and Cao 2002). In reality, stock prices have been observed to have large instantaneous jumps. Such behavior can be modeled by the risk-neutral process (Merton 1976)

$$\frac{dS}{S} = (r - \lambda\kappa)\,dt + \sigma\,dZ + (\eta - 1)dq, \tag{2.5}$$

where dq is a Poisson process (independent of the Brownian motion), and $\eta - 1$ is an impulse function producing a jump from S to $S\eta$. If λ is the arrival intensity of the Poisson process, then $dq = 0$ with probability $1 - \lambda dt$, and $dq = 1$ with probability λdt. The expected jump size can be denoted by $\kappa = E[\eta - 1]$, where E is the expectation operator.

As is well known, the fair price of a contingent claim $V(S,t)$ under a process of the form (2.5) is given by the following partial integro differential equation (PIDE):

$$V_\tau = \frac{1}{2}\sigma^2 S^2 V_{SS} + (r - \lambda\kappa)\,SV_S - rV + \lambda \int_0^\infty V(S\eta)g(\eta)\,d\eta - \lambda V. \tag{2.6}$$

In equation (2.6), $g(\eta)$ is the probability density function of the jump amplitude η. The probability density function is assumed to have the usual distribution properties, such as $\forall\eta$, $g(\eta) \geq 0$ and $\int_0^\infty g(\eta)\,d\eta = 1$. Letting

$$\hat{\mathcal{L}}V = \frac{1}{2}\sigma^2 S^2 V_{SS} + (r - \lambda\kappa)SV_S - (r + \lambda)V, \tag{2.7}$$

equation (2.6) can be written as

$$V_\tau = \hat{\mathcal{L}}V + \lambda \int_0^\infty V(S\eta)g(\eta)\,d\eta. \tag{2.8}$$

As with $\mathcal{L}V$, the behavior of $\hat{\mathcal{L}}V$ is well understood. Further, it should be straightforward to modify any reasonably designed software that can handle numerically $\mathcal{L}V$ to compute $\hat{\mathcal{L}}V$. Of a more difficult nature is the integral term in equation (2.8).

The obvious approach for the numerical computation of the integral term is to use standard numerical integration methods such as Simpson's rule or Gaussian quadrature. Unfortunately, for a numerical grid of size n, these techniques are $\mathcal{O}(n^2)$. For real-time pricing software, and especially for calibration routines, quicker algorithms are desirable.

To this end, the integral term of equation (2.8) should be computed in a way that is

- efficient (better than $\mathcal{O}(n^2)$),
- robust,
- flexible (can be used with non-linear pricing models, and/or exotic options),
- easily added to existing option pricing software.

All of these properties are satisfied if

- the integral term is evaluated by FFTs, thereby only requiring $\mathcal{O}(n \log n)$ operations per timestep,
- the integral term is applied implicitly, thereby increasing stability and allowing the possibility of second-order convergence.

The FFT evaluation of the integral and the implicit treatment of the resulting terms will be discussed separately below. Following these, an extension to American options will be provided, as well as a brief description of credit risk. Examples which use the techniques described below are provided in section 3.

It should be noted that in some cases, the integral term can be evaluated directly in $\mathcal{O}(n)$ time using fast Gauss transform (FGT) techniques (Greengard and Strain 1991). While this technique works for the case where jump sizes are lognormally distributed, it is not clear if it works for more general distributions. Furthermore, numerical experiments show that for any practical grid size the FFT approach for evaluating the integral term is faster than the FGT method. (Note that the integral needs only to be evaluated with an accuracy consistent with the discretization of the PDE.)

2.1 FFT evaluation

Before the integral term of equation (2.8) can be evaluated by FFTs, it must be manipulated into the form of a correlation integral. Once this process is done, at least two numerical issues remain. First, standard FFT algorithms require an equally spaced grid, whereas an efficient PDE grid will be unequally spaced. Interpolation must be used to move from one grid to the other. Second, since the input functions to the FFT routines will be non-periodic, wrap-around pollution can negatively affect the solution. These numerical issues are discussed in section 2.1.2.

Manipulation Ignoring the leading λ, the integral term in equation (2.8) is

$$I(S) = \int_0^\infty V(S\eta)g(\eta)\,d\eta. \tag{2.9}$$

The goal is to turn this expression into a correlation product which can be evaluated by FFT techniques. Letting $x = \log(S)$ and applying the change of variable $y = \log(\eta)$, we obtain

$$I = \int_{-\infty}^{+\infty} \overline{V}(x+y)\overline{f}(y)dy, \tag{2.10}$$

where $\overline{f}(y) = g(e^y)e^y$ and $\overline{V}(y) = V(e^y)$. The $\overline{f}(y)$ term can be interpreted as the probability density of a jump of size $y = \log \eta$. Conveniently, equation (2.10) corresponds to the correlation product $\overline{V}(y) \otimes \overline{f}(y)$. In discrete form, equation (2.10) becomes

$$I_i = \sum_{j=-N/2+1}^{j=N/2} \overline{V}_{i+j} \overline{f}_j \Delta y + \mathcal{O}\left((\Delta y)^2\right), \tag{2.11}$$

where $I_i = I(i\Delta x)$, $\overline{V}_j = \overline{V}(j\Delta x)$, and

$$\overline{f}_j = \overline{f}(j\Delta y) = \frac{1}{\Delta x} \int_{x_j-\Delta x/2}^{x_j+\Delta x/2} \overline{f}(x)\, dx \tag{2.12}$$

It has been assumed that $\Delta y = \Delta x$.

Assuming that $\overline{f}$ is real (a safe assumption for financial applications), the discrete correlation of equation (2.11) can be evaluated using FFTs since

$$I_i = IFFT\left(\left(FFT(\overline{V})\right)\left(FFT(\overline{f})\right)^*\right)_i \tag{2.13}$$

where $(\cdot)^*$ denotes the complex conjugate. For efficiency, $FFT(\overline{f})$ can be pre-computed and stored. During each timestep (or each iteration of an iterative method), an *FFT* and an inverse *FFT* must be computed.

Numerical issues A typical grid for the discretization of $\hat{\mathcal{L}}V$ in equation (2.8) will be unequally spaced in S coordinates. For example, small mesh spacing will be used near strikes or barriers, with large mesh spacing elsewhere. However, the discrete form of the correlation integral (2.11) requires an equally spaced grid in $\log(S)$ coordinates. It is highly unlikely that these two grids are fully compatible. Hence, values must be interpolated between the two grids.

In particular, values of V on the unequally spaced S grid must be interpolated onto an equally spaced $\log(S)$ grid. The computation of equation (2.13) can then be performed.[1] Finally, the resulting equally spaced $\overline{V}$ data needs to be interpolated back onto the unequally spaced S grid. The overall process is summarized in algorithm (1). If linear or higher order interpolation is used, algorithm (1) is second-order correct. This is consistent with the discretization error in the PDE and the midpoint rule used to evaluate the integral in equation (2.11).

Algorithm 1 Method for computing the integral term of equation (2.8) by FFTs.

Interpolate the discrete values of V onto an equally spaced $\log(S)$ grid. This generates the required values of $\overline{V}_j$.
Carry out the FFT on the $\overline{V}$ data.
Compute the correlation in the frequency domain (with pre-computed $FFT(\overline{f})$ values), using equation (2.13).
Invert the FFT of the correlation.
Interpolate the discrete values of $I(x_i)$ back onto the original S grid.

For the actual FFT evaluation, standard algorithms assume periodic input data. If the input data is not periodic (as with the current application), then the discrete Fourier transform is effectively applied to the periodic extension of the input functions. This can lead to undesirable 'wrap-around pollution', which manifests itself with erroneous values in the solution.

To avoid wrap-around effects, the domain of the integral in equation (2.8) can be extended to the left and right by amounts Δy^- and Δy^+. The integral then becomes

$$I_{ext} = \int_{y_{\min}-\Delta y^-}^{y_{\max}+\Delta y^+} \overline{V}(x+y)\overline{f}(y)dy, \tag{2.14}$$

where $y_{\max} = \log(S_{\max})$, $y_{\min} = \log(S_{\min})$, and $[S_{\min}, S_{\max}]$ are selected appropriately. Unknown values in the range $[y_{\max}, y_{\max} + \Delta y^+]$ can be obtained by linear extrapolation. This assumes that the far field behavior of the option pricing problem is linear. Values in the range $[y_{\min} - \Delta y^-, y_{\min}]$ can be obtained from interpolation on the original S grid, assuming an $S_0 = 0$ grid point has been maintained.

Once the FFT has been performed in the extended domain, values in the extensions are discarded. Because of the extension, values in the original domain will have been less affected by wrap-around pollution.

2.2 Implicit evaluation

We now look at the numerical evaluation of equation (2.8). Let Ψ_i^n denote the discrete form of the integral evaluated at timestep n using data V^n (one can think of Ψ as an application of algorithm (1)). To solve equation (2.8), the $\hat{\mathcal{L}}V$ term must also be discretized. This can be done by any standard method, such as finite differences, finite volumes, or finite elements. Let the discrete form of $\hat{\mathcal{L}}V$ at timestep n be given by $\left(\hat{\mathcal{L}}V\right)_i^n$. A general discretized form of equation (2.8) can then be written as

$$\frac{V_i^{n+1} - V_i^n}{\Delta\tau} = (1-\theta)\left(\hat{\mathcal{L}}V\right)_i^{n+1} + \theta\left(\hat{\mathcal{L}}V\right)_i^n + (1-\theta_J)\lambda\Psi_i^{n+1} + \theta_J\lambda\Psi_i^n, \tag{2.15}$$

where

θ is a time-weighting parameter for $\hat{\mathcal{L}}$

$\theta = 0$ is fully implicit

$\theta = 1/2$ is Crank Nicolson

$\theta = 1$ is fully explicit

θ_J is time-weighting for the jump term Ψ

$\theta_J = \{0, 1/2, 1\}$.

Let M denote the discretization matrix stencil such that

$$-[MV]_i^n = \left(\hat{\mathcal{L}}V\right)_i^n. \tag{2.16}$$

Algorithm 2 Fixed point iteration.

Let $(V^{n+1})^0 = V^n$
Let $\hat{V}^k = (V^{n+1})^k$
Let $\hat{\Psi}^k = (\Psi^{n+1})^k$
Construct vector Ψ^n using algorithm (1)
for $k = 0, 1, 2, \ldots$ until convergence **do**
  Construct vector $\hat{\Psi}^k$ using algorithm (1)
  Solve $[I - (1-\theta)M]\hat{V}^{k+1} = [I + \theta M]V^n + (1-\theta_J)\lambda\hat{\Psi}^k + \lambda\theta_J\Psi^n$
  if $\max_i \frac{|\hat{V}_i^{k+1} - \hat{V}_i^k|}{\max(1, |\hat{V}_i^{k+1}|)} < tolerance$ **then**
    quit
  end if
end for

Equation (2.15) becomes

$$[I - \Delta\tau(1-\theta)M]V^{n+1} = [I + \Delta\tau\theta M]V^n + (1-\theta_J)\lambda\Delta\tau\Psi_i^{n+1} + \theta_J\lambda\Delta\tau\Psi_i^n. \tag{2.17}$$

For standard PDE discretization techniques, the matrix M in equation (2.17) is tridiagonal. Tridiagonal systems are quick and easy to solve. However, an implicit treatment of the jump term ($\theta_J \neq 1$) causes Ψ_i^{n+1} to lead to a highly undesirable dense matrix (all nodal values are coupled in equation (2.10)). On the other hand, a fully explicit treatment of the jump term is easy to adapt to existing code, since only the right-hand side vector needs to be updated. However, while still stable, only first-order convergence is possible.

To allow for an implicit treatment of jumps, a fixed point iteration method must be used. A description of the method is given in algorithm (2). At iteration k known data is used to construct the jump term. Since only the right-hand side is affected, a simple tridiagonal system needs to be solved at each iteration.

Under some fairly mild assumptions—that the discretization of $\hat{\mathcal{L}}$ forms an M-matrix, the probability density function has certain standard properties, the interpolation weights are positive, and that r and λ are positive—it can be proven that algorithm (2) is globally convergent d'Halluin *et al*. (2004). Further, the error at each iteration is reduced by approximately $(1-\theta)\lambda\Delta\tau$, indicating convergence in a small number of iterations (i.e. for typical values, three iterations are sufficient).

2.3 American options

American options can be solved by a simple penalty approach. Details of the penalty approach can be found in Forsyth and Vetzal (2002). Further details with regards to jump diffusion models can be found in d'Halluin *et al*. (2004). Briefly, the penalty approach involves adding a penalty term to the pricing PDE. Equation (2.8) then becomes

$$V_\tau = \hat{\mathcal{L}}V + \lambda\int_0^\infty V(S\eta)g(\eta)d\eta + \rho\max(V^* - V, 0). \tag{2.18}$$

In the limit as $\rho \to \infty$, the solution satisfies $V \geq V^*$. The American constraint is enforced by setting V^* to the payoff of the option.

In the discrete equations, ρ is set independently at each node. If the value at a node i drops below V_i^* (the payoff), then ρ_i is set to a large number. This essentially adds an extra source term to the PDE, thereby increasing the value at the particular node. If the value at a node is greater than V^*, then ρ_i is set to zero, and the regular PDE is solved. This can also be thought of as constraint switching. Wherever the value drops below the V^* threshold, the constraint is switched on and applied. If the value is above the threshold, the constraint is switched off.

As with the evaluation of the integral term, the penalty constraint can be applied explicitly or implicitly. An explicit evaluation simply uses data at the previous timestep to determine when the constraint is activated. An implicit evaluation could use a fixed point iteration (or other non-linear solving method) to apply the constraint using data at the current timestep. If the jump term is already being evaluated using an iterative method, little or no extra cost is incurred by the penalty method. Convergence of the penalty approach for American options in a jump diffusion framework was proven in d'Halluin *et al*. (2004).

2.4 Credit risk

Until this point, jumps in stock price associated with the jump diffusion model have been assumed to occur for arbitrary exceptional events. However, a special jump in asset level occurs in the case of bankruptcy. In pricing corporate and convertible bonds, it is of interest to determine the risk adjusted hazard rate of bankruptcy. If it is assumed that the stock price of a firm jumps to zero on default, then λ_h can be interpreted as the risk adjusted hazard rate of bankruptcy (or default in the case of bonds). In this case, the PDE satisfied by vanilla puts/calls in the presence of a single jump to bankruptcy is given by

$$V_\tau = \frac{1}{2}\sigma^2 S^2 V_{SS} + (r + \lambda_h) S V_S - (r + \lambda_h) V + \lambda_h V(0, \tau). \tag{2.19}$$

Equation (2.19) can be derived by hedging arguments, or by setting κ to -1 and the jump probability density function $g(\eta)$ to the delta function $\delta(0)$ (1 at $\eta = 0$, zero elsewhere) in the usual Merton jump diffusion model.

It is usually assumed that $\lambda_h = \lambda_h(S, t)$, with $\lambda_h(S, t)$ being determined by calibration to observed market prices for vanilla options and credit instruments. Since option prices are usually available for a range of strikes, more information is provided about default rates than is usually available from simply examining credit instruments. Note that equation (2.19) suggests that default risk has an effect on the pricing of vanilla options. As well, if the possibility of a single jump to bankruptcy is assumed, then a hedging portfolio consisting of the option, an underlying asset, and an additional option can be constructed which eliminates both the diffusion risk (a delta hedge) as well as the jump risk (since the jump has only one possible outcome).

3 Results

The examples of this section are intended to compare the regular Black–Scholes model and the jump diffusion model. To ensure a consistent basis for comparison, the following procedure is used:

1. Given some jump diffusion parameters, compute the (numerical) at-the-money price V_{jump} of a European put option.

TABLE 1: INPUT DATA USED TO VALUE VARIOUS OPTIONS UNDER THE LOGNORMAL JUMP DIFFUSION PROCESS. THESE PARAMETERS ARE APPROXIMATELY THE SAME AS THOSE REPORTED IN ANDERSEN AND ANDREASEN (2000) USING EUROPEAN CALL OPTIONS ON THE S&P500 STOCK INDEX IN APRIL OF 1999

volatility: σ	0.15
risk-free rate: r	0.05
jump standard deviation: γ	0.45
jump mean: μ_{mean}	−0.90
jump intensity: λ	0.10
time to expiry: T	0.25
strike: K	1.00

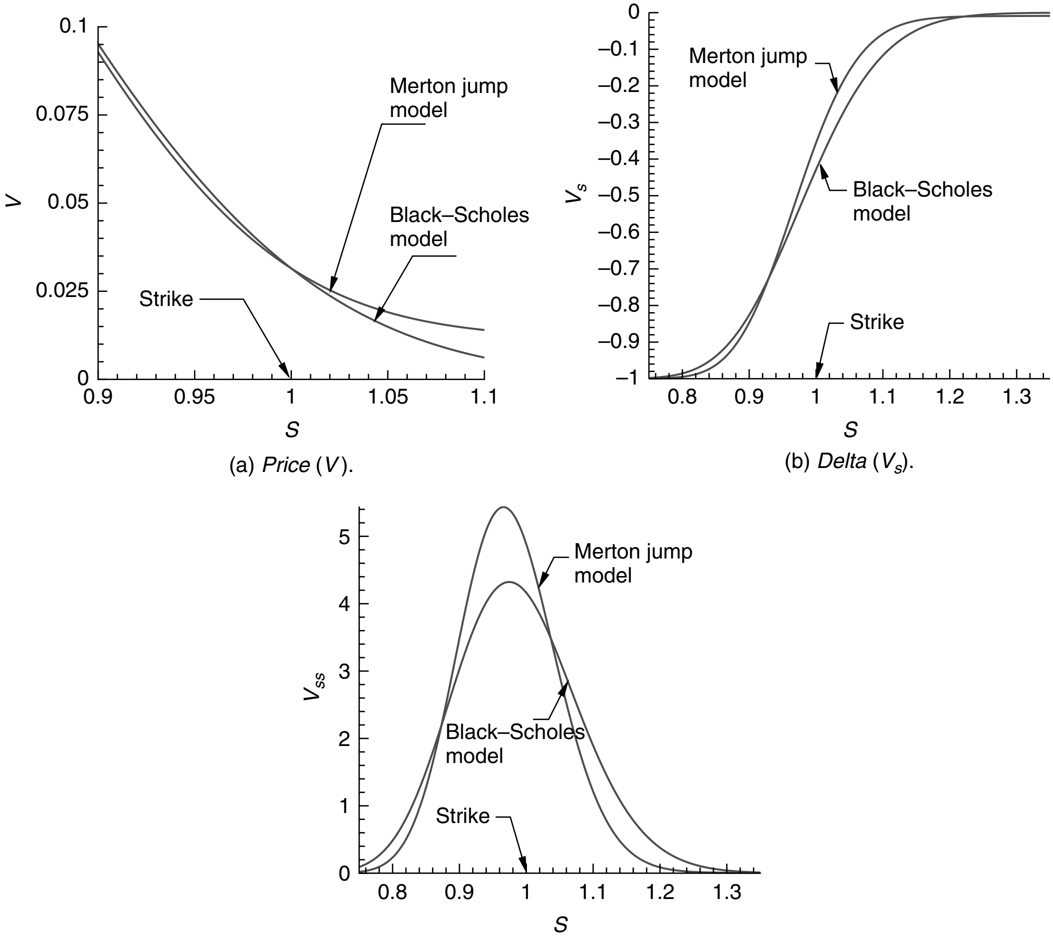
2. Using a constant volatility Black–Scholes model, determine the implied volatility σ_{implied} which matches the option price to the jump diffusion value V_{jump} at the strike K.
3. Value the option using a constant volatility model (no jumps) using the implied volatility σ_{implied} estimated in Step 2.

The first example prices a European put option with and without jumps. Parameters are provided in Table 1. Results are shown in Figure 1. The implied volatility value for the Black–Scholes model is 0.1886. By construction, the prices of the Black–Scholes model and the Merton jump model are equal at the strike price. In-the-money values are larger for the Black–Scholes model, but only slightly. Of interest is the fact that the jump model prices deep out-of-the-money options significantly higher. This reflects the fact that a jump event can dramatically change the moneyness of an option to a much larger extent than a simple diffusion only model.

The delta and gamma plots for the two models are similar, although the jump model plots show greater variation. This indicates that a delta hedge of the jump model may need more frequent rebalancing. Nevertheless, jumps introduce market incompleteness, and simple delta hedging will definitely fail. Optimal hedging in incomplete markets is preferred (Henrotte 2002, Ayache *et al.* 2004). In any case, hedging will require accurate delta and gamma information. It is essential that the numerical scheme produce smooth delta and gamma values.

The second example is a repeat of the first, except that an American put option is priced instead of a European put option. The implied volatility value used is the same as in the previous example: $\sigma_{\text{implied}} = 0.1886$. Results are similar, except that delta values now reach and remain at -1 for low stock prices, while gamma values jump to zero. This jump to zero occurs at the free boundary between the early exercise region and the regular pricing region. The early exercise region is further to the right for the Merton jump model, indicating that jumps cause an increase in the probability that the option should be exercised early.

The last example is for a Parisian knock-out call option. The particular case considered here is an up-and-out call with daily discrete observation dates. This contract ceases to have value if S is above a specified barrier level for a specified number of consecutive monitoring dates. This can be valued by solving a set of one-dimensional problems which exchange information

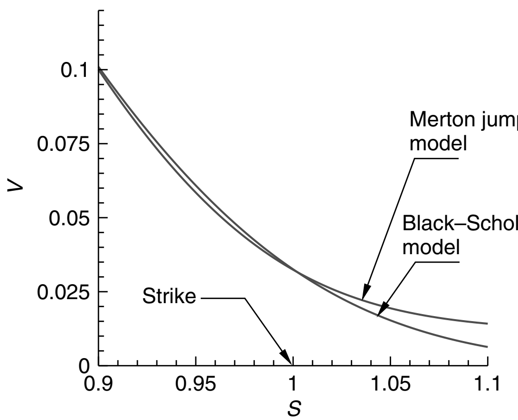


(a) *Price* (V). (b) *Delta* (V_s). (c) *Gamma* (V_{ss}).

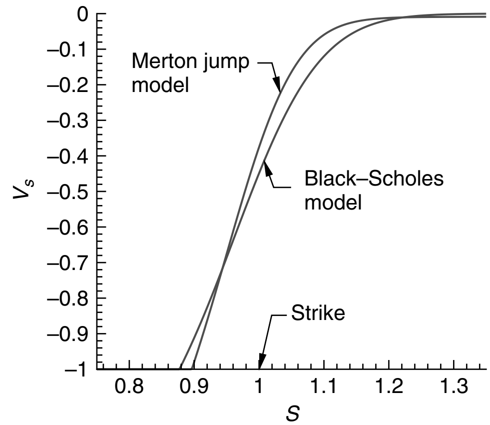
Figure 1: Put option price (V), delta (V_S) and gamma (V_{SS}). The input data is contained in Table 1

at monitoring dates (Vetzal and Forsyth 1999). Base parameters are the same as in Table 1. The knock-out barrier is placed at $S = 1.20$, while the number of consecutive days above the barrier until knock-out is set to 10. The implied volatility value is 0.1886. It is interesting to note that the Merton jump model gives smaller prices for stock values below the strike and above the barrier. This is somewhat in contradiction to the put options, for which deep out-of-the-money prices were higher for the jump model. Nevertheless, the differences are small, and the delta and gamma plots show the far field behavior to be quite similar.

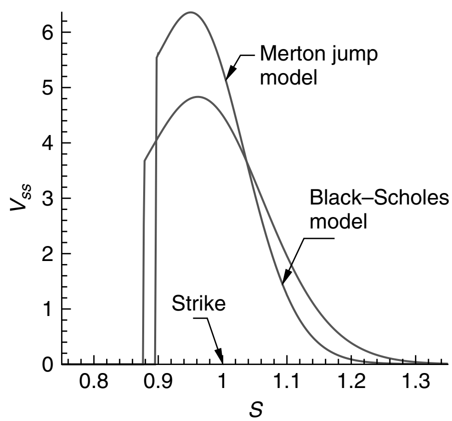
The greatest price difference occurs between the strike and barrier levels. Presumably a jump in this region hides the effect of the (upper) barrier, whereas a pure diffusion model will have its value decreased by the barrier. However, it is difficult to intuitively predict the effect of jumps on prices. For convex payoffs, jumps increase the value of an option. For non-convex payoffs,



(a) *Price* (V).



(b) *Delta* (V_s).



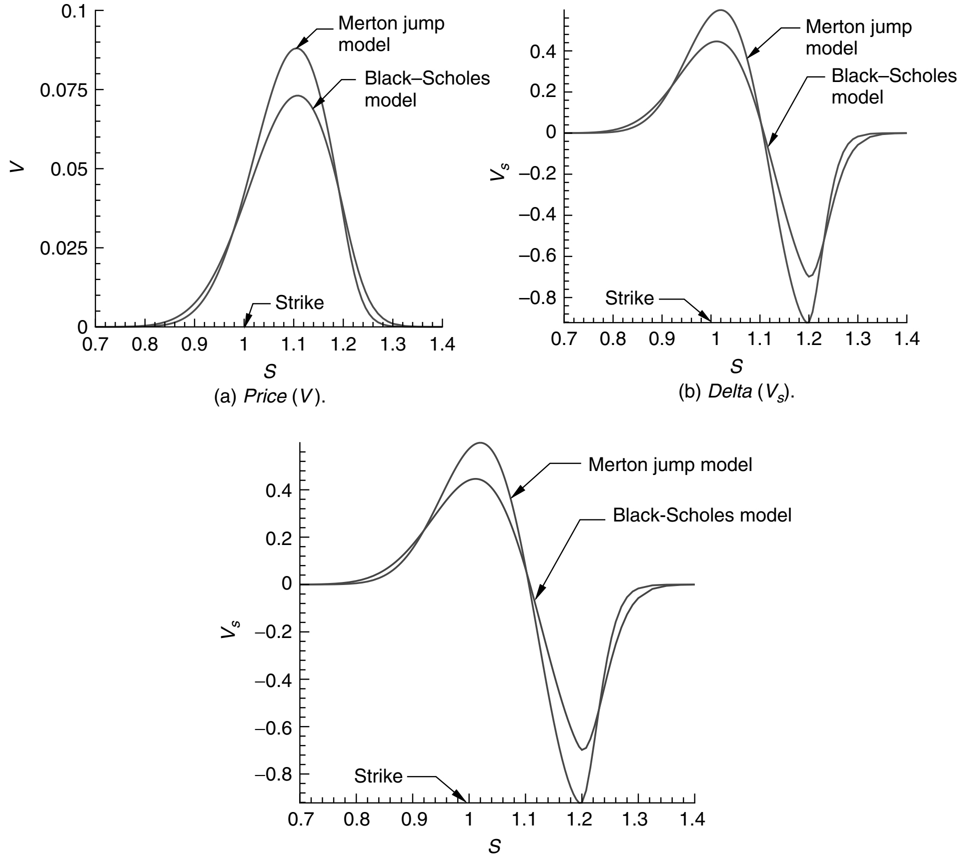
(c) *Gamma* (V_{ss}).

Figure 2: American put option price (V), delta (V_S) and gamma (V_{SS}). The input data is contained in Table 1

as is the case for the Parisian knock-out call, it is not clear what effect jumps will have on the price.

4 Conclusion

This chapter has demonstrated the numerical evaluation of the PIDE resulting from the Merton jump diffusion model in option pricing. The integral term of the pricing equation was evaluated using efficient FFT techniques. The issues of interpolation between unequally spaced PDE grids and equally spaced FFT grids, as well as wrap-around pollution effects, were briefly discussed. A fixed point iteration method was used to obtain an implicit timestepping method without resorting to a full dense matrix solve. Extensions to American options and credit risk were also mentioned.



(a) *Price* (V). (b) *Delta* (V_s). (c) *Gamma* (V_{ss}).

Figure 3: Parisian knock-out call option (V), delta (V_S) and gamma (V_{SS}) with discrete daily observation dates with and without jumps. The barrier is set at $S = 1.20$ and the number of consecutive daily observations to knock-out is 10. The input data is contained in Table 1

Perhaps the biggest advantage of the techniques described in this chapter is the ease with which they can be added to an existing exotic option pricing library. All that is required is that a function be added to the library which, given the current vector of discrete option prices, returns the vector value of the correlation integral. This vector is then added to the right-hand side of the fixed point iteration. This method can even be applied to any jump size probability density function.

The numerical examples showed the effect of jumps on various option values. For European and American put options, the jump diffusion model increases deep out-of-the-money prices. Changes to the hedging parameters—delta and gamma—were also noted. The stability of the methods was alluded to by the smooth delta and gamma plots. An example of a Parisian knock-out option was also provided.

An important issue not addressed in this chapter is hedging jump diffusion models. Since the market is incomplete, simple delta hedging can give large errors. In this case optimal hedging in incomplete markets must be used (Ayache *et al*. 2004).

FOOTNOTE & REFERENCES

1. Methods exist for computing an FFT on unequally spaced data. However, these methods do not appear to be more efficient than the straightforward approach suggested here.

■ Andersen, L. and Andreasen, J. (2000) Jump-diffusion processes: volatility smile fitting and numerical methods for option pricing. *Review of Derivatives Research*, 4, 231–262.
■ Ayache, E., Henrotte, P., Nassar, S. and Wang, X. (2004) Can anyone solve the smile problem? Wilmott, January.
■ Bakshi, G. and Cao, C. (2002) Risk-neutral kurtosis, jumps, and option pricing: evidence from 100 most actively traded firms on the CBOE. Working paper, Smith School of Business, University of Maryland.
■ Bakshi, G., Cao, C. and Chen, Z. (1997) Empirical performance of alternative option pricing models. *Journal of Finance*, 52, 2003–2049.
■ Bates, D. S. (1996) Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options. *Review of Financial Studies*, 9, 69–107.
■ Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–659.
■ d'Halluin, Y., Forsyth, P. A. and Labahn, G. (2004) A penalty method for American options with jump diffusion processes, *Numerische Mathematik*, 97, 321–352.
■ d'Halluin, Y., Forsyth, P. A. and Vetzal, K. R. (2005) Robust numerical methods for contingent claims under jump diffusion processes, *IMA Journal of Numerical Analysis*, 28, 87–112.
■ Forsyth, P. A. and Vetzal, K. R. (2002) Quadratic convergence of a penalty method for valuing American options. *SIAM Journal on Scientific Computation*, 23, 2096–2123.
■ Greengard, L. and Strain, J. (1991) The fast Gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12, 79–94.
■ Henrotte, P. (2002) Dynamic mean variance analysis. Working paper, July.
■ Lewis, A. (2002) Fear of jumps. *Wilmott*, 60–67, December.
■ Merton, R. C. (1976) Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3, 125–144.
■ Scott, L. O. (1997) Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: applications of Fourier inversion methods. *Mathematical Finance*, 7, 413–426.
■ Vetzal, K. R. and Forsyth, P. A. (1999) Discrete Parisian and delayed barrier options: a general numerical approach. *Advances in Futures and Options Research*, 10, 1–16.
■ Zhang, X. L. (1997) Numerical analysis of American option pricing in a jump-diffusion model. *Mathematics of Operations Research*, 22, 668–690.

Index

Index compiled by Terry Halliday