With an unmatched breadth of products and depth of knowledge, FIS is uniquely capable of helping you build and maintain a distinctive position in your market.

We are the trusted partner that offers you the greatest range of choices – not only in what technologies and services to deploy, but also how. This can include customizing software functions, private branding and tailoring customer interfaces or leveraging our deep suite of outsourcing and managed services.

With 30,000 dedicated people in over 100 countries and the recent addition of the Capco organization, we can apply unprecedented power to your unique market strategy. Look closer – and discover the power and value of FIS. Go to fisglobal.com

Visit our new Web site
Learn about our industry-leading products, services and solutions.
Go to fisglobal.com
## Part 1

9   Economists’ Hubris – The Case Of Award Winning Finance Literature  
    Shahin Shojai, George Feiger

19   Tracking Problems, Hedge Fund Replication, and Alternative Beta  
    Thierry Roncalli, Guillaume Weisang

31   Empirical Implementation of a 2-Factor Structural Model for Loss-Given-Default  
    Michael Jacobs, Jr.

45   Regulatory Reform: A New Paradigm for Wealth Management  
    Haney Saadah, Eduardo Diaz

53   The Map and the Territory: The Shifting Landscape of Banking Risk  
    Sergio Scandizzo

63   Towards Implementation of Capital Adequacy (Pillar 2) Guidelines  
    Kosrow Dehnad, Mani Shabrang

67   The Failure of Financial Econometrics: Estimation of the Hedge Ratio as an Illustration  
    Imad Moosa

73   Systemic Risk Seen from the Perspective of Physics  
    Udo Milkau

83   International Supply Chains as Real Transmission Channels of Financial Shocks  
    Hubert Escaith, Fabien Gonguet

## Part 2

101  Chinese Exchange Rates and Reserves from a Basic Monetary Approach Perspective  
    Bluford H. Putnam, Stephen Jay Silver, D. Sykes Wilford

115  Asset Allocation: Mass Production or Mass Customization?  
    Brian J. Jacobsen

123  Practical Attribution Analysis in Asset Liability Management of a Bank  
    Sunil Mohandas, Arjun Dasgupta

133  Hedge Funds Performance Ratios Adjusted to Market Liquidity Risk  
    Pierre Clauss

141  Regulating Credit Ratings Agencies: Where to Now?  
    Amadou N. R. Sy

151  Insurer Anti-Fraud Programs: Contracts and Detection versus Norms and Prevention  
    Sharon Tennyson

157  Revisiting the Labor Hoarding Employment Demand Model: An Economic Order Quantity Approach  
    Harlan D. Platt, Marjorie B. Platt

    Jannis Bischof, Michael Ebert

173  Indexation as Primary Target for Pension Funds: Implications for Portfolio Management  
    Angela Gallo
Driving change in a post-crisis world

Exactly ten years ago, we published the first edition of the Journal. Our objective at the time was to establish a publication that allowed senior financial executives to keep abreast of the latest thinking in finance in a format that they could read and understand. That remains our objective today.

However, we did not envision at the time that we would face a situation where a good number of the established theories in finance would come under such severe criticism and that we would be forced to rebuild the discipline with a new body of knowledge. The two major crises that we witnessed since the inception of the Journal have slowly eroded our trust in financial models, and specifically in how risk is measured and managed.

Today, we find ourselves in a situation where our industry is heavily criticized for its actions and responsibilities during these two crises. It is now time to move on and to think about how we can come together to rethink traditional finance, and its many complex supporting models. It is now time that we recognize that finance academics have as much to learn from practitioners as the latter did from the former. It is also time we start looking forward and begin to develop ideas that can make our industry stronger and more resilient against future crises. The time for looking back is over.

This edition of the Journal aims to introduce some of those ideas that we feel will be part of the foundations of tomorrow’s finance. This is also the first edition in which practical finance receives as much attention as it deserves.

We hope that you will join us on this journey of discovery and help us form the future of finance.

Rob Heyvaert, Founder and CEO, Capco
For the last couple of years we have all been inundated with analyses of why the recent financial crisis took place, who the real culprits were, and what steps are needed to prevent future crises. While these are useful from an academic perspective, what we really need to do is not discuss how old models can be modified so that they can be reliably applied to future crises, but to actually question their validity. What we need to do is ask whether the models that have become the cornerstones of modern finance are actually practically viable.

For over forty years, the world of finance has been managed like the many dictatorships we see being replaced around the world. A series of eloquent models were developed and it was deemed sacrilegious to question their validity. While the current crisis has cost us dearly, it has been beneficial in that it is now possible to question whether four decades of academic research has resulted in models that can actually work in practice, and to challenge those who have been fighting very hard to maintain the status quo.

A number of articles in this issue of the Journal question the accepted logic of academic finance and submit ideas that financial institutions and regulators could actually use in their work. This edition is consequently of immense value. It sets a new benchmark for the publication of future articles: the need for the ideas presented to be of practical use. It also challenges other publications to position themselves either as proponents of new, more applied ideas, or as bastions of older ideas that have produced a total separation between academic models and real-world relevance.

Most academic publications demand a religious adherence to citing previous articles, many of which were discredited during the recent crisis. In contrast, we at the Journal hope to become the medium of choice for those wishing to challenge the status quo by developing ideas that are based on solid practical perspectives.

That is the aim of this publication going forward. It is an ambitious objective, one that the world of finance demands and that we are confident we can meet, thanks to the outstanding support of our contributing authors. There are now enough financial researchers who share our ambitions to ensure a healthy flow of exceptional articles in practical finance.

We hope that you enjoy this edition of the Journal and that you join our endeavor to narrow the enormous gap between academic finance and what financial executives really need.
Part 1

Economists’ Hubris – The Case of Award Winning Finance Literature

Tracking Problems, Hedge Fund Replication, and Alternative Beta

Empirical Implementation of a 2-Factor Structural Model for Loss-Given-Default

Regulatory Reform: A New Paradigm for Wealth Management

The Map and the Territory: The Shifting Landscape of Banking Risk

Towards Implementation of Capital Adequacy (Pillar 2) Guidelines

The Failure of Financial Econometrics: Estimation of the Hedge Ratio as an Illustration

Systemic Risk Seen from the Perspective of Physics

International Supply Chains as Real Transmission Channels of Financial Shocks
Abstract
In this fifth article in the Economists’ Hubris series, we investigate the practical applications of eight papers that won best-article awards in 2008 and 2009 from the Journal of Finance or the Journal of Financial Economics, the two leading journals in finance. We find that these articles are unlikely to help financial executives improve the way they evaluate risk or manage either risk or their institutions. Finance academics appear to live in a parallel universe, completely oblivious to the nature of the financial services sector that they purport to study. Some of the papers do challenge long-held beliefs, which is very encouraging, but academics still need to go much further than that to write articles that are of any practical value.

1 The views expressed in this paper reflect only those of the authors and are in no way representative of the views of Capco, Contango Capital Advisors, Institute of Management Technology, Dubai, where Shahin Shojai is also a Senior Professor of Finance and Strategic Management, or any of their partners.
This article is the fifth in the Economists’ Hubris series of papers that have so far investigated the shortcomings of academic thinking in the fields of mergers and acquisitions [Shojai (2009)], asset pricing [Shojai and Feiger (2009)], risk management [Shojai and Feiger (2010)], and equity asset management [Shojai et al. (2010)]. In this article, we will focus on the practical contributions of the articles deemed to have made the most significant contributions to the science of finance in 2008 and 2009 by the editors of the Journal of Finance and the Journal of Financial Economics.

We feel that such reality checks are essential if the contributions of academic finance are to migrate from the classroom to the boardroom. Those who are familiar with our previous articles in the Economists’ Hubris series know that we are not very confident about the practical benefits of academic work in this discipline. Students who study medicine are to a large extent able to apply what they learn at medical schools within their roles as doctors. We are not sure that the same applies to finance, the most practically focused of economics disciplines. Yet, we shall do our utmost to be as objective as we can.

The eight papers we are examining were written in an era (2008 and 2009) when complaints about the shortcomings of academic literature were at their loudest. We began with the hope that these papers would be somewhat more practically focused than their peers in the previous years. (We apologize in advance to the editors of financial journals that were neglected by our study, but there is wide agreement that these are the two leading journals of academic finance.)

The articles that have been selected are: Almeida and Philippon (2007), winner of the 2008 Brattle Group Prize in Corporate Finance; Axelson et al. (2008), winner of the 2009 Brattle Group Prize in Corporate Finance; Bargeron et al. (2008), winner of the 2008 Jensen Prizes for the Best Papers Published in the Journal of Financial Economics in the Areas of Corporate Finance and Organizations; Caballero and Krishnamurthy (2008), winner of the 2008 Smith Breeden Prize; Duarte and Young (2009), winner of the 2009 Fama-DFA Prizes for the Best Papers Published in the Journal of Financial Economics in the Areas of Capital Markets and Asset Pricing; Hertzel et al. (2008), winner of the 2008 Fama-DFA Prizes for the Best Papers Published in the Journal of Financial Economics in the Areas of Capital Markets and Asset Pricing; Kondor (2009), winner of the 2009 Fama-DFA Prizes for the Best Papers Published in the Journal of Financial Economics in the Areas of Corporate Finance and Organizations; McLean et al. (2009), winner of the 2009 Smith Breeden Prize; and Pan and Singleton (2008). (We apologize in advance to the editors of financial journals that were neglected by our study; there is wide agreement that these are the two leading journals of academic finance.)

The eight papers we are examining were written in an era (2008 and 2009) when complaints about the shortcomings of academic literature were at their loudest. We began with the hope that these papers would be somewhat more practically focused than their peers in the previous years. (We apologize in advance to the editors of financial journals that were neglected by our study, but there is wide agreement that these are the two leading journals of academic finance.)

The first is that they take account of risk premia when calculating the cost of financial distress. In other words, they calculate the net present value (NPV) of financial distress using “observed credit spreads to back out the market-implied risk-adjusted (or risk-neutral) probabilities of default.” Previous studies have typically used the risk-free rate to discount such costs.

They find that “risk-adjusted probabilities of default and, consequently, the risk-adjusted NPV of distress costs, are considerably larger than historical default probabilities and the non-risk-adjusted NPV of distress, respectively.” For example, the authors find that the risk-adjusted cost of distress for BBB-rated bonds averages 4.5% using their NPV formula, compared to the non-risk-adjusted NPV of distress of only 1.4%.

Their second important contribution is that the tax benefits of leverage are offset by the additional risk-adjusted costs of distress. For example, they find that “using our benchmark assumptions the increase in risk-adjusted distress costs associated with a ratings change from AA to BBB is 2.7% of pre-distress firm value.” “The implied gain in tax benefits as the firm moves from an AA to a BBB rating is 2.67% of firm value. Thus, it is not clear that the firm gains much by increasing leverage from AA to BBB levels. These large estimated distress costs may help explain why many U.S. firms appear to be conservative in their use of debt.”

There is no doubt that taking account of risk in any cost/income calculations is a good thing. However, what is not clear is whether it can in fact be calculated with any degree of accuracy [Shojai and Feiger (2009), (2010)]. Consequently, the main assumption upon which this article is premised is highly questionable. It also requires a huge leap of faith to assume that the rating agencies are able to accurately rate the chances of default, which is essential for many of the calculations in this article to hold. Even if credit risk premia were accurate, they are not stable over time [Berndt et al. (2005), Pan and Singleton (2008)].

Furthermore, if it is true that the tax benefits of additional leverage are offset by the higher cost of borrowing, then the reverse would also hold true. If companies realized that, they would not work so hard to move up the ratings table. They would be very happy with a BBB rating because they would know that their higher cost of borrowing would be offset by lower taxes.

More important, this proves that financial executives have completely ignored previous academic studies in this space: if costs of financial distress had been underestimated, most companies would have borrowed...
to the hilt. The fact that they did not proves that practitioners never even look at studies of this kind.

However, the issue is not so much that the accurate evaluation of risk is almost impossible, or that ratings by ratings agencies have been proven to be anything but suitable for scientific examinations. The main issue is that even if the findings of this study are 100% accurate, which they certainly are not, all it is saying is that after 40 years or so academics are now able to understand why the spread between corporate and government bonds are higher than academic theses suggest. What is the point now able to understand why the spread between corporate and government bonds are higher than academic theses suggest. What is the point of telling bankers what they already know? Not much.

Article 2 – “Why are buyouts levered? The financial structure of private equity funds,” Axelson et al. (2008)

Axelson et al. (2008) focus their attention on the private equity industry. They try to find out why private equity funds have chosen the financial structures they have, and whether that has any impact on their investment choices and performance. “Why is most private equity activity undertaken by funds where LPs commit capital for a number of investments over the fund’s life? Why are the equity investments of these funds complemented by deal-level financing from third parties? Why do GP compensation contracts have the nonlinear incentive structure commonly observed in practice? What should we expect to observe about the relation among industry cycles, bank lending practices, and the prices and returns of private equity investments? Why are booms and busts in the private equity industry so prevalent?”

The authors have nicely thought through the problem of adverse selection when you delegate your investment decisions to others. Indeed, the entire paper could have been better expressed in a couple of pages of clear prose. The ‘mathemtization’ of the paper exemplifies an academic disease in economics, namely attempting to make one’s thoughts seem more significant by expressing them in pompous and useless mathematical formulae. Profound thinkers like Keynes or Friedman did not do such things. The authors have concluded that the structure of private equity payouts and deal funding seems to minimize the decision problems in an agency context. Good, that is what we would have expected in an industry that has been around as long as this one and that is so lightly regulated that all forms of private experimentation are possible. Consequently, similar to the first paper, the authors have discovered that the market works in terms of capintal structure. What a surprise.

Stepping back, the authors have essentially modeled the private equity market as it is. It would have been more useful if they could also have suggested alternative models that might be more efficient. Simply describing in fancy terms what the industry is doing is not really of much use to those who are already managing these businesses.

Trying to identify causes of business models, or financing structures, that are completely different for each entity is very different from publishing an article in the Journal of Neurosurgery about how meningiomas can be debulked using the latest methodologies. The methodologies are developed by people who actually practice them on patients and can be employed by other surgeons all over the world. However, when a group of academics try to dissect a business or financing model, they have no idea whether it will apply to the next private equity fund, nor have they tested their findings with the people who actually manage these businesses.

Moreover, as the authors themselves acknowledge, their analysis is incomplete in important ways. “This paper presents a model of the financial structure of a private equity firm. In the model, a firm can finance its investments either ex ante, by pooling capital across future deals, or ex post, by financing deals when the GP finds out about them. The financial structure chosen is the one that maximizes the value of the fund. Financial structure matters because managers have better information about deal quality than potential investors. Our model suggests that a number of contractual features common to private equity funds arise as ways of partially alleviating these agency problems. However, our model fails to address a number of important features of private equity funds. First, private equity funds tend to be finitely-lived; we provide no rationale for such a finite life. Second, our model does not incorporate the role of general partners’ personal reputations. Undoubtedly these reputations, which provide the ability for GPs to raise future funds, are a very important consideration in private equity investment decisions and a fruitful topic for future research.”

Sadly, despite the fact that nothing in this article is of much relevance to the industry or to its investors or regulators, the paper won the best paper of the year award.

Article 3 – “Why do private acquirers pay so little compared to public acquirers?” Bargeron et al. (2008)

Bargeron et al. (2008) compare the target shareholder wealth gains of acquisitions made by public firms with those made by private firms and find that the difference in premiums between these two types of acquisitions is sizeable and significant. The authors find that the average gain for target shareholders when the bidder is a public firm is 31.74% over...
the three days surrounding the announcement of the acquisition, 22.20% when the acquirer is a private firm, and 20.47% when it is a private equity fund. When they try to determine the causes for such differences, the authors find that the differences can be explained neither by synergistic reasons (around 40% of the latter group are also operating companies and can similarly benefit from synergies), nor by the specific characteristics of the target or the deal. Finally, they find that management of target firms are unlikely to sell the company cheaply to private owners, simply due to potentially more lucrative post-acquisition contracts, because institutional shareholders will prevent them from doing so. Moreover, if they have a large ownership pre-acquisition, they would lose out on the premium on the shares sold to the bidding firm.

What the authors assert from their research is that where agency costs are high, managers pay over and above what the target is actually worth simply for self-aggrandizing reasons. As a result, the gains to a seller seem to be much greater when a public firm is making the acquisition. In support of their hypothesis, they find that as managerial ownership of the public bidder increases, the gap between target shareholder gains of public and private acquisitions decreases. When managerial ownership exceeds 20%, the authors find no difference between shareholder gains of public and private acquisitions. They also find that private bidders are much more likely to walk away from bids than public firms, with around 36% of offers made by private firms withdrawn as compared to around 14% for public bidders.

It is astonishing that the authors believe that the financial markets are so analytically efficient that they are able to work out within a day or so of an announcement that simply because the management of the bidding firm has a large ownership stake in the public company it manages or because it is a private buyer, it will not be willing to overpay for the target. This is the Efficient Market Hypothesis par excellence. We doubt that even Eugene Fama believes that the markets are so incredibly efficient.

There can be many reasons why public firms might seem to be paying more than what private buyers might be willing to pay. It could be that they are involved in larger deals, which the authors also find to be the case. These deals attract greater attention and are more likely to result in a more drawn-out contest with other bidders. It could also be that the bidder is trying to protect market share, which also results in a more contested environment. The authors also find that the private firms undertake more diversifying transactions than their publicly quoted peers. Finally, the authors find that the targets of the private firms are more likely to have been underperforming, which means that they are more likely to benefit from improved management. Given that these targets are also found to have greater operating cash flows, they make excellent choices for management buy-ins [Shojai (2004)]. However, if the markets expect the benefits from the acquisition of badly performing firms to be greater, then the abnormal returns should be greater for these transactions: as is the case when buy-outs (transactions involving incumbent management) are compared with buy-ins (which involve a replacement of the management by a group that is expected to improve the management of the company’s assets). The reason that this is not found in the article by Bargeron et al. (2008) is probably due to the fact that the private deals are smaller and have remained under the radar. The buy-in of RJR Nabisco by KKR proves that when the deals do become public and contested, even if the owners are private, the premiums can become astronomical.

Last, and by no means least, the single most damaging contribution to the field of finance has been the advent of event-study methodologies. The presumption that one can find answers to highly complex and uniquely different questions through an aggregated regression analysis is beyond bizarre. The tremendous reliance that academics place on the reliability of these methodologies has meant that they no longer spend the time looking for genuine answers to tough questions. They allow the data to find a range of potential responses and then try very hard to find explanations that would fit those findings.

Event-study methodologies are not remotely as reliable as academics think they are. Making small changes here and there, such as the use of non-parametric data, management of heteroscedasticity, or application of non-linear regressions, will not solve the overall problem. These methodologies are a useful tool but cannot be viewed as the source of indisputable facts. So-called financial scholars have used the same dataset and obtained different results, simply by changing a small part of the methodology. When different datasets are used, well, then the whole thing falls to pieces. Shojai (2009) presents a small sample of studies that have used event-study methodologies to investigate the sources of gains from mergers and acquisition transactions. The results make for very interesting reading.

As Shojai (2009) states: “a question that few of the experts who have written on this subject have asked themselves is what value their research has to corporate executives. Just how much credence would these executives give to knowing that Dodd and Ruback (1977) or Jarrell and Poulsen (1989) find that target shareholders in the U.S. on average experience returns of around 21.2% and 28.9%, respectively, during acquisitions or that Franks and Harris (1989) find that they make 23.3% in the U.K. or that Husson (1988) finds that they make 36.7% in France? Would that in anyway impact the premium that Company A pays to Company B? Given that the economic environment is different for every deal,

---

3 It should be added that event study methodologies have also been applied within the article, and the reader is cautioned about the causalities mentioned therein, similar to cautions found in other papers that use such techniques.
the availability of capital [Franks et al. (1988)] and level of competition is different, and the industry would be at different stages of its maturity, the premium could be very different.

Essentially, aggregate data cannot take into account the implications of Company A’s acquisition of Company B on the long-term state of play in that specific sector, and if the sector is important enough in the entire index that was used as the proxy for the market. This interconnection of factors and environments is completely overlooked by event-study approaches. For example, it is almost impossible to determine how strategically a given acquisition has impacted a given competitor or competitors. If you now add the personal attributes of the executives involved, you can appreciate such data become completely useless. For example, is the CEO of Company A expected to pay the same premium in 2009, when credit has pretty much dried up, that he/she paid in 2006, when private equity firms where swallowing up most of the major corporations? Should the premium be the same when markets have gone mad and economists’ efficient markets are nowhere to be found as it would be when the global economy is on the edge of the precipice? Should he offer cash, as Franks et al. (1988) suggest, if he is the CEO of Yahoo! in 1999, at a time when his company’s shares are significantly more king than cash? For whose benefit exactly are these numbers compiled and all this effort spent to update old numbers with new data and methodology? A scientific study would first create comparable situations, and only then examine any differences between public and private buyers.


Caballero and Krishnamurthy (2008) investigate the potential benefits of central bank interventions during flight-to-quality episodes, which are triggered by severe unexpected events, using two models: capital/liquidity shortages and Knightian uncertainty [Knight (1921)], with the latter receiving a lot of attention from the practitioners and less from the academicians.

The authors suggest that while crises caused by liquidity shortages and Knightian uncertainty might have some similarities, they do not always result in similar outcomes. Consequently, depending on which is the cause of the crisis, if not both together, the repercussions and the need for central bank intervention would differ. Where the crisis is isolated and does not result in a precipitous uncertainty in other market participants, there should be no need for central bank intervention.

An important distinction is raised between situations of liquidity shortages and situations of Knightian uncertainty: moral hazard. Whereas in the former private and public insurance are substitutes, in the latter they are complementary. That is why in the liquidity-crisis model, ex-ante policy recommendations typically focus on prudential risk management, such as regulations to reduce leverage, increase liquidity ratios, or tighten capital requirements. In the case of Knightian uncertainty, however, since there are no precedents for what is taking place, neither the public nor private institutions can take precautionary actions that would prevent the crisis. In such cases, the authors suggest that transparency and collaboration are essential. If parties faced with the crisis are open about the extent of their exposures and collaborate with the support of the central bank, in its role as the lender of last resort, then the crisis might be contained somewhat.

Reading through the many interesting and impressive mathematical models in this article, one cannot but feel some comfort about our abilities to react to future crises in what is being presented. That is until one tries to imagine how it would be of help in a real crisis scenario. The authors are right: because major crises have no precedents, it is almost impossible to prepare for them. You cannot use knowledge and data from previous crises to aver future crises. Having a mathematical model that illustrates this is of little use. Anyone who has lived through crises such as the bursting of the Internet bubble or the Russian debt crisis is fully aware that public policy responses cannot be predicted with any degree of accuracy until the crisis flares up. More important, no one knows the implications of such responses until the dust has settled. The recent crisis, which seems to have taken hold just after this article was published, has demonstrated the limits of our understanding. Different central banks across the world undertook different approaches that resulted in different outcomes. The reason for the differences was partly because they had different views on how resilient their domestic/regional banking systems and economies were. Some were lucky to find that a good number of banks within their jurisdictions were better capitalized than their peers across the Atlantic. However some were less confident in the ability of financial mathematicians to accurately price the risk of complex assets. They actually questioned the contributions of academic finance.

Notwithstanding these niceties, there is absolutely nothing in this paper that was not available to a reader of the Financial Times, which expressed its ideas in far more lucid prose. Simply stating the obvious does not make a contribution, no matter how fascinating the mathematics employed to state it.

Digging deeper, this paper is a very good fit for the joke about looking for your lost keys under the streetlight rather than where you dropped them, because that’s where the light is. Their argument is essentially that a situation where everyone assumes the worst outcome and acts accordingly represents market irrationality because not everyone can simultaneously be as badly off as in the worst-case scenario. Consequently, a central bank acting as lender of last resort restores ‘rationality’ to the situation by compensating for these irrational fears of the individual,
private participants. Why is this? Because their model of risk/uncertainty assumes that as the premise. Their model assumes that there is a first shock and then subsequent ones of decreasing likelihood. But of course that is not what happened between late 2007 and the beginning of 2009 – things got progressively worse in ways that were not anticipated and, if you look at what is happening to the euro and to China, the uncertainty continues to accumulate. In essence, market participants have come to agree with Donald Rumsfeld that what you need to fear is not the ‘known unknowns’ but the ‘unknown unknowns.’ Who knew that all these leveraged off-balance-sheet structures were out there? Who had any idea about how governments and central banks would react? Who could predict how obstinate trade unions would be in Greece or how willing German taxpayers would be to bail out the Irish? Who knew that these questions were even relevant?

And, to top it all off, as with most academic articles, this article concludes with a more difficult question than the one it tries to answer. The main question that the authors hope to answer is mentioned in the closing paragraph of the conclusion: “Finally, as we note, Knightian uncertainty may often be associated with financial innovations. This suggests that crises surrounding financial innovations may be fertile ground to look empirically for the effects we have modeled, and disentangle them from other more well-understood effects. It also suggests a new perspective on the costs and benefits of innovation. For example, our model suggests that in a dynamic context with endogenous financial innovation, it is the pace of this innovation that inherently creates uncertainty and hence the potential for a flight to quality episode. Financial innovation facilitates risk sharing and leverage, but also introduces sources of uncertainty about the resilience of the new system to large shocks. This uncertainty is only resolved once the system has been tested by a crisis or near-crisis, after which the economy may enjoy the full benefits of the innovation. We are currently exploring these issues.”

Nice to hear they are working on it.

**Article 5 – “Why is PIN priced?” Duarte and Young (2009)**

Duarte and Young (2009) look at why investors demand to be compensated by higher returns from shares in which informational asymmetry is higher. Some studies find that this informational asymmetry is diversifiable while others find that it is not. For example, Easley et al. (2002) find that a ten percent difference in the PINs (probability of informed trading) of two stocks results in a 250-basis-point difference in their annual expected returns.

The authors also find that, assuming we are able to identify periods of private information via abnormal order flow imbalances motivated by sequential trade models, information-based trading does not affect expected stock returns. Rather, it is the microstructure and liquidity effects unrelated to information asymmetry that influence expected returns.

We were a bit befuddled by the logic of their work. We let the authors’ own words explain why: “It is worth noting that this interpretation relies on the assumption that periods of asymmetric information can be identified as periods with abnormal order flow imbalances. It is possible the relation between private information and order flow is more complex than the one implied by sequential trade models, in which case private information could indeed be related to expected returns. However, in this case, both PIN and adjPIN would be inappropriate proxies for information asymmetry.”

The assumption that trading patterns can be used to discern asymmetric information, which somehow people could access and benefit from, is actually a bold assertion that should be validated rather than taken as given. How would a private investor know if it was facing a more severe form of informational asymmetry in Company A than Company B?

So far, levels of asymmetry cannot be tested. However, we have learned that facts cannot be allowed to prevent an entire area of literature from being developed.

**Article 6 – “Inter-firm linkages and the wealth effects of financial distress along the supply chain,” Hertzel et al. (2008)**

Hertzel et al. (2008) investigate the implications of financial distress of companies, before and after filing for bankruptcy, on their customers, suppliers, and industry rivals. The question is an important one because few can deny that firms undertake certain actions prior to full-blown bankruptcy filings that can have unintended consequences, sometimes to the detriment of their shareholders.4 More importantly, suppliers and customers do change the way they deal with a firm that is deemed to be experiencing financial difficulties, such as shortening credit terms, reducing the amount of credit provided, or delaying payments. The value of the firm’s after-sale service and warranties is also severely affected when the firm is perceived to be experiencing financial difficulties.

The authors find significant pre-filing and filing-date contagion effects, which impact not only industry rivals but also suppliers of the firm. They do not find, however, a significant contagion effect between the firm and its customers, which they attribute to the greater likelihood that customers can anticipate, or even cause, the firm to experience financial distress.

---

4 When the management is facing the real prospect of corporate bankruptcy, it might be tempted to undertake riskier endeavors since it has an option on the company. If the option pays off, management will have saved the company and survived; if it does not, the only loser are the shareholders and debt holders.
They also find that supplier contagion effects are more severe when the filing firm’s industry also suffers contagion, which is attributed to the difficulties that suppliers face in switching to other customers. When the announcement of the filing causes the share price of its competitors to increase, or at least not fall, the impact on suppliers and customers is insignificant. The authors attribute this to shifts in market share and not increased market power. When the bias of the sample firms that are reliant on the failing company is removed, the authors find that “contagion effects spread beyond reliant suppliers and major customers to firms in their respective industries.”

Investigations into whether various cross-sectional characteristics of the filing firms, customers, and suppliers affect the returns of customers and suppliers did not yield significant findings.

Well, the findings of this study would have been of huge importance had it not been possible to arrive at the same conclusion with a course in strategic management 101. Almost anyone who has ever studied management, or been involved in managing a company, knows that when a company goes bankrupt it will have an impact on its suppliers and customers. The fact that the authors of this study did not find any impact on the customers is the main surprise of this study. And to attribute that finding to the notion that customers can somehow foresee the firm’s bankruptcy while its suppliers cannot beggars belief.

Given the competitive environment we are in, it would not be hard for firms to find replacement customers and suppliers, but of course that takes time.

And, while we do not have the highest regard for the skills of investment analysts, we are sure that even they can work out the implications that a firm facing financial difficulties would have on its suppliers and customers, the international nature of its trading partners, how easy or difficult it would be for the firm to be replaced, whether it will be allowed to face full-blown bankruptcy and the implications thereof, etc.

It is also stating the blimin obvious when the authors suggest that in situations where the supplier/customer is not easily replaceable, the implications are greater. Anyone who has been watching the news in recent years is familiar with the case of Delphi, the car parts maker, and the risks it would have faced had GM been allowed to fail. When GM was facing financial difficulties and it was looking as though the end was nigh, the management of Ford Motor Company came out in support of GM and stated that since most of the major U.S. car manufacturers rely on Delphi for a good number of their parts, it must not be allowed to fail. If GM had been allowed to fail, there would have been a good chance that Delphi, formerly a division of GM, would have also failed, and Ford would have had a difficult time finding suitable replacements in a short period of time.

They worked this out without academic help.


Kondor (2009) investigates the risks that arbitrageurs with limited capital resources face when speculating on the convergence of prices of similar assets if, because of limited resources, they are forced to unwind their positions if the prices diverged. Kondor suggests that because risk-neutral arbitrageurs have to decide how to allocate their limited capital across uncertain future arbitrage opportunities, when arbitrage opportunities arise – predominantly due to temporary pressure on local demand curves of two very similar assets traded in segmented markets – their actions, along with the uncertain duration of the local demand pressure, will determine the future distribution of the price gap between the two assets. He finds that mere action of the arbitrageurs will result in potential losses because their individually optimal strategies will increase price gaps.

This article actually makes a very clever point that does, indeed, dispel some naive assumptions about the beneficial effects of arbitrage on price behavior. If you took away all the equations and kept the verbal discussion, it would be remarkably lucid. Essentially, it says that arbitrage opportunities attract traders, but that these traders face a difficult and totally realistic decision. They have to collateralize their trades, one side of which is a short, and there is a net cost of carry of the short. So, they have to decide when to make the bet and when to pull the bet because they are running out of money. Because the opportunity widens and narrows randomly, they can end up investing too early and running out of money. So, their demand and supply comes and goes from the market, and the price can move a lot before the arbitrage is finally eliminated. There are no ‘sure things’ to be found – only bets.

This is exactly what happened in the downward spiral that started in early 2008 where, due to the collapse of liquidity, what were essentially arbitrage bets failed in staggering numbers because all the collateral was used up. That is true, and even more important, it is not what is taught in economics textbooks. If Kondor’s argument were put in plain language it could be a chapter in a textbook. As it is, it will serve primarily as a cure for insomnia.

**Article 8 – “Share issuance and cross-sectional returns: international evidence,” McLean et al. (2009)**

McLean et al. (2009) basically apply a number of tests undertaken in the U.S. market to an international setting to investigate whether issuance effect is also present among non-U.S. firms and to determine whether the cross-country differences can be explained.

Their findings are similar to the studies of U.S. firms, though not of the
same magnitude. They find that issuance predictability is more statistically significant than either size or momentum, and is of the same magnitude as book-to-market. They also find that the issuance effect is robust across both small and large firms.

The authors also find that the issuance effect is stronger in countries where it is cheaper to issue and repurchase shares. In more developed markets, where issuance costs are lower, firms are able to issue shares frequently to take advantage of either market mispricings or changes in exposure to priced risk. In less developed markets, where share issuance is more costly, the benefits of market timing are exceeded by issuance costs, and share issuance is undertaken predominantly to take advantage of market timings. Overall, they find that share “issuance will be both more frequent, and more highly correlated with future returns in well-developed markets.”

Of immediate interest is that the findings of this paper are different from those of so many other studies that have focused on specific markets rather than comparing across markets. We might note that many of the other studies do not corroborate each other’s results either. However, we are prepared to accept the statistical work. What does it mean?

The only useful contribution of this study is that it argues, in effect, that the notion of an ‘efficient market’ giving the ‘true price’ is a load of nonsense. Companies and their advisers know when the mob is overpricing their company and that is when they choose to issue new stock, thereby using the new buyers to subsidize the existing owners with cheap capital. Where capital markets are more fluid, it is easier for companies to do this and they do it more. Shame on them, really.

Conclusion

In this fifth article in the Economists’ Hubris series, we have investigated the practical applications of eight articles that won best article awards for 2008 and 2009 from the Journal of Finance or the Journal of Financial Economics, the two leading journals in the field.

We find that these articles provide little to help financial executives improve the way they evaluate risk, or manage risk or their institutions. Finance academics seem to live in a parallel universe completely oblivious to the needs of the financial services sector. Some of the papers do challenge long-held beliefs of academic finance, which is very good to see, but none provides anything that can be of practical use to market participants.

We sympathize with the editors of these two journals because we at the Journal of Financial Transformation are also constantly struggling with the difficulties of attracting articles, or persuading academics to write papers, that are of practical benefit to financial executives.

But, what we believe is different in the case of these articles is that they have won prizes for their contributions despite making no genuine effort to be practical. We believe that the academics got so engulfed in looking at the methodologies and applying impressive mathematical or statistical models that they forgot that the point was to provide information that would benefit financial executives. In many cases, and proven by the huge variations in results of studies of the same subject matter, it simply comes down to a beautiful model applied to useless data.

We argue that the gap between academic and practical finance remains as large today as it has ever been. Financial economists probably think practical finance is too boring and irrelevant to their objectives to merit much attention. It probably is very hard to make a genuine contribution that can be applied in practice. It is far easier to write about an imaginary world in which it is possible to improve air quality by separating carbon from oxygen than to devise a tool that actually does that. We believe that the gap between academic and practical finance is no less severe.

The recent crisis has given academic finance a remarkable opportunity to throw away many of the models that were until recently viewed as gospel, and many of the useless publications that simply publish new ways of analyzing the same old useless data and models. Instead, we could develop a smaller number of genuinely useful publications that work in collaboration with those who actually apply these models.

Now is the perfect time to write articles that financial executives can actually use and understand, so that they can be involved in the refereeing process. What is the harm in having financial executives assess the genuine viability of a new idea? Finance academics are petrified of such an examination since they know that more than 40 years of so-called scholarly thinking will have to be discarded and they will have to start afresh. But isn’t that what all good industries do when they realize the old models simply do not work anymore? It is time that academic finance also does the same. Throw out all the discredited theories, stop requiring that future generations of researchers cite them as gospel, and force academics to give their ideas a much-needed reality check.

Our hope is that this article will get at least some of the more influential academic institutions to start questioning the practical validity of some of the issues their researchers work on. There is no harm in having lots of theoretical publications on the side, but the world of finance requires that theoretical finance laboratories be segregated from financial think-tanks, and that the latter start behaving more like the genuine scientists they can and should be.
References

- Knight, F., 1921, Risk, uncertainty and profit, Houghton Mifflin, Boston
Tracking Problems, Hedge Fund Replication, and Alternative Beta

Thierry Roncalli — Professor of Finance, University of Evry, and Head of Research and Development, Lyxor Asset Management

Guillaume Weisang — Doctoral Candidate, Bentley University

Abstract

As hedge fund replication based on factor models has encountered growing interest among professionals and academics, and despite the launch of numerous products (indexes and mutual funds) in the past year, it has faced many critics. In this paper, we consider two of the main critiques, namely the lack of reactivity of hedge fund replication, its deficiency in capturing tactical allocations, and the lack of access to the alpha of hedge funds. To address these problems, we consider hedge fund replication as a general tracking problem which may be solved by means of Bayesian filters. Using the example provided by Roncalli and Teiletche (2008), we detail how the Kalman filter tracks changes in exposures, and show that it provides a replication methodology with a satisfying economic interpretation. Finally, we address the problem of accessing the pure alpha by proposing a core/satellite approach of alternative investments between high-liquid alternative beta and less liquid investments. Non-normality and non-linearities documented on hedge fund returns are investigated using the same framework in a companion paper [Roncalli and Weisang (2009)].
Over the past decade, hedge fund replication has encountered a growing interest both from an academic and a practitioner perspective. Recently, Della Casa et al. (2008) reported the results of an industry survey showing that, even though only 7% of the surveyed institutions had invested in hedge fund replication products in 2007, three times as many were considering investing in 2008. Despite this surge in interest, the practice still faces many critics. If the launch of numerous products (indexes and mutual funds) by several investment banks in the past year can be taken as proof of the attraction of the ‘clones’ of hedge funds (HF) as investment vehicles, there remain nonetheless several shortcomings which need to be addressed. For instance, according to the same survey, 13% of the potential investors do not invest because they do not believe that replicating hedge funds’ returns was possible; 16% deplore the lack of track record of the products; another 16% consider the products as black boxes. Finally, 25% of the same investors do not invest for a lack of understanding of the methodologies employed, while 31% of them were not interested for they see the practice as only replicating an average performance, thus failing to give access to one of the main attractive features of investing in one hedge fund, namely its strategy of management.

As a whole, the reasons put forward by these institutions compound different fundamental questions left unanswered by the literature. Since the seminal work of Fung and Hsieh (1997), most of the literature [Agarwal and Naik (2000), Amenc et al. (2003, 2007), Fung and Hsieh (2001), inter alia] has focused on assessing and explaining the characteristics of HF returns in terms of their (possibly time-varying) exposures to some underlying factors. Using linear factor models, these authors report the incremental progress in the explanatory power of the different models proposed. Yet, for now, the standard rolling-windows OLS regression methodology, used to capture the dynamic exposures of the underlying HF’s portfolio, has failed to show consistent out-of-sample results, stressing the difficulty of capturing the tactical asset allocation (TAA) of HF’s managers. More recently, more advanced methodologies, in particular Markov-Switching models and Kalman Filter (KF), have been introduced [Amenc et al. (2008), Roncalli and Teiletche (2008)] and show superior results to the standard rolling-windows OLS approach. From the point of view of investors, however, the complexity of these algorithms certainly does not alleviate the lack of understanding in the replication procedure. Furthermore, despite superior dynamic procedures and an ever expanding set of explanatory factors, some nonlinear features of HF returns [Diez de los Rios and Garcia (2008)] as well as a substantial part of their performance remain unexplainable, unless surmising ultrahigh frequency trading and investments in illiquid assets or in derivative instruments by HF managers. To our knowledge, while commonly accepted by most authors, because of practical difficulties, these explanations have not led to a systematic assessment nor have they been subject to systematic replication procedures. In this paper, we address two of the main critiques formulated on hedge fund replication. First, using the notion of tracking problems and Bayesian filters and their associated algorithms, we address the alleged failure of HF replication to capture the tactical allocations of the HF industry. Using the linear Gaussian model as a basis for the discussion, we provide the readers with an intuition for the inner tenets of the Kalman Filter. We illustrate how one can obtain sensible results, in terms of alternative betas. Second, we address the problem of accessing the part of the HF performances attributed to uncaptured dynamic strategies or investments in illiquid assets, i.e., the alpha of HF.

Framework

Although HF replication is at the core of this paper, we would like to inscribe our contribution in a larger framework, albeit limited to a few financial perspectives. Thus, after a description of HF replication, this section introduces the notion of tracking problems. After a brief and succinct formal definition, we show how this construct indeed underpins many different practices in finance, including some hedge fund replication techniques and some investment strategies such as, for example, Global Tactical Asset Allocation (henceforth, GTAA). It is armed with this construct and the tools associated to it that we tackle three of the main critiques heard in the context of hedge fund replication in subsequent sections.

Hedge fund replication

Rationale behind HF replication

Even though HF returns’ characteristics make them an attractive investment, investing in hedge funds is limited for many investors due to regulatory or minimum size constraints, in particular for retail and institutional investors. Hedge funds as an investment vehicle have also suffered from several criticisms: lack of transparency of the management’s strategy, making it difficult to conduct risk assessment for investors; poor liquidity, particularly relevant in periods of stress; and the problem of a fair pricing of their management fees. It is probably the declining average performance of the hedge fund industry coupled with a number of interrogations into the levels of fees [Fung and Hsieh (2007)] which led many major investors to seek means of capturing hedge fund investments strategies and performance without investing directly into these alternative investment vehicles [Amenc et al. (2007)]. Hence, the idea of replicating hedge funds’ portfolios, already common in the context of equity portfolios, gained momentum.

Factor models2

Starting with the work of Fung and Hsieh (1997) as an extension of Sharpe’s style regression analysis [Sharpe (1992)] to the world of hedge funds, factor-based models were first introduced as tools for performance analysis.

2 With the growing interest in hedge fund replication over the last decade, it is not surprising to find that there exists a rich literature which is almost impossible to cover extensively. A comparison of the factor and the pay-off distribution approaches can be found in Amenc et al. (2007). We also refer the interested reader to Amin and Kat (2003), Kat (2007), or Kat and Palaro (2008) for a more detailed account of the pay-off distribution approach.
The underlying assumption of Sharpe’s style regression is that there exists, as in standard Arbitrage Pricing Theory (APT), a return-based style (RBS) factor structure for the returns of all the assets that compose the investment world of the fund’s manager [Fung and Hsieh (1997), Sharpe (1992)]. Factor-based models for hedge fund replication make a similar assumption but use asset-based style (ABS) factors. While RBS factors describe risk factors and are used to assess performance, ABS factors are directly selected with the purpose of being directly transposable into investment strategies. ABS factors have been used to take into account dynamic trading strategies with possibly nonlinear pay-off profiles [Agarwal and Naik (2000), Fung and Hsieh (2001)]. The idea of replicating a hedge fund’s portfolio is therefore to take long and short positions in a set of ABS factors suitably selected so as to minimize the error with respect to the individual hedge fund or the hedge fund index.

A generic procedure for HF replication using factor models [Agarwal and Naik (2000), Fung and Hsieh (2001), Sharpe (1992)] can be decomposed in two steps. In step 1, one estimates a model of the HF returns as $r_k = \sum_{i=1}^{m} w_i r^{(i)}_k + \epsilon_k$. Given the estimated positions $\hat{w}(i)$ (on the ABS factor $r(i)$) resulting from step 1, step 2 simply constructs the ‘clone’ of the hedge fund by $\sum_{i=1}^{m} \hat{w}(i) r^{(i)}_k$. The factor-based approach is thus very intuitive and natural. There are, however, several caveats to this exercise. Contrary to the passive replication of equity indices, the replication of hedge funds returns must take into account key unobservable determinants of hedge fund investment strategies such as the returns from the assets in the manager’s portfolio; dynamic trading strategies; or the use of leverage [Fung and Hsieh (1997, 2001)].

Fairly recently, attempts to capture the dynamic nature of the HF portfolio allocation have been explored in the literature in order to improve the in-sample explanatory power and the quality of the out-of-sample replication. One method, used extensively [Fung and Hsieh (2004), Hasanhodzic and Lo (2007), Jaeger (2008), Lo (2008), inter alia], is to use rolling-windows OLS where the coefficients $\hat{w}(i)$ are estimated by running the OLS regressions of $r^{(i)}_k$ on the set of factors $r^{(j)}_k$ for $i = 1, \ldots, m$. A common choice for the window length L is 24 months, even though one could consider a longer time-span trading-off the dynamic character of the coefficients for more stable and more robust estimates. By means of an example, Roncalli and Teiletche (2008) have demonstrated that such a methodology poorly captures the dynamic allocation in comparison with the Kalman filter (KF). The use of KF estimation, however, requires caution in its implementation, making the estimation of the positions $\hat{w}(i)$ a non-trivial matter. Markov regime-switching models have also been considered [Amenc et al. (2008)]. The idea therein is that HF managers switch from one type of portfolio exposure to another depending on some state of the world, assumed to be discrete in nature. One possible interpretation is to consider that the active management consists of changing the asset allocation depending on two states of the economy (high and low). Justifying the number of states or their interpretation is, however, tricky.

**Definition of the tracking problem**

We follow Arulampalam et al. (2002) and Ristic et al. (2004) in their definition of the general tracking problem. We note $x_k \in \mathbb{R}^N$ the vector of states and $z_k \in \mathbb{R}^M$ the measurement vector at time index $k$. In our setting, we assume that the evolution of $x_k$ is given by a first-order Markov model $x_k = f(t_k, x_{k-1}, \nu_k)$, where $f$ is a (non-)linear function and $\nu_k$ a noise process. In general, the state $x_k$ is not observed directly, but partially through the measurement vector $z_k$. It is further assumed that the measurement vector is linked to the target state vector through the following measurement equation $z_k = h(t_k, x_k, \eta_k)$, where $h$ is a (non-)linear function, and $\eta_k$ is a second noise process independent from $\nu_k$. Our goal is thus to estimate $x_k$ from the set of all available measurements $z_{1:k} = \{z_i, i=1,\ldots,k\}$.

**Remark 1** – In the rest of the paper, a system in the following format will be referred to as a tracking problem (henceforth TP) $\{x_k = f(t_k, x_{k-1}, \nu_k); z_k = h(t_k, x_k, \eta_k)\}$ (1)

**Link between GTAA, HF replication, and tracking problems**

The two problems of replicating a global tactical asset allocation (GTAA) strategy and HF replication can be seen as belonging to the same class of approaches. For the clarity of our exposition, we decompose the return of a hedge fund into two components

$$r_k^{(HF)} = \sum_{l=1}^{m} w_l^{(HF)} r_l^{(HF)} + \sum_{l=m+1}^{n} w_l^{(HF)} r_l^{(HF)}.$$

GTAA is an investment strategy that attempts to exploit short-term market inefficiencies by establishing positions in an assortment of markets with a goal to profit from relative movements across those markets. This top-down strategy focuses on general movements in the market rather than on performance of individual securities. Beside GTAA, hedge fund managers may invest in a larger universe. A part of the universe is composed of the asset classes found in GTAA strategy and another part of the universe is composed of other alternative asset classes and strategies, such as stock picking strategies (which may be found in equity market neutral, long/short event driven hedge funds), high frequency trading, non-linear exposures using derivatives, and illiquid assets (corresponding to distressed securities, real estate or private equity).
The idea of HF replication, in particular to create investment vehicles, is to replicate the first term on the RHS of (2). If we note $n_k = \sum_{p=m+1}^{n_k} w_k^{(p)} n_k^{(p)}$, then HF replication can be described as a TP $w_k = W_{k-1} + \eta_k; t_k^{(HF)} \Rightarrow n_k = n_k^{(k)} + \eta_k$ (3). We must, however, stress two points before continuing. First, HF replication will work best at the industry level using aggregates of hedge funds’ performances as the replication benchmark. Díez de los Ríos and García (2008) report a large proportion of the HF industry to be following long/short equity strategies (about 30%)3. The performance of a single HF following an L/S equity strategy is explained by its proprietary model of stock picking and its proprietary model to choose its beta, such that its portfolio will be long of a 100% of the selected stocks, and short of x% of its benchmark index. It is almost impossible to determine without inside information the portfolio of stocks picked by the HF manager as it depends on its targeted risk profile and the private views of the managers. However, because of the efficiency of liquid markets, as an aggregate, the performance of all the L/S equity HF will be proportional to $1-x$, where $x$ is the average taken over all L/S equity funds of their exposure. In other words, the performance of the aggregate will be proportional to the beta of the entire industry, and the idiosyncratic decisions of each manager are averaged out. It is worth noting that in this case, as the underlying asset classes are standard, replicating an aggregate of L/S equity HF is about the similar to replicating a GTAA strategy. This point is all the more salient since other HF strategies are not represented in a proportion equivalent to the L/S equity HF [Fung and Hsieh (2004)].

Seemingly, one weakness of the approach we propose is that only the beta of HF strategies seems to matter. One could rightly argue, however, that an attractive feature of investing in single HF is the promise of absolute performance. Even in the case of L/S equity strategies, Fung and Hsieh (2004) further argued that they produce ‘portable’ absolute overperformances, which they termed ‘alternative alphas,’ that are not sensitive to traditional asset classes. We contend nonetheless, as our decomposition above between GTAA ABS factors and HF ABS factors hinted at, that one must be realistic between what can and cannot be replicated. If HF performances can be divided between a beta component and a non-replicable alpha component, it is because HF managers engage in trading at high-frequencies or in illiquid assets, thus benefiting from local and transient market inefficiencies or illiquidity premia. Moreover, if considering these typical HF ABS factors is very useful in explaining the performance of the HF industry, these items cannot in good measure be replicated from an investment perspective. Thus, we already need to point out that not all of the HF strategies can be successfully replicated using the method we advocate in this paper. This is perhaps the one good news for the HF industry. Even though we will demonstrate one can truly capture a substantial part of the performance of the industry as a whole, still they individually retain some edge, particularly those practicing true alternative strategies. The next sections expose and provide the tools to capture the tactical allocation of a manager’s portfolio.

Capturing tactical allocation with Bayesian filters

The prior density of the state vector at time k is given by the following equation $p(x_k | z_{1:k-1}) = p(x_k | x_{1:k-1})p(x_{k-1} | z_{1:k-1}) dx_{k-1}$, where we used the fact that our model is a first-order Markov model to write $p(x_k | x_{1:k-1}, z_{1:k-1}) = p(x_k | x_{k-1})$. This equation is known as the Bayes prediction step. It gives an estimate of the probability density function of $x_k$ given all available information until $t_{k-1}$. At time $t_k$, a new measurement value $z_k$ becomes available, one can update the probability density of $x_k$: $p(x_k | z_{1:k}) \propto p(z_k | x_k)p(x_k | z_{1:k-1})$ (5). This equation is known as the Bayes update step. The Bayesian filter corresponds to the system of the two recursive equations (4) and (5). In order to initialize the recurrence algorithm, we assume the probability distribution of the initial state vector $p(x_0)$ to be known.

Using Bayesian filters, we do not only derive the probability distributions $p(x_k | z_{1:k-1})$ and $p(x_k | z_{1:k})$, but we may also compute the best estimates $\hat{x}_{k|k}$ and $\hat{x}_{k|k-1}$ which are given by $\hat{x}_{k|k} = \mathbb{E}[x_k | z_{1:k}] = x_k p(x_k | z_{1:k}) dx_k$ and $\hat{x}_{k|k-1} = \mathbb{E}[x_k | z_{1:k-1}] = x_k p(x_k | z_{1:k-1}) dx_k$. To gain better understanding of the advantages of using the tracking problem’s formalization as well as Bayesian filters to answer the problem at hand, we examine here HF replications in a Gaussian linear framework using KF. In a companion paper, we also considered the use of particle filters to allow for more flexible specification of the density function [Roncalli and Weisang (2009)].

Hedge fund replication: the Gaussian linear case

In this section, in order to substantiate our claim that the tactical asset allocation of a portfolio is retrievable, we start by providing an intuition of the inner workings of the KF algorithm. We also test, with the aid of an example, the capacity of KF to determine plausible weights for a replicating portfolio of a standard HF index. Furthermore, we show that the replicating portfolio provides a qualitatively sensible explanation for the behavior of the HFRI index over the period 1994-2008, while enabling us to capture a significant part of its performance.4 Finally, we look into the types of strategies that one could consider when replicating in the HF industry.

Understanding linear Gaussian approach and Bayesian filtering to replication strategies

While the KF algorithm described in appendix is well known to many engineers and econometricians, the classic contemporaneous representation (A1) provides little insight into how KF dynamically modifies the estimated weights to track the exposures of the portfolio as described

---

3 Fung and Hsieh (2004) report further that in March 2003 about 40% of the HFs reported in the TASS database list long/short equity as their primary investment style. There are historical reasons for that. L/S equity strategy was the strategy used by the first HF on record, created in 1949 by A.W. Jones.

4 To be more precise, the study period for all the computations done in the rest of this paper begins in January 1994 and ends in September 2008.
in TP (3). In the following, using the innovations representation of the KF algorithm, we explain with finer details the dynamic adjustments of the recursion.

Innovation representation of linear state-space models
The dynamic described by the equations (A1) can be re-written in terms of the tracking error $\hat{e}_k$. It suffices to recombine (A1) into $\tilde{x}_{k+1|k} = c_{k+1} + F_{k+1}x_{k+1|k-1} + K_k e_k$, where $K_k = F_{k+1}P_{k|k-1}H_k^{-1}$ is called the gain matrix. The state-space is then represented as $\{Z_k = d_k + H_kx_{k|k-1} + e_k; x_{k+1|k} = c_{k+1} + F_{k+1}x_{k|k-1} + K_k e_k\}$ (6), where the two noise processes $v_k$ and $w_k$ have been replaced by the process $\hat{e}_k$ and the transition equation is defined on the estimate of the state vector $\tilde{x}_{k|k-1}$, and not directly on the state vector $x_k$.

In the case of the tracking problem (3), the innovation representation yields $\{\tilde{x}_{k|k-1} = d_k + \hat{e}_k; \tilde{x}_{k+1|k} = \tilde{x}_{k|k-1} + K_k e_k\}$. It can be shown that the gain matrix $K_k$ can be construed as the matrix of ‘regression’ coefficients of $\tilde{x}_{k+1|k}$ on $\hat{e}_k$ the innovation at time $t_k$ (cf. appendix).

Interpretation of the correction mechanism of the Kalman filter
At time $t_k$, KF performs an update of the previous weights estimates $\tilde{w}_{k|k-1}$ by applying the correction term $K_k \hat{e}_k = \tilde{P}_{k|k-1}^{-1} \hat{e}_k$, where $\hat{e}_k$ is the normalized tracking error. Recall also that $\tilde{P}_{k|k-1}^{-1} = [w_k - \tilde{w}_{k|k-1}]^T [\tilde{H}_k]^T [\tilde{H}_k]^{-1}$ is the variance matrix of the state estimation error $w_k - \tilde{w}_{k|k-1}$.

We are now in a position to explain how KF adjusts the weights between two rebalancing dates. Here are some facts to understand the statistical prediction-correction system behind KF.

1. First, notice that the larger the normalized tracking error $\hat{e}_k$, the larger the change in the allocation $\Delta w_{k+1|k}$. This remark compounds three smaller ones.

   a) The size of $\hat{e}_k$ takes into account the relative size of $\hat{e}_k$ with respect to its covariance $V_k$.
   b) Note that $V_k = \tilde{P}_{k|k-1}^{-1} + S_k$. Thus, the variance of the tracking error depends on the covariance matrix $\tilde{P}_{k|k-1}$ of the past state estimation error and the variance of the current observation noise $\eta_k$. Hence, the larger the recent past errors of the Kalman Filter, the smaller the normalized tracking error $\hat{e}_k$ will be. In other words, ceteris paribus, the smaller the recent past errors, the ‘stronger’ is the algorithm’s reaction to the last observed tracking error. We have $\hat{e}_k > 0 => |\Delta w_{k+1|k}| > 0$ or $\Delta w_{k+1|k} > 0$.

2. Second, assume that $\tilde{P}_{k|k-1}$ is a diagonal matrix. The errors on the estimated weights are not correlated. The direction of change for the asset class I will then be given by the sign of $r_k^i \times \hat{e}_k$

   $r_k^i \times \hat{e}_k > 0 => \Delta w_{k+1|k} > 0$

   The directions are then adjusted to take into account the volatility of the Kalman filter errors on the estimated weights. For the $i^{th}$ factor, we have

   $\Delta w_{k+1|k}^i = \tilde{P}_{k|k-1}^{-1} r_k^i \hat{e}_k^i$.

   If KF has made a lot of errors on the weight of one factor (which means that the weights have highly changed in the past), it will perform a large correction $\tilde{P}_{k|k-1}^{-1} r_k^i \hat{e}_k^i$.

3. Third, assume that $\tilde{P}_{k|k-1}$ is not a diagonal matrix. The correction done by KF takes into account of the correlations between the errors on the estimated weights $\Delta w_{k+1|k}^i = \tilde{w}_{k+1|k}^i - \tilde{w}_{k|k-1}^i = \hat{e}_k^i \sum_{m=1}^{M} (\tilde{P}_{k|k-1})^{-1}_{i,m} r_k^m$.

   Suppose that $\hat{e}_k < 0$ and $r_k^i > 0$. According to point 2 above, the weight of the first factor should be reduced. However, because of the correlations between the errors on the estimated weights, there may be an opposite correction $\Delta w_{k+1|k}^i$, because for instance the errors on the other factors are negatively correlated with the error on the first factor and the performances of the other factors are negative.

4. Finally, notice that when, at time $t_k$, the replication strategy has the same performance as the fund’s strategy, KF does not change the estimated weights $\hat{e}_k = 0 => \tilde{w}_{k+1|k} = \tilde{w}_{k|k-1}$.

An example with a well-diversified Hedge Fund index
As in Roncalli and Teiletche (2008), we consider replicating the HFR Fund Weighted Composite index as an example. The model considered (6F) is

$$
\begin{align*}
\tau_k^{(nF)} &= \sum_{i=1}^{6} w_k^{(i)} \tau_k^{(i)} + \eta_k \\
 w_k &= w_{k-1} + \tau_k \\
\theta_k &= \text{diag}(\sigma_1^2, ..., \sigma_6^2)
\end{align*}
$$

where the set of factors that served as a basis for this exercise is: an equity exposure in the S&P 500 index (SPX), a long/short position between Russell 2000 and S&P 500 indexes (RTY/SPX), a long/short position between DJ Eurostoxx 50 and S&P 500 indexes (SX5E/SPX), a long/short position between Topix and S&P 500 indexes (TPX/SPX), a bond position in the 10-year U.S. Treasury (UST), and an FX position in the EUR/USD.
Results

To present realistic results, we assumed that replication of the exposures to each factor was done using futures5 (hedged in USD) and that the sampling period is one month. The study period begins in January 1994 and ends in September 2008. We estimated the model described in (8)6. The estimates of the parameters are (in %) $\hat{\eta}^2 = 5.48 \times 10^{-5}$, $\hat{\sigma}_1^2 = 7.34 \times 10^{-4}$, $\hat{\sigma}_2^2 = 2.83 \times 10^{-4}$, $\hat{\sigma}_3^2 = 2.09 \times 10^{-3}$, $\hat{\sigma}_4^2 = 5.25 \times 10^{-4}$ and $\hat{\sigma}_6^2 = 6.26 \times 10^{-4}$. The resulting estimated exposures are presented in Figure 1.

Interpretation of the results

A closer look at the results of the previous estimation demonstrates, as we show below, that replication using KF provides better replicators than traditional methods in the sense that it captures a better part of the performance of the HF benchmark while providing estimated weights that possess a sensible explanation of the dynamic investment strategy of the underlying index. To do so, we first introduce the alternative beta concept, before moving to an attribution of performance (AP) of the replicating strategy.

The alternative beta concept

As mentioned in Hasanhodzic and Lo (2007), Lo (2008), and Roncalli and Teiletche (2008), we may compute the attribution of performance of the return $r_k^{(HF)}$ of hedge funds indices in several ways. In practice, the attribution of performance is often done directly on the absolute returns.7 First, rewrite the return of the hedge fund portfolio using the following decomposition

$$r_k^{(HF)} = \sum_{i=1}^{m}(w_k^{(i)} - w_{k-1}^{(i)})t^{(i)} + \sum_{i=1}^{m}w_{k-1}^{(i)}r_k^{(i)} + \sum_{i=1}^{m}w_{k-1}^{(i)}\eta_k.$$  
(3)

where $w_{k-1}^{(i)}$ are the fixed weights on the different asset classes, e.g. $w_{k-1}^{(i)} = E(w_k^{(i)}/\hat{\sigma})$ and $r_k^{(i)}$ is the return of the risk-free asset.

The first approach to consider is the traditional alpha/beta decomposition derived from the CAPM $r_k^{(HF)} = \alpha_k + \beta_k$, where $\beta_k$ is the component of return attributed to the benchmarks and where the sensitivities of the fund’s portfolio to the benchmarks are considered constant. In this alpha/beta decomposition, we thus have

$$\begin{cases} \alpha_k = r_k^{(HF)} - \beta_k \beta_k = \left(1 - \sum_{i=1}^{m} w_k^{(i)} \right)^{\tau_k} + \sum_{i=1}^{m} w_k^{(i)} r_k^{(i)} \end{cases}$$  
(4)

When taking the mathematical expectations in (10), one finds that the traditional alpha/beta decomposition will always underestimate the systematic part of the performance – the beta – and overestimate the idiosyncratic part – the alpha – as the contribution of the covariance between the factors and the exposures is lumped into the idiosyncratic part. Instead, one may consider another decomposition $r_k^{(HF)} = \alpha^{AB} + \beta^{TB} + \beta^{AB}$ where $\beta^{TB}$ is the traditional beta and $\beta^{AB}$ is called alternative beta. We have

$$\begin{cases} \alpha^{AB} = \sum_{i=1}^{m} w_k^{(i)} \hat{r}_k^{(i)} - \sum_{i=1}^{m} w_k^{(i)} \hat{\sigma}_k^{(i)} \end{cases}$$  
(5)

The alternative beta $\beta^{AB}$ thus captures the part of the performance of the fund due to an active management of the portfolio’s expositions to the different benchmarks. For a discussion of active versus passive management, we refer the reader to Roncalli and Teiletche (2008) and Lo (2008). After approximating $w_k$ with $\hat{w}_{k-1}$, the clone gives access to the sum of the traditional beta and the alternative beta $r_k^{(clone)} = (1 - \sum_{i=1}^{m} \hat{w}_{k-1}^{(i)})\eta_k + \sum_{i=1}^{m} \hat{w}_{k-1}^{(i)} \hat{r}_k^{(i)}$. The term $\alpha^{AB}$ is called the alternative alpha. It is computed as $\alpha^{AB} = r_k^{(HF)} - r_k^{(clone)}$. We have reported the performance attribution of $\alpha^{AB}$ components in Figure 2. Notice that a large part of the HF returns are not explained by the traditional alpha but by the alternative beta. For the

5 When the future does not exist, we approximate the monthly performance by the monthly return of the corresponding TR index minus the one-month domestic Libor and the hedging cost.

6 The parameters $w_0$ and $P_0$ are initialized at $w_0 = 0$; $P_0 = I_{6x6}$.

7 In this case, we assume that the cash investment is part of the beta component.

8 There are several ways to compute the fixed weights. One approach is to consider the mean of the dynamic weights $w_k^{(i)} = \frac{1}{m} \sum_{i=1}^{m} w_k^{(i)}$. Another approach is to compute the OLS regression on the entire period $t_{-1} \rightarrow t_k$, $w_k^{(i)} = \frac{1}{m} \sum_{i=1}^{m} \hat{w}_k^{(i)} + w_k^{(i)}$. Finally, we may estimate the weights using the Kalman filter by imposing that $Q_k = 0_{m \times m}$. In this case, the weights $\hat{w}_k^{(i)}$ correspond to the recursive OLS estimates.

### Table 1 – Estimated yearly alpha (in %)

<table>
<thead>
<tr>
<th>Period</th>
<th>Traditional Alpha</th>
<th>Traditional Beta</th>
<th>Alternative Alpha</th>
<th>Alternative Beta</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994-2008</td>
<td>3.80</td>
<td>5.92</td>
<td>2.22</td>
<td>7.55</td>
<td>9.94</td>
</tr>
<tr>
<td>1997-2008</td>
<td>3.14</td>
<td>5.46</td>
<td>1.14</td>
<td>7.55</td>
<td>8.77</td>
</tr>
<tr>
<td>2000-2008</td>
<td>2.20</td>
<td>4.02</td>
<td>1.48</td>
<td>4.75</td>
<td>6.30</td>
</tr>
</tbody>
</table>

![Figure 1 – Estimated weights of the 6F model (Jan 1994 - Sep 2008)](image-url)
entire period, the alternative alpha explains about 23% of the HF returns whereas the alternative beta explains about 77%. The decomposition between alpha and beta over several periods is reported in Table 1. Note that the alpha is overestimated using traditional beta.

**Performance attribution of the replicated strategy**

In Table 2, we report the performance attribution of the ABS factors’ exposures for our example. The main contributor to the replicated strategy is the long equity exposure. It is interesting to note that three other strategies have a significant contribution. They are the two L/S equity strategies on small caps and Eurozone and the FX position EUR/USD. Finally, the last two positions have a small absolute contribution to the performance: the L/S equity on Japan and the 10-year U.S. bond position. In a first approach, one may consider the elimination of these factors. However, they may help track the volatility of the HF index, therefore contributing to the performance as well.

Interestingly, using the KF estimates, we are now able to explain the success of the HF industry between 2000 and 2003. Notice in Figure 3 that the highest exposure of the HF industry to the directional equity market was in March 2000 and represented more than 60% of the overall exposure. After March 2000, the HF industry decreased the leverage on equity and modified the bets on L/S equity. In the right graph, we compare the performance of the alternative beta strategy with respect to two other strategies. The first one uses the fixed allocation of March 2000 for all the asset classes and the second corresponds to the alternative beta, except for the directional equity exposure which is fixed and equal to the equity beta of March 2000. It appears that the relative good performance of the HF industry may be explained by two components: equity deleverage and good bets on L/S equity on RTY/SPX and SX5E/SPX. We estimate that with respect to the allocation of March 2000, the equity deleverage explains 40% of the outperformance whereas the reallocation of the L/S equity explains about 60% of the outperformance.

**Which strategies may be replicated?**

The example provided above is of course no proof that the methodology we have exposed so far is the panacea to the replication problem. Rather, the preceding example could almost be taken as a teaching case used to demonstrate the aptitudes of this formulation of the replication problem to provide satisfying answers. It is, however, important to better understand what types of strategies followed by the HF industry may subject themselves well to this replication process. To try to provide an answer to this problem, we thus estimated the 6F\(^9\) on a series of HF

---

\(^9\) We also estimated a factor model using seven factors (7F) including some nonreplicable factors traditionally used in the literature [Hasanhodzic and Lo (2007)]. If this (7F) model performs better on a number of accounts, providing better performances, lower volatility, lower volatility of the tracking error, better correlation of the returns of the tracker with its benchmark, one must however make note of three facts. First, any gain is in general small and parsimony considerations suggest a smaller model. Second, from an investment point of view some of the factors in (7F) are not easily implementable, and any gain in performance may be offset by additional implementation costs these factors could involve. Third, the gain in the tracking performance is reflected, even if only slightly, by higher drawdowns. Detailed results are available from the authors on demand.
indexes representing general categories of strategies. The HFRI index trackers using the 6F model were compared to their benchmarks (detailed results are available from the authors). The key points of an analysis of our results can be summarized in the following way. Overall, HF trackers have smaller Sharpe ratios than their respective indexes, even though they generally exhibit lower volatilities. However, they also present a smaller risk if one measures risk as the maximum drawdown or as excess kurtosis of the returns. Some strategies present low correlation with their respective trackers and one can thus conclude that they are difficult to replicate by the method employed here. This concerns mainly illiquid strategies (i.e., distressed securities), strategies with small betas (i.e., relative value), and strategies based on stock picking (like merger arbitrage or equity market neutral). Also of note, some tracker may not have a high correlation with their respective index, but may still exhibit similar performance. This is, for example, the case of funds of funds (FOF in the tables). One reason for this may be that part of the alternative betas of the underlying funds is captured by the fee structure of the FOF and thus do not appear in their performance, while the replicating process provides a direct access to this part of the performance.

Finally, on a more particular note, it is worth taking a look at two particular strategies. First, on the “emerging market: Russia/E. Europe” HFRI index, it is worth noting that the model performs particularly poorly, pointing at the fact that in our pool of factors, none had a strong relation with the economy of that region of the world. Second, the “macro: syst. diversified” is the one case where the model produces a clone with higher drawdowns than the actual HFRI macro: syst. diversified. One reason behind these poor results is probably the set of factors used. Another reason could be the inadequacy of factor models in this case, but one could ask why, if the concept of factor model is the underlying problem, our results do not show more results similar to these. This illustrates that the better results obtained with our replication methodology cannot replace a careful choice of the set of factors. It is also a sign that if a better selection methodology is found, it would still have to rely on some economic insight, echoing results found in the literature [Amenc et al. (2007)].

**Alpha considerations**

In the previous sections, we have developed and demonstrated the use of Bayesian filters to answer the question of HF replication. In this section, we focus on the part of the HF performance left unexplained by the methods presented above. We thus look into the alternative alpha component, and look for possible explanations of its origin. In the previous sections, we suggested possible sources including high frequency trading and investments in illiquid assets. To these two, we add here another component which stems not from specific strategies but from the fact that, by construction, a replicating portfolio implements its exposure with a time lag with respect to the replicated HF profile. We focus here on the impacts of the implementation lag and the illiquid investments, in this respective order. Nonlinearities are addressed in a companion paper [Roncalli and Weisang (2009)].

Starting with the impact of the implementation lag, note first that replication clones are obtained using lagged exposures with a lag \(d = -1\). If one uses \(d = 0\), one assumes that one can implement at time \(t_k\) the true exposures of the period \([t_k, t_{k+1}]\) and for \(d > 0\), the implemented exposures are those estimated for the period \([t_k + d, t_{k+1} + d + 1]\). Putting to test our claim that the implementation lag contributes to the alpha, we computed backtests of the portfolios obtained for \(d = 0, 1, 2\) using the 6F model presented above and the HFRI Fund Weighted Composite Index and compared them with the case \(d = -1\). The results obtained are provided in Table 3. Unsurprisingly, with the added information, the results are substantially better, with the best results for the contemporaneous implementation (\(d = 0\)). The part of the HF performance explained by the alternative beta clone jumps by about 10% to 85%, reducing the alpha
component from around 25% to about 15%. In other words, in our example, 40% of the alternative alpha is explained by the implementation delay. In this particular case, we can therefore propose a new breakdown on the HF performance.

75% of the performance corresponds to alternative beta which may be reproduced by the tracker and 25% is the alternative alpha of which 10% corresponds in fact to alternative beta which may not be implemented and are lost due to the dynamic allocation and 15% makes up a component that we call the pure alternative alpha. It is also interesting to note that the volatility of the pure alpha component \(\sigma_{TE}\) for \(d = 0, 1, 2\) is lower and is half of the volatility of the alternative alpha. We represent in Figure 4 the evolution of the two components of the alternative alpha, with \(a_1\) representing the contribution of the implementation lag to the alternative alpha and \(a_2\) the pure alternative alpha.

We now turn to our second claim that the alternative alpha stems from the illiquidity premia associated with investment in illiquid assets. Using the results of our previous experiment on implementation delay, we focus on explaining the pure component of the alternative alpha. One possible way to substantiate this claim would be, for example, to extract the pure alpha component and run an analysis in the same fashion as it was done at first for HF replication using regressions to determine whether factors representing different illiquid assets, such as distressed securities or private equity, are able to explain the returns of the pure alternative alpha. We proceed differently here by keeping in mind the idea to demonstrate that it is possible to access the performance of this pure component from an investment perspective. One idea then is to build a core/satellite portfolio where the core is the alternative beta and the satellite is a basket of illiquid or optional strategies. The previous construction of alternative investments has some important advantages. For example, one could consider a portfolio with 70% of alternative beta, 10% of optional or quantitative strategies, 10% of real estate, and 10% of private equity. The core/satellite approach permits us to distinguish clearly liquid and illiquid investments, small term and long term investments. In our example, these three satellite strategies are respectively proxied by equally weighted portfolios of the SGI volatility premium index and JP Morgan carry max index, UK IPD TR all property index and NCREIF property index, and LPX buyout index and LPX venture index. The results of this approach are displayed in Figure 5.

After obtaining these results, there is no doubt in our mind that, in this case at least, the pure alternative alpha component can be replicated by means of this core/satellite strategy. One may wonder, however, why there is apparently no need to take into account a high frequency factor. Beside the fact that it is rather good news from a practitioner point-of-view, one must point out that in our example, we replicated the HFRI Fund Weighted Composite Index, which is the most general industry aggregate provided by Hedge Fund Research, Inc. As such, in light of the results presented by Diez de los Rios and Garcia (2008), we surmise that the effect of high frequency trading, which would appear as nonlinear, is negligible.

---

**Table 3 – Results of time lags implementation on the replicating portfolios**

<table>
<thead>
<tr>
<th>d</th>
<th>(\mu_{HF}^{\text{TE}})</th>
<th>(\mu_{\text{Tracker}}^{\text{TE}})</th>
<th>(\mu_{\text{AB}})</th>
<th>(\sigma_{\text{AB}})</th>
<th>(\rho)</th>
<th>(\tau)</th>
<th>(\theta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>9.94</td>
<td>7.55</td>
<td>75.93</td>
<td>3.52</td>
<td>87.35</td>
<td>67.10</td>
<td>84.96</td>
</tr>
<tr>
<td>0</td>
<td>9.94</td>
<td>8.39</td>
<td>84.45</td>
<td>1.94</td>
<td>96.17</td>
<td>80.18</td>
<td>94.55</td>
</tr>
<tr>
<td>1</td>
<td>9.94</td>
<td>8.42</td>
<td>84.77</td>
<td>2.05</td>
<td>95.71</td>
<td>80.09</td>
<td>94.42</td>
</tr>
<tr>
<td>2</td>
<td>9.94</td>
<td>8.26</td>
<td>83.11</td>
<td>2.22</td>
<td>94.96</td>
<td>78.42</td>
<td>93.59</td>
</tr>
</tbody>
</table>

\(\mu_{HF}\) is the annualized performance; \(\mu_{AB}\) the proportion of the HF index performance explained by the tracker and \(\mu_{\text{AB}}\) the yearly tracking error. \(\rho\), \(\tau\) and \(\theta\) are respectively the linear correlation, the Kendall tau and the Spearman rho between the monthly returns of the HF index and the tracker. All statistics are expressed in percents.
Discussion
In the sections above, we demonstrated the efficiency of Bayesian filters – in particular the Kalman filter – in capturing the tactical asset allocation. Furthermore, completed by a core/satellite portfolio strategy, we showed this approach would be enough to replicate a general HF index like HFRI. Nonetheless, one could legitimately ask ‘so what?’ question. Hedge returns are renowned to be generated using complex financial instruments generating highly nonlinear returns. Obviously the exercise above does not include any complex product, and tactical asset allocation is more the realm of ‘traditional’ managers than hedge funds. Thus, it seems, it falls short of answering the question at hand: replicating any hedge fund track records. Our answer is threefold.

First, the philosophy of replication that we pursued here is a readily available methodology that directly translates into implementable investments. One could take issue and point out that it does not take into account the risk management perspective of hedge fund replication, i.e., using the methodology for risks assessment. But, the issue then is: who is the end user of clones and hedge fund replicates? In what we presented above, nothing forbids the inclusion of ‘rule-based’ factors mechanically reproducing an alternative strategy to represent a certain type of risk. Unfortunately, as we demonstrate in a companion paper [Roncalli and Weisang (2009)], these types of factors are often difficult to implement as they are extremely dependent on the data available, which are themselves not necessarily representative of investable opportunities.

Second, although it has been documented from the beginning of hedge fund replication [Fung and Hsieh (1997)], the existence and presence of nonlinearities in hedge fund returns seem persistent in only a handful of strategies [Diez de los Rios and Garcia (2008)]. Thus, a core/satellite approach capturing on one hand the tactical allocation between different asset classes, combined on the other hand with buy-and-hold strategies to capture risk premia of illiquid investments presents clear advantages, transparency not being last on the list. Again, nothing prevents the inclusion of rule-based factors in the tactical allocation part if the goal is risk assessment. Finally, the framework of tracking problems and their solvability using Bayesian filters provides readily available extensions. For example, using particle filters, one can try to integrate some nonlinearity in the replication methods [Roncalli and Weisang (2009)]. From the academics’ point of view, introducing particle filters opens a door for a better understanding of HF returns and the underlying risks of the HF strategies. If it already has direct implications from a risk management perspective, we also surmise that particles filters are one of the main avenues toward a better monitoring of, for now, unaccounted risks, as they are contained in the higher moments of the returns’ distribution.

Conclusion
In this paper, after providing a formal statistical framework to hedge fund replication, we limited ourselves to demonstrate that linear factor models can efficiently recover the tactical allocation using an adequate methodology. Furthermore, we considered how hedge fund replicates could reproduce the alpha. For sake of space, and because they present completely different challenges, we left the study of the replication of nonlinearities in hedge fund returns to another paper [Roncalli and Weisang (2009)]. Nevertheless, we believe the results presented in here to be very interesting both for the practitioners and the academics. From the practitioners’ point of view, by grounding all of our approaches into a general and coherent framework, and by meticulously adding complexity to the methodology, we demonstrated that a robust replication process can be obtained by means of mainstream statistical methods, such as the Kalman filter, provided that careful thought is given to the specification of the model and the type of instruments used in the replication process (particularly with respect to liquidity or other trading considerations). It is perhaps necessary to remind the reader again that as an investment toolbox to manage HF exposures (both long and short) and liquidity, the first quality of a HF clone should not be to be a hedge fund in itself. As such, and in line with this HF replication philosophy, our core/satellite approach showed that this robust approach (Kalman filter and liquid instruments) can still be supplemented by other illiquid investments to capture and reproduce more efficiently the risk profile of the hedge fund industry. Incidentally, it also hints at the efficiency of the ‘core’ method to capture the HF betas to classic asset classes. From the academic’s point of view, the new framework provided allows for readily available extensions, with similar problems having been already studied in other disciplines like engineering and signal processing.
The Capco Institute Journal of Financial Transformation

Tracking Problems, Hedge Fund Replication, and Alternative Beta

References


• Agarwal V., and N-Y. Naik, 2000, “Funds and portfolio decisions involving hedge funds,” Review of Financial Studies, 17, 63-98


• Hamilton, J. D., 1994, Time series analysis, Princeton University Press


• Hamilton, J. D., 1994, Time series analysis, Princeton University Press


Appendix 1

Contemporaneous representation of the Kalman Filter

If one assumes the tracking problem to be linear and Gaussian, one may prove that the optimal algorithm to estimate the state vector is the Kalman filter. The state space model is then given by

\[ x_k = c_k + F_k x_{k-1} + \nu_k \]

\[ z_k = h_k x_k + \epsilon_k \]

with \( \nu_k \sim N(0, Q_k) \) and \( \epsilon_k \sim N(0, R_k) \). Moreover, the initial distribution of the state vector is \( p(x_0) = \phi(x_0, m_0, P_0) \), where \( \phi(x, m, P) \) is the Gaussian pdf with argument \( x \), mean \( m \) and covariance matrix \( P \). The Bayes filter is then described by the following recursive equations

\[
\begin{align*}
\hat{x}_{k|k-1} &= \hat{x}_{k|k-1} + K_k (z_k - \hat{y}_{k|k-1}) \\
\hat{y}_{k|k-1} &= \hat{y}_{k|k-1} + K_k (z_k - \hat{y}_{k|k-1}) \\
\hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k (z_k - \hat{y}_{k|k-1}) \\
\hat{y}_{k|k} &= \hat{y}_{k|k-1} + K_k (z_k - \hat{y}_{k|k-1}) \\
\end{align*}
\]

where \( K_k = P_k A_{k|k-1} + R_k \) and \( P_k = P_k A_{k|k-1} P_k A_{k|k-1}^T + Q_k \).

The set of equations (A1) describes the Kalman filter algorithm. The quantities can be interpreted as follows:

- \( \hat{x}_{k|k-1} = \hat{X}_k | Z_{1:k-1} \) is the estimate of \( x_k \) based on all available information until time index \( t_k \);
- \( \hat{P}_{k|k-1} \) is the covariance matrix of the estimator \( \hat{x}_{k|k-1} \);
- \( z_{k|k-1} = E(z_k | z_{1:k-1}) \) is the estimate of \( z_k \) based on all available information until time index \( t_k \);
- \( \hat{V}_k \) is the covariance matrix of the tracking error \( \hat{V}_k = E(\hat{e}_k|e_k) \);
- \( \hat{\xi}_k \) is the covariance matrix of \( \hat{\xi}_k : \hat{\xi}_k = E(\hat{\xi}_k | z_{1:k}) \) is the estimate of \( \xi_k \) based on all available information until time index \( t_k \).

Finally, \( \hat{P}_{k|k} = \hat{P}_{k|k-1} + \hat{V}_k \hat{V}_k^T \) is the covariance matrix of \( \hat{P}_{k|k} \).

Interpretation of KF estimates updates

The joint density of the observational vectors \( z_1, \ldots, z_k \) can be written as \( p(z_1, \ldots, z_k) = p(z_1) \prod_{l=2}^{k} p(z_l | z_{l-1}) \). Transforming from \( z_1 \) to \( \hat{e}_1 = z_1 - \hat{z}_{1|1} \), we have \( p(\hat{e}_1, \ldots, \hat{e}_k) = p(\hat{e}_1) \prod_{l=2}^{k} p(\hat{e}_l) \) since \( p(z_1) = p(\hat{e}_1) \) and the Jacobian of the transformation is unity because each \( \hat{e}_l \) is \( z_l \) minus a linear function of \( z_{1:l-1} \) for \( l = 2, \ldots, k \). We deduce then that \( \hat{e}_1, \ldots, \hat{e}_k \) are independent from each other and that \( \hat{e}_1, \ldots, \hat{e}_k \) are independent from \( z_{1:1} \). This last property, combined with some well known results of multivariate regression, provides us with an interpretation of the gain matrix and the dynamical adjustment of the weights. Noticing that

\[
E(x_{k+1} | z_1, \ldots, z_k) = \sum_{i=1}^{n} \left( E(x_{k+1} | z_1, \ldots, z_k, \hat{e}_i) \right)
\]

where the second equality is a useful result from multivariate regression.

Hence, we see that in equation (6), the gain matrix \( K_k \) can be construed as the matrix of ‘regression’ coefficients of \( x_{k+1} \) on \( \hat{e}_k \) the innovation at time \( t_k \) [cf. e.g., Durbin and Koopman (2001) or Hamilton (1994)].

10 By definition, \( \hat{e}_k = z_k - (z_k | z_{1:k}) \), i.e. \( \hat{e}_k \) is the part of \( z_k \) that cannot be predicted from the past. For this reason, the process \( \hat{e}_k \) is sometimes called the innovation process.
Empirical Implementation of a 2-Factor Structural Model for Loss-Given-Default

Michael Jacobs, Jr. — Senior Financial Economist, Credit Risk Analysis Division, Department of Economic and International Affairs, Office of the Comptroller of the Currency

Abstract

In this study we develop a theoretical model for ultimate loss-given default in the Merton (1974) structural credit risk model framework, deriving compound option formulae to model differential seniority of instruments, and incorporating an optimal foreclosure threshold. We consider an extension that allows for an independent recovery rate process, representing undiversifiable recovery risk, having a stochastic drift. The comparative statics of this model are analyzed and in the empirical exercise, we calibrate the models to observed LGDs on bonds and loans having both trading prices at default and at resolution of default, utilizing an extensive sample of losses on defaulted firms (Moody’s Ultimate Recovery Database™), 800 defaults in the period 1987-2008 that are largely representative of the U.S. large corporate loss experience, for which we have the complete capital structures and can track the recoveries on all instruments from the time of default to the time of resolution. We find that parameter estimates vary significantly across recovery segments, that the estimated volatilities of recovery rates and of their drifts are increasing in seniority (bank loans versus bonds). We also find that the component of total recovery volatility attributable to the LGD-side (as opposed to the PD-side) systematic factor is greater for higher ranked instruments and that more senior instruments have lower default risk, higher recovery rate return and volatility, as well as greater correlation between PD and LGD. Analyzing the implications of our model for the quantification of downturn LGD, we find the ratio of the later to ELGD (the “LGD mark-up”) to be declining in expected LGD, but uniformly higher for lower ranked instruments or for higher PD-LGD correlation. Finally, we validate the model in an out-of-sample bootstrap exercise, comparing it to a high-dimensional regression model and to a non-parametric benchmark based upon the same data, where we find our model to compare favorably. We conclude that our model is worthy of consideration to risk managers, as well as supervisors concerned with advanced IRB under the Basel II capital accord.

1 The views expressed herein are those of the author and do not necessarily represent a position taken by of the Office of the Comptroller of the Currency or the U.S. Department of the Treasury.
Loss given default (LGD), the loss severity on defaulted obligations, is a critical component of risk management, pricing and portfolio models of credit. This is among the three primary determinants of credit risk, the other two being the probability of default (PD) and exposure of default (EAD). However, LGD has not been as extensively studied, and is considered a much more daunting modeling challenge in comparison to other components, such as PD. Starting with the seminal work by Altman (1968), after many years of actuarial tabulation by rating agencies, predictive modeling of default rates is currently in a mature stage. The focus on PD is understandable, as traditionally credit models have focused on systematic components of credit risk which attract risk premia, and unlike PD determinants of LGD have been ascribed to idiosyncratic borrower specific factors. However, now there is an ongoing debate about whether the risk premium on defaulted debt should reflect systematic risk, in particular whether the intuition that LGDs should rise in worse states of the world is correct, and how this could be refuted empirically given limited and noisy data [Carey and Gordy (2007)]. The recent heightened focus on LGD is evidenced the flurry of research into this relatively neglected area [Acharya et al. (2007), Carey and Gordy (2007), Altman et al. (2001, 2003, 2004), Altman (2006), Gupton et al. (2000, 2005), Araten et al. (2003), Frye (2000 a,b,c, 2003), Jarrow (2001)]. This has been motivated by the large number of defaults and near simultaneous decline in recovery values observed at the trough of the last two credit cycle circa 2000-2002 and 2008-2009, regulatory developments such as Basel II [BIS (2003, 2005, 2006), OCC et al. (2007)], and the growth in credit markets. However, obstacles to better understanding and predicting LGD, including dearth of data and the lack of a coherent theoretical underpinning, have continued to challenge researchers. In this paper, we hope to contribute to this effort by synthesizing advances in financial theory to build a model of LGD that is consistent with a priori expectations and stylized facts, internally consistent and amenable to rigorous validation.

In addition to answering the many questions that academics have, we further aim to provide a practical tool for risk managers, traders, and regulators in the field of credit.

LGD may be defined variously depending upon the institutional setting or modeling context, or the type of instrument (traded bonds versus bank loans) versus the credit risk model (pricing debt instruments subject to the risk of default versus expected losses or credit risk capital). In the case of bonds, one may look at the price of traded debt at either the initial credit event4, the market values of instruments received at the resolution of distress5 [Keisman et al. (2000), Altman and Kishore (1996)], or the actual cash-flows incurred during a workout.5 When looking at loans that may not be traded, the eventual loss per dollar of outstanding balance at default is relevant [Asarnow and Edwards (1995), Araten et al. (2003)]. There are two ways to measure the latter – the accounting LGD refers to nominal loss per dollar outstanding at default5 while the economic LGD refers to the discounted cash flows to the time of default taking into consideration when cash was received. The former is used in setting reserves or a loan loss allowance, while the latter is an input into a credit capital attribution and allocation model. In this study we develop various theoretical models for ultimate loss-given default in the Merton (1974) structural credit risk model framework. We consider an extension that allows for differential seniority within the capital structure, an independent recovery rate process, representing undiversifiable recovery risk, with stochastic drift. The comparative statics of this model are analyzed in a framework that incorporates an optimal foreclosure threshold [Carey and Gordy (2007)]. In the empirical exercise, we calibrate alternative models for ultimate LGD on bonds and loans having both trading prices at default and at resolution of default, utilize an extensive sample of rated defaulted firms in the period 1987-2008 (Moody’s Ultimate Recovery DatabaseTM - URD™), 800 defaults (bankruptcies and out-of-court settlements of distress) that are largely representative of the U.S. large corporate loss experience, for which we have the complete capital structures and can track the recoveries on all instruments to the time of default to the time of resolution. We find that parameter estimates vary significantly across recovery segments. We find that the estimated volatilities of the recovery rate processes, as well as of their random drifts are increasing in seniority, in particular for bank loans as compared to bonds. We interpret this as reflecting greater risk in the ultimate recovery for higher ranked instruments having lower expected loss severities (or ELGDs). Analyzing the implications of our model for the quantification of downturn LGD, we find the latter to be declining in expected LGD, higher for worse ranked instruments, increasing in the correlation between the processes driving firm default and recovery on collateral, and increasing in the volatility of the systematic factor specific to the recovery rate process or the volatility of the drift in such. Finally, we validate the leading model derived herein in an out-of-sample bootstrap exercise, comparing it to a high-dimensional regression model, and to a non-parametric benchmark based upon the same data, where we find our model to compare favorably. We conclude that our model is worthy of consideration to risk managers, as well as supervisors concerned with advanced IRB under the Basel II capital accord.

---

4. This is equivalent to one minus the recovery rate, or dollar recovery as a proportion of par, or EAD assuming all debt becomes due at default. We will speak in terms of LGD as opposed to recoveries with a view toward credit risk management applications.
5. By default we mean either bankruptcy (Chapter 11) or other financial distress (payment default). In a banking context, this is defined as being synonymous with respect to non-accrual on a discretionary or non-discretionary basis. This is akin to the notion of default in Basel, but only proximate.
6. Note that this may be either the value of pre-petition instruments received valued at emergence from bankruptcy, or the market values of new securities received in settlement of a bankruptcy proceeding, or as the result of a distressed restructuring.
7. Note that the former may be viewed as a proxy to this, the pure economic notion.
8. In the context of bank loans, this is the cumulative net charge-off as a percent of book balance at default (the net charge-off rate).
Review of the literature

In this section we will examine the way in which different types of theoretical credit risk models have treated LGD – assumptions, implications for estimation and application. Credit risk modeling was revolutionized by the approach of Merton (1974), who built a theoretical model in the option pricing paradigm of Black and Scholes (1973), which has come known to be the structural approach. Equity is modeled as a call option on the value of the firm, with the face value of zero coupon debt serving as the strike price, which is equivalent to shareholders buying a put option on the firm from creditors with this strike price. Given this capital structure, log-normal dynamics of the firm value and the absence of arbitrage, closed form solutions for the default probability and the spread on debt subject to default risk can be derived. The LGD can be shown to depend upon the parameters of the firm value process as is the PD, and moreover is directly related to the latter, in that the expected residual value to claimants is increasing (decreasing) in firm value (asset volatility or the level of indebtedness). Consequently, LGD is not independently modeled in this framework; this was addressed in much more recent versions of the structural framework [Frye (2000), Dev and Pykhtin (2002), Pykhtin (2003)]. Extensions of Merton (1974) relaxed many of the simplifying assumptions of the initial structural approach. Complexity to the capital structure was added by Black and Cox (1976) and Geske (1977), with subordinated and interest paying debt, respectively. The distinction between long- and short-term liabilities in Vasicek (1984) was the precursor to the KMVä model. However, these models had limited practical applicability, the standard example being evidence of Jones et al. (1984) that these models were unable to price investment grade debt any better than a naive model with no default risk. Further, empirical evidence in Franks and Touros (1989) showed that the adherence to absolute priority rules (APR) assumed by these models are often violated in practice, which implies that the mechanical negative relationship between expected asset value and LGD may not hold. Longstaff and Schwartz (1995) incorporate into this framework a stochastic term structure with a PD-interest rate correlation. Other extensions include Kim at al. (1993) and Hull and White (2002), who examine the effect of coupons and the influence of options markets, respectively.

Partly in response to this, a series of extensions ensued, the so-called ‘second generation’ of structural form credit risk models [Altman (2003)]. The distinguishing characteristic of this class of models is the relaxation of the assumption that default can only occur at the maturity of debt – now default occurs at any point between debt issuance and maturity when the firm value process hits a threshold level. The implication is that LGD is exogenous relative to the asset value process, defined by a fixed (or exogenous stochastic) fraction of outstanding debt value. This approach can be traced to the barrier option framework as applied to risky debt of Black and Cox (1976). All structural models suffer from several common deficiencies. First, reliance upon an unobservable asset value process makes calibration to market prices problematic, inviting model risk. Second, the limitation of assuming a continuous diffusion for the state process implies that the time of default is perfectly predictable [Duffie and Lando (2001)]. Finally, the inability to model spread or downgrade risk distorts the measurement of credit risk. This gave rise to the reduced form approach to credit risk modeling [Duffie and Singleton (1999)], which instead of conditioning on the dynamics of the firm, posit exogenous stochastic processes for PD and LGD. These models include (to name a few) Litterman and Iben (1991), Madan and Unal (1995), Jarrow and Turnbull (1995), Lando (1998), and Duffie (1998). The primitives determining the price of credit risk are the term structure of interest rates (or short rate), and a default intensity and an LGD process. The latter may be correlated with PD, but it is exogenously specified, with the link of either of these to the asset value (or latent state process) not formally specified. However, the available empirical evidence [Duffie and Singleton (1999)] has revealed these models deficient in generating realistic term structures of credit spreads for investment and speculative grade bonds simultaneously. A hybrid reduced - structural form approach of Zhou (2001), which models firm value as a jump diffusion process, has had more empirical success, especially in generating a realistic negative relationship between LGD and PD [Altman et al. (2006)]. The fundamental difference of reduced with structural form models is the unpredictability of defaults; PD is non-zero over any finite time interval, and the default intensity is typically a jump process (e.g., Poisson), so that default cannot be foretold given information available the instant prior. However, these models can differ in how LGD is treated. The recovery of treasury assumption of Jarrow and Turnbull (1995) assumes that an exogenous fraction of an otherwise equivalent default-free bond is recovered at default. Duffie and Singleton (1999) introduce the recovery of market value assumption, which replaces the default-free bond by a defaultable bond of identical characteristics to the bond that defaulted, so that LGD is a stochastically varying fraction of market value of such bond the instant before default. This model yields closed form expressions for defaultable bond prices and can accommodate the correlation between PD and LGD; in particular, these stochastic parameters can be made to depend on common systematic or firm specific factors. Finally, the recovery of face value assumption [Duffie (1998), Jarrow et al. (1997)] assumes that LGD is a fixed (or seniority specific) fraction of par, which allows the use of rating agency estimates of LGD and transition matrices to price risky bonds. It is worth mentioning the treatment of LGD in credit models that attempt to quantify unexpected losses analogously to the Value-at-Risk (VaR) market risk models, so-called credit VaR models – Creditmetrics™ [Gupton et al. (1997)], KMV CreditPortfolioManager™ [KMV Corporation (1984)], CreditRisk+™ [Credit Suisse Financial Products (1997)], and CreditPortfolioView™ [Wilson (1998)]. These models are widely employed by financial institutions to determine expected credit losses as well as economic capital (or unexpected losses) on credit portfolios. The main output of these models is a probability distribution function for future credit losses.
over some given horizon, typically generated by simulation of analytical approximations, as it is modeled as highly non-normal (asymmetrical and fat-tailed). Characteristics of the credit portfolio serving as inputs are LGDs, PDs, EADs, default correlations, and rating transition probabilities. Such models can incorporate credit migrations (mark-to-market mode – MTM), or consider the binary default versus survival scenario (default mode – DM), the principle difference being that in addition an estimated transition matrix needs to be supplied in the former case. Similar to the reduced form models of single name default, LGD is exogenous, but potentially stochastic. While the marketed vendor models may treat LGD as stochastic (e.g., a draw from a beta distribution that is parameterized by expected moments of LGD), there are some more elaborate proprietary models that can allow LGD to be correlated with PD.

We conclude our discussion of theoretical credit risk models and the treatment of LGD by considering recent approaches, which are capable of capturing more realistic dynamics, sometimes called ‘hybrid models.’ These include Frye (2000a, 2000b), Jarrow (2001), Bakshi et al. (2001), Jokivuolle et al. (2003), Pykhtin (2003), and Carey and Gordy (2007). Such models are motivated by the conditional approach to credit risk modeling, credited to Finger (1999) and Gordy (2000), in which a single systematic factor derives defaults. In this more general setting, they share in common the feature that dependence upon a set of systematic factors can induce an endogenous correlation between PD & LGD. In the model of Frye (2000a, 2000b), the mechanism that induces this dependence is the influence of systematic factors upon the value of loan collateral, leading to lower recoveries (and higher loss severity) in periods where default rates rise (since asset values of obligors also depend upon the same factors). In a reduced form setting, Jarrow (2001) introduced a model of codependent LGD and PD implicit in debt and equity prices.  

Theoretical model

The model that we propose is an extension of Black and Cox (1976). The baseline mode features perpetual corporate debt, a continuous and a positive foreclosure boundary. The former assumption removes the time dependence of the value of debt, thereby simplifying the solution and comparative statics. The latter assumption allows us to study the endogenous determination of the foreclosure boundary by the bank, as in Carey and Gordy (2007). We extend the latter model by allowing the coupon on the loan to follow a stochastic process, accounting for the effect of illiquidity. Note that in this framework, we assume no restriction on asset sales, so that we do not consider strategic bankruptcy, as in Leland (1994) and Leland and Toft (1996).

Let us assume a firm financed by equity and debt, normalized such that the total value of perpetual debt is 1, divided such that there is a single loan with face value \( \lambda \) and a single class of bonds with a face value of \( 1 - \lambda \). The loan is senior to that bond, and potentially has covenants which permit foreclosure. The loan is entitled to a continuous coupon at a rate \( c \), which in the baseline model we take as a constant, but may evolve randomly. Equity receives a continuous dividend, having a constant and a variable component, which we denote as \( \lambda + \rho V_t \), where \( V_t \) is the value of the firm’s assets at time \( t \). We impose the restriction that \( 0 < \rho c < c \), where \( r \) is the constant risk-free rate. The asset value of the firm, net of coupons and dividends, follows a Geometric Brownian Motion with constant volatility \( \sigma \):  

\[
\frac{dV_t}{V_t} = (r - \rho - C/V_t)dt + \sigma dZ_t
\]

where in (3.1) we denote the fixed cash outflows per unit time as: \( C = c\lambda + \gamma (1 - \lambda) + \delta \) (3.2), where in (3.2) \( \gamma \) and \( \delta \) are the continuous coupon rate on the bond and dividend yield on equity, respectively. Default occurs at time \( t \) and is resolved after a fixed interval \( \tau \), at which point dividend payments cease, but the loan coupon continues to accruce through the settlement period. At the point of emergence, loan holders receive \( (\lambda, \exp(\kappa \tau), V_{1,\tau}) \), or the minimum of the legal claim or the value of the firm at emergence. We can value the loan at resolution, under either physical or risk neutral measure, using the standard Merton (1974) formula. Denote the total legal claim at default by:  

\[
D = \lambda \exp(\kappa \tau) + (1 - \lambda)
\]

This follows from the assumption that the coupon \( c \) on the loan with face value \( \lambda \) continues to accrue at the contractual rate throughout the resolution period \( \tau \), whereas the bond with face value \( 1 - \lambda \) does not.

Thus far we have taken the solved for LGD under the assumption that the senior bank creditors foreclose on the bank when the value of assets is \( V_t \), where \( t \) is the time of default. However, this is not realistic, as firm value fluctuates throughout the bankruptcy or workout period, and we can think that there will be some foreclosure boundary (denoted by \( s \)) below which foreclosure is effectuated. Furthermore, in most cases there exists a covenant boundary, above which foreclosure cannot occur, but below which it may occur as the borrower is in violation of a contractual provision. For the time being, let us ignore the latter complication, and focus on the optimal choice of \( s \) by the bank. In the general case of time dependency in the loan valuation equation \( F(V_t | \lambda, \delta, r, \tau) \), following Black and Cox (1976), we have to solve a following second order partial differential equation. Following Carey and Gordy (2007), we modify this such that the value of the loan at the threshold is not a constant, but simply equal to the recovery value of the loan at the default time. Second, we remove the time dependency in the value of the perpetual debt. It is shown in Carey and Gordy (2007) that under these assumptions, so long as there are positive and fixed cash flows to claimants other than the bank, \( \gamma(1-\lambda) > 0 \) or \( \delta > 0 \), then there exists a finite and positive solution \( \kappa^* \), the optimal foreclosure boundary (and the solution reduces to a 2nd order ordinary differential equation, which can be solved using standard numerical techniques.)

\[7\] Jarrow (2001) also has the advantage of isolating the liquidity premium embedded in defaultable bond spreads.
We model undiversifiable recovery risk by introducing a separate process for recovery on debt, $R_t$. This can be interpreted as the state of collateral underlying the loan or bond. $R_t$ is a geometric Brownian process that depends upon the Brownian motion that drives the return on the firm's assets $Z_t$, an independent Brownian motion $W_t$ and a random instantaneous mean $\alpha_t$: 

$$
\frac{dT_t}{T_t} = \alpha_t dt + \beta \frac{dT_t}{T_t} + \nu \frac{dW_t}{W_t} (3.6);
$$

$$
\frac{d\alpha_t}{\alpha_t} = \kappa \left( \alpha - \alpha \bar{a} \right) dt + \eta dB_t (3.7)
$$

Where the volatility parameter $\beta$ represents the sensitivity of recovery to the source of uncertainty driving asset returns (or the systematic factor), implying that the instantaneous correlation between asset returns and recovery is given by $\frac{1}{dt} \text{Corr}(dA_t, dR_t) = \nu$. On the other hand, the volatility parameter $\nu$ represents the sensitivity of recovery to a source of uncertainty that is particular to the return on collateral, also considered a 'systematic factor,' but independent of the asset return process. The third source of recovery uncertainty is given by (3.7), where we model the instantaneous drift on the recovery rate by an Ornstein-Uhlenbeck mean-reverting process, with $\kappa$ the speed of mean-reversion, $\alpha$ the long-run mean, $\eta$ the constant diffusion term, and $B_t$ is a standard Wiener process having instantaneous correlation with the source of randomness in the recovery process, given heuristically by $\varsigma = \frac{1}{dt} \text{Corr}(dB_t, dW_t)$. The motivation behind this specification is the overwhelming evidence that the mean LGD is stochastic.

Economic LGD on the loan is given by following expectation under physical measure:

$$
\text{LGD}_{	ext{E}}(R, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau) =
$$

$$
= 1 - \frac{\exp(-\gamma \tau)}{B} \left( \text{min} \left( R \exp \left( \alpha - \frac{\beta^2 + \nu^2}{2} \right) + \beta Z_{t+\tau} + \nu W_{T+\tau} \right), 0 \right) (3.8)
$$

Where the modified option theoretic function $B(\cdot)$ is given by:

$$
B(R, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau) =
$$

$$
= \Phi \left( \frac{R - \lambda}{\alpha} \right) - R \Phi \left( -\frac{\alpha}{\alpha} \right) + \exp \left( \left( -\frac{\alpha}{\alpha} \right) \tau \right) \lambda \Phi \left( \frac{\alpha}{\alpha} \right) (3.9)
$$

having arguments to the Gaussian distribution function

$$
\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2} du
$$

$$
\Phi(\gamma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-u^2} du
$$

$$
\frac{d\gamma}{\gamma} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-u^2} du
$$

A well-known result [Bjerksund (1991)] is that the maturity-dependent volatility is given by:

$$
\sigma_t = \left( \beta^2 + \nu^2 + \frac{1}{\kappa^2} \left( \sqrt{\beta^2 + \nu^2 + \frac{1}{\kappa^2}} - \frac{1}{\kappa^2} \right) \right) \left( \frac{T + \tau}{T} \right) + \frac{2\eta}{\kappa} \left( \frac{T + \tau}{T} \right) \left( 1 - e^{-\kappa \tau} \right) + \frac{1}{2} \left( \frac{T + \tau}{T} \right) \left( 1 - e^{-\kappa \tau} \right) (3.11)
$$

The recovery to the bondholders is the expectation of the minimum of the positive part of the difference in the recovery and face value of the loan $[R_{t+\tau} - \lambda \exp(\alpha t)]^+$ and the face value of the bond $B$, which is structurally identical to a compound option valuation problem [Geske (1977)]:

$$
\text{LGD}_{	ext{E}}^B (R, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau) =
$$

$$
= 1 - \frac{\exp(-\gamma \tau)}{B} \left( \text{min} \left( B \exp \left( \alpha - \frac{\beta^2 + \nu^2}{2} \right) + \beta Z_{t+\tau} + \nu W_{T+\tau}, 0 \right), \lambda \exp(\gamma \tau) \right) -
$$

$$
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-u^2} du (3.12)
$$

where $R_{t+\tau} = R_{t+\tau} \exp \left[ \left( \beta^2 + \nu^2 \right) T + \beta Z_{T+\tau} + \nu W_{T+\tau}, \tau \right]$ is the value of recovery on the collateral at the time of resolution. We can easily write down the closed-form solution for the LGD on the bond according to the well-known formula for a compound option: here the ‘outer option’ is a put, and the ‘inner option’ is a call, and the expiry dates are equal. Let $R^*$ be the critical level of recovery such that the holder of the loan is just breaking even:

$$
\lambda \exp(\gamma \tau) = 1 - \text{LGD}_{\text{E}}^B (R^*, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau) (3.13)
$$

where $\tau$ is the time-to-resolution for the loan, which we assume to be prior to that for the bond, $\tau < \tau_B$. Then the solution is given by:

$$
\text{LGD}_{\text{E}}^B (R, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau) =
$$

$$
= 1 - \frac{\exp(-\gamma \tau)}{B} \left( \text{B}(R, \alpha_t | \lambda, c, \gamma, \beta, \nu, \kappa, \eta, \zeta, \tau), \right) (3.14)
$$

having arguments to the Gaussian distribution function

$$
\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2} du
$$

$$
\frac{d\gamma}{\gamma} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-u^2} du
$$

$$
\frac{d\gamma}{\gamma} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-u^2} du
$$

A well-known result [Bjerksund (1991)] is that the maturity-dependent volatility is given by:

$$
\sigma_t = \left( \beta^2 + \nu^2 + \frac{1}{\kappa^2} \left( \sqrt{\beta^2 + \nu^2 + \frac{1}{\kappa^2}} - \frac{1}{\kappa^2} \right) \right) \left( \frac{T + \tau}{T} \right) + \frac{2\eta}{\kappa} \left( \frac{T + \tau}{T} \right) \left( 1 - e^{-\kappa \tau} \right) + \frac{1}{2} \left( \frac{T + \tau}{T} \right) \left( 1 - e^{-\kappa \tau} \right) (3.11)
$$

Where $\Phi_2(X, Y; \rho_{XY})$ is the bivariate normal distribution function for Brownian increments the correlation parameter is given by $\rho_{XY} = (T_X/ T_Y)^{1/2}$ for respective ‘expiry times’ $T_X$ and $T_Y$ for $X$ and $Y$, respectively. Note that this assumption, which is realistic in that we observe in the data that on average earlier default on the bond even if it emerges from bankruptcy or resolve a default at a single time (which in addition is random), is matter of necessity in the log-normal setting in that the bivariate normal distribution is not defined for $\rho_{XY} = (T_X/ T_Y) = 1$ in the case that $T_X = T_Y = \tau$. We can extend this framework to arbitrary tranches of debt, such as for a subordinated issue, in which case we follow the same procedure in order to arrive at an expression that involves trivariate cumulative normal
distributions. In general, a debt issue that is subordinated to the dth degree results in a pricing formula that is a linear combination of d+1 variate Gaussian distributions. These formulae become cumbersome very quickly, so for the sake of brevity we refer the interested reader to Haug (2006) for further details.

**Comparative statics**

In this section we discuss and analyze the sensitivity of ultimate LGD in to various key parameters. In Figures 1 and 2 we examine the sensitivity of the ultimate LGD in the 2-factor model mentioned above, incorporating the optimal foreclosure boundary. In Figure 1, we show the ultimate LGD as a function of the volatility in the recovery rate process attributable to the LGD side systematic factor $\eta$, fixing firm value at default at $V_t = 0.5$. We observe that ultimate LGD increases at an increasing rate in this parameter, that for higher correlation between firm asset value and recovery value return the LGD is higher and increases at a faster rate, and that for bonds these curves lie above and increase at a faster rate. In Figure 2 we show the ultimate LGD as a function of the volatility $\beta$ in the recovery rate process attributable to the PD side systematic factor, fixing LGD side volatility $\nu = 0.5$, for different firm values at default at $V_t = (0.3, 0.5, 0.8)$. We observe that ultimate LGD increases at an increasing rate in this parameter, that for lower firm asset values the LGD is higher but increases at a slower rate, and that for bonds these curves lie above and increase at a lower rate.

**Empirical analysis – calibration of models**

In this section we describe our strategy for estimating parameters of the models for ultimate LGD by full-information maximum likelihood (FIML.) This involves a consideration of the LGD implied in the market at time of default $t_D$ for the ith instrument in recovery segment s, denoted $LGD_{i,s,t_D}$. This is the expected, discounted ultimate loss-given-default $LGD_{i,s,t_E}$ at time of emergence $t_E$ as given by any of our models $m$, $LGD_{s,m}(\theta_{s,m})$ over the resolution period $t_{E,s} - t_{D,s}$.

$$LGD_{s,m}(\theta_{s,m}) = \frac{E_P \left[ LGD_{s,m}(\theta_{s,m}) \right]}{(1 + r_{D,s})^{t_{E,s} - t_{D,s}}} = LGD_{s,m}(\theta_{s,m})$$

(4.1)

Where $\theta_{s,m}$ is the parameter vector for segment s under model m, expectation is taken with respect to physical measure $P$, discounting is at risk adjusted rate appropriate to the instrument $r_{D,s}$ and it is assumed that the time-to-resolution $t_{E,s} - t_{D,s}$ is known. In order to account for the fact that we cannot observe expected recovery prices ex-ante, as only by coincidence would they coincide with expectations, we invoke market rationality to postulate that for a segment homogenous with respect to recovery risk the difference between expected and average realized recoveries should be small. We formulate this by defining the normalized forecast error as:

$$\tilde{\epsilon}_{s,m} = \frac{LGDP_{s,m}(\theta_{s,m}) - LGD_{i,s,t_E}}{LGD_{s,m}(\theta_{s,m}) \times \sqrt{t_{E,s} - t_{D,s}}}$$

(4.2)

This is the forecast error as a proportion of the LGD implied by the market at default (a ‘unit-free’ measure of recovery uncertainty) and the square root of the time-to-resolution. This is a mechanism to control for the likely increase in uncertainty with time-to-resolution, which effectively puts more weight on longer resolutions, increasing the estimate of the loss-severity. The idea behind this is that more information is revealed as the emergence point is approached, hence a decrease in risk. Alternatively, we can analyze $\epsilon_{i,s} = [LGDP_{s,m}(\theta_{s,m}) - LGD_{i,s,t_E}] / LGD_{i,s,t_D}$, the forecast error that is non-time adjusted, and argue that its standard error
is proportional to \((t_i - t_{D_i})^{1/2}\), which is consistent with an economy in which information is revealed uniformly and independently through time [Miu and Ozdemir (2005)]. Assuming that the errors \(\varepsilon_{i,s}\) in (4.2) are standard normal, we may use full-information maximum likelihood (FIML), by maximizing the log-likelihood (LL) function:

\[
(4.3)
\]

This turns out to be equivalent to minimizing the squared normalized forecast errors:

\[
(4.4)
\]

We may derive a measure of uncertainty of our estimate by the ML...
standard errors from the Hessian matrix evaluated at the optimum:

\[
\hat{\theta}^* = \left[ \frac{\partial^2 L}{\partial \theta_{\alpha,\beta} \partial \theta_{\alpha,\beta}} \right]^{-1} \left| \frac{\partial L}{\partial \theta_{\alpha,\beta}} \right|
\]  

\[(4.5)\]

Data and estimation results

We summarize basic characteristics of our dataset in Table 1 and the maximum likelihood estimates are shown in Table 2. These are based upon our analysis of defaulted bonds and loans in the Moody’s Ultimate Recovery (MURD™) database release as of August, 2009. This contains the market values of defaulted instruments at or near the time of default9, as well as the values of such pre-petition instruments (or of instruments received in settlement) at the time of default resolution. This database is largely representative of the U.S. large-corporate loss experience, from the mid 1980s to the present, including most of the major corporate bankruptcies occurring in this period. Table 1 shows summary statistics of various quantities of interest according to instrument type (bank loan, bond, term loan, or revolver) and default type (bankruptcy under Chapter 11 or out-of-court renegotiation). First, we annualized the return or yield on defaulted debt from the date of default (bankruptcy filing or distressed renegotiation date) to the date of resolution (settlement of renegotiation or emergence from Chapter 11), henceforth abbreviated as ‘RDD.’ Second, the trading price at default implied LGD ('TLGD'), or par minus the trading price of defaulted debt at the time of default (average 30-45 days after default) as a percent of par value. Third, our measure of ultimate loss

<table>
<thead>
<tr>
<th>Recovery segment</th>
<th>Parameter</th>
<th>(\alpha) (1)</th>
<th>(\beta) (2)</th>
<th>(\nu) (3)</th>
<th>(\beta_s) (4)</th>
<th>(\nu_s) (5)</th>
<th>(\sigma) (6)</th>
<th>(\eta_{\alpha}) (7)</th>
<th>(\kappa_a) (8)</th>
<th>(\eta_{\nu}) (9)</th>
<th>(\nu) (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revolving credit / term loan</td>
<td>Est.</td>
<td>4.32%</td>
<td>18.63%</td>
<td>18.16%</td>
<td>36.83%</td>
<td>41.06%</td>
<td>19.55%</td>
<td>80.47%</td>
<td>12.82%</td>
<td>3.96%</td>
<td>37.08%</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.5474%</td>
<td>0.9177%</td>
<td>0.7310%</td>
<td>1.3719%</td>
<td>0.4190%</td>
<td>0.0755%</td>
<td>4.2546%</td>
<td>3.2125%</td>
<td>0.9215%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior secured bonds</td>
<td>Est.</td>
<td>5.47%</td>
<td>16.99%</td>
<td>16.54%</td>
<td>30.41%</td>
<td>34.62%</td>
<td>22.83%</td>
<td>77.17%</td>
<td>11.64%</td>
<td>4.40%</td>
<td>33.66%</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.5314%</td>
<td>0.8613%</td>
<td>0.6008%</td>
<td>1.3104%</td>
<td>0.7448%</td>
<td>0.0622%</td>
<td>3.5085%</td>
<td>2.6903%</td>
<td>0.8297%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior unsecured bonds</td>
<td>Est.</td>
<td>6.82%</td>
<td>14.16%</td>
<td>13.82%</td>
<td>24.38%</td>
<td>28.02%</td>
<td>24.30%</td>
<td>75.70%</td>
<td>9.71%</td>
<td>5.50%</td>
<td>28.07%</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.5993%</td>
<td>1.0813%</td>
<td>1.3913%</td>
<td>1.9947%</td>
<td>0.6165%</td>
<td>0.0281%</td>
<td>2.8877%</td>
<td>2.2441%</td>
<td>0.6504%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior subordinated bonds</td>
<td>Est.</td>
<td>8.19%</td>
<td>11.33%</td>
<td>12.02%</td>
<td>17.35%</td>
<td>21.11%</td>
<td>32.43%</td>
<td>67.57%</td>
<td>7.76%</td>
<td>4.42%</td>
<td>22.45%</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.6216%</td>
<td>1.0087%</td>
<td>1.0482%</td>
<td>1.0389%</td>
<td>0.9775%</td>
<td>0.0181%</td>
<td>2.0056%</td>
<td>2.0132%</td>
<td>1.0016%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subordinated bonds</td>
<td>Est.</td>
<td>9.05%</td>
<td>9.60%</td>
<td>10.24%</td>
<td>12.37%</td>
<td>16.06%</td>
<td>25.73%</td>
<td>69.34%</td>
<td>5.97%</td>
<td>3.34%</td>
<td>28.07%</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.6192%</td>
<td>1.0721%</td>
<td>1.0128%</td>
<td>1.0771%</td>
<td>0.9142%</td>
<td>0.0106%</td>
<td>2.049%</td>
<td>2.014%</td>
<td>1.0124%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Value log-likelihood function: -371.09
Degrees of freedom: 1391
P-value of likelihood ratio statistic: 4.69E-03
Area under ROC curve: 93.14%
Hosmer-Lemeshow chi-squared (P-values): 0.63

1 – The volatility of the firm-value process governing default.
2 – The drift of the firm-value process governing default.
3 – The sensitivity of the recovery-rate process to the systematic governing default in (or the component of volatility in the recovery process due to PD-side systematic risk).
4 – The sensitivity of the recovery-rate process to the systematic governing collateral value (or the component of volatility in the recovery process due to LGD-side systematic risk).
5 – The total volatility of the recovery rate process: \(\sqrt{\beta^2 + \nu^2}\)
6 – Component of total recovery variance attributable to PD-side (asset value) uncertainty: \(\frac{\beta^2}{\beta^2 + \nu^2}\)
7 – Component of total recovery variance attributable to LGD-side (collateral value) uncertainty: \(\frac{\nu^2}{\beta^2 + \nu^2}\)
8 – The speed of the mean-reversion in the random drift in the recovery rate process.
9 – The long-run mean of the random drift in the recovery rate process.
10 – The volatility of the random drift in the recovery rate process.
11 – The correlation of the random processes in drift of and the level of the recovery rate process.

Table 2 – Full information maximum likelihood estimation of option theoretic two-factor structural model of ultimate loss-given-default with optimal foreclosure boundary, systematic recovery risk and random drift in the recovery process (Moody’s Ultimate Recovery Database 1987-2009)

---

9 This is an average of trading prices from 30 to 45 days following the default event. A set of dealers is polled every day and the minimum/maximum quote is thrown out. This is done by experts at Moody’s.
severity, the dollar loss-given-default on the debt instrument at emergence from bankruptcy or time of final settlement (ULGD), computed as par minus either values of pre-petition or settlement instruments at resolution. We also summarize two additional variables in Table 1, the total instrument outstanding at default, and the time in years from the instrument default date to the time of ultimate recovery. The preponderance of this sample is made up of bankruptcies as opposed to out-of-court settlements, 1322 out of a total of 1398 instruments. We note that out-of-court settlements have lower LGDs by either the trading or ultimate measures, 37.7% and 33.8%, as compared to Chapter 11’s, 55.7% and 51.6%, respectively; and the heavy weight of bankruptcies are reflected in how close the latter are to the overall averages, 54.7% and 50.6% for TLGD and ULGD, respectively. Interestingly, not only do distressed renegotiations have lower loss severities, but such debt performs better over the default period than bankruptcies, RDD of 37.3% as compared to 28.1%, as compared to an overall RDD of 28.6%. We also note that the TLGD is higher than the ULGD by around 5% across default and instrument types, 55.7% (37.7%) as compared to 51.6% (33.8%) for bankruptcies (renegotiations). Finally, we find that loans have better recoveries by both measures as well higher returns on defaulted debt, respective average TLGD, ULGD and RDD are 52.5%, 49.35% and 32.2%.

In Table 2 we present the full-information maximum likelihood estimation (FIML) results of the leading model for ultimate LGD derived in this paper, the two-factor structural model of ultimate loss-given-default, with systematic recovery risk and random drift (2FSM-SR&RD) on the recovery. The model is estimated along with the optimal foreclosure boundary constraint. The model is estimated with stochastic drift (MSM-SD) are available upon request.

The next set of results concern the random drift in the recovery rate process. The MLE point estimates of the parameter \( \kappa \), the speed of the mean-reversion, is hump-shape in seniority class, ranging from 3.3% to 19.6% from subordinated bonds to senior bank loans. Consequently, more senior instruments not only exhibit greater recovery volatility than less senior instruments, but a larger component of this volatility is driven by the collateral rather than the asset value process.

The model calibration. The MLE point estimates of the parameter \( \mu \), which is the drift of the firm-value process governing default, we observe that estimates are decreasing in seniority class, ranging from 9.1% to 4.3% from subordinated bonds to senior loans, respectively. As standard errors range in 1% to 2%, increasing in seniority rank, these differences across seniority classes and models are generally statistically significant. Regarding the MLE point estimates of the parameter \( \alpha \), which is the drift of the firm-value process governing default, we observe that estimates are increasing in seniority class, ranging from 9.6% to 18.6% from subordinated bonds to loans, respectively. These too are statistically significant across seniorities. The fact that we are observing different estimates of a single firm value process across seniorities is evidence that models which attribute identical default risk across different instrument types are misspecified – in fact, we are measuring lower default risk (i.e., lower asset value volatility and greater drift in firm-value) in loans and senior secured bonds as compared to unsecured and subordinated bonds. A key finding concerns the magnitudes and composition of the components of recovery volatility across maturities inferred from the model calibration. The MLE point estimates of the parameter \( \beta \), the sensitivity of the recovery-rate process to the systematic factor governing default (or due to PD-side systematic risk), increases in seniority class, from 10.2% for subordinated bonds to 18.2% for senior bank loans. On the other hand, estimates of the parameter \( \kappa \), the sensitivity of the recovery-rate process to the systematic factor governing collateral value (or due to LGD-side systematic risk), are greater than \( \beta \) across seniorities, and similarly increases from 12.4% for subordinated bonds to 36.8% for bank loans. This monotonic increase in both \( \beta \) and \( \kappa \) as we move up in the hierarchy of the capital structure from lower to higher ranked instruments has the interpretation of a greater sensitivity in the recovery rate process attributable to both systematic risks, implying that total recovery volatility \( \sigma_R = (\beta^2 + \kappa^2)^{1/2} \) increases from higher to lower ELGD instruments, from 16.1% for subordinated bonds to 41.1% for senior loans. However, we see that the proportion of the total recovery volatility attributable to systematic risk in collateral (firm) value, or the LGD (PD) side, is increasing (decreasing) in seniority from 59.3% to 80.5% (40.7% to 19.6%) from subordinated bonds to senior bank loans. Consequently, more senior instruments not only exhibit greater recovery volatility than less senior instruments, but a larger component of this volatility is driven by the collateral rather than the asset value process.
We conclude this section by discussing the quality of the estimates and model performance measures. Across seniority classes, parameter estimates are all statistically significant, and the magnitudes of such estimates are in general distinguishable across segments at conventional significance levels. The likelihood ratio statistic indicates that we can reject the null hypothesis that all parameter estimates are equal to zero across all ELGD segments, a p-value of 4.7e-3. We also show various diagnostics that assess in-sample fit, which show that the model performs well-in sample. The area under receiver operating characteristic curve (AUROC) of 93.1% is high by commonly accepted standards, indicating a good ability of the model to discriminate between high and low LGD defaulted instruments. Another test of discriminatory ability of the models is the Kolmogorov-Smirnov (KS) statistic, the very small p-value 2.1e-8 indicating adequate separation in the distributions of the low and high LGD instruments in the model.11 We also show 2 tests of predictive accuracy, which is the ability of the model to accurately quantify a level of LGD. The McFadden pseudo r-squared (MPR2) is high by commonly accepted standards, 72.1%, indicating a high rank-order correlation between model and realized LGDs of defaulted instruments. Another test of predictive accuracy of the models is the Hoshmer-Lemeshow (HL) statistic, high p-values of 0.63 indicating high accuracy of the model to forecast cardinal LGD.

**Downturn LGD**

In this section we explore the implications of our model with respect to downturn LGD. This is a critical component of the quantification process in Basel II advanced IRB framework for regulatory capital. The Final Rule (FR) in the U.S. [OCC et al. (2007)] requires banks that either wish, or are required, to qualify for treatment under the advanced approach to estimate a downturn LGD. We paraphrase the FR, this is an LGD estimated during an historical reference period during which default rates are elevated within an institution’s loan portfolio. In Figures 3 we plot the ratios of the downturn LGD to the expected LGD. This is derived by conditioning on the 99.9th quantile of the PD side systematic factor in the model for ultimate LGD. We show this for loans and bonds, as well as for different settings of key parameters (\(\sqrt{\eta} \), \(\eta \), \(\eta_{a} \)) in the plot, with other parameters set to the MLE estimates. We observe that the LGD mark-up for downturn is monotonically declining in ELGD, which is indicative of lower tail risk in recovery for lower ELGD instruments. It is also greater than unity in all cases, and approaches 1 as ELGD approaches 1. This multiple is higher for bonds than for loans, as well as for either higher PD-LGD correlation \(\sqrt{\eta} \) or collateral specific volatility \(\nu \), although these differences narrow for higher ELGD.

**Model validation**

In this final section we validate our model, in particular, we implement an out-of-sample and out-of-time analysis, on a rolling annual cohort basis for the final 12 years of our sample. Furthermore, we augment this by resampling on both the training and prediction samples, a non-parametric bootstrap [Efron (1979), Efron and Tibshirani (1986), Davison and Hinkley (1997)]. The procedure is as follows: the first training (or estimation) sample is established as the cohorts defaulting in the 10 years 1987-1996, and the first prediction (or validation) sample is established as the 1997 cohort. Then we resample 100,000 times with replacement from the training sample the 1987-1996 cohorts and for the prediction sample 1997 cohort, and then based upon the fitted model in the former we evaluate the model based upon the latter. We then augment the training sample with the 1997 cohort, and establish the 1998 cohort as the prediction sample, and repeat this. This is continued until we have left the 2008 cohort as the holdout. Finally, to form our final holdout sample, we pool all of our out-of-sample resampled prediction cohorts, the 12 years running from 1997 to 2008. We then analyze the distributional properties (such as median, dispersion, and shape) of the two key diagnostic statistics: the Spearman rank-order correlation for discriminatory (or classification) accuracy, and the Hoshmer-Lemeshow chi-squared P-values for predictive accuracy, or calibration.

Before discussing the results, we briefly describe the two alternative frameworks for predicting ultimate LGD that are to be compared to the 2-factor structural model with systematic recovery and random drift (2FSM-SR&RD) developed in this paper. First, we implement a full-information maximum likelihood simultaneous equation regression model (FIMLE-SERM) for ultimate LGD, which is an econometric model built...
upon observations in MURD at both the instrument and obligor level. FIMLE is used to model the endogeneity of the relationship between LGD at the firm and instrument levels in an internally consistent manner. This technique enables us to build a model that can help us understand some of the structural determinants of LGD, and potentially improve our forecasts of LGD. This model contains 199 observations from the MURD™ with variables: long term debt to market value of equity, book value of assets quantity, intangibles to book value of assets, interest coverage ratio, free cash flow to book value of assets, net income to net sales, number of major creditor classes, percentage of secured debt, Altman Z-Score, debt vintage (time since issued), Moody’s 12 month trailing speculative grade default rate, industry dummy, filing district dummy and pre-packaged bankruptcy dummy. Detailed discussion of the results can be found in Jacobs and Karagozoglu (2010). The second alternative model we consider addresses the problem of non-parametrically estimating a regression relationship, in which there are several independent variables and in which the dependent variable is bounded, as an application to the distribution of LGD. Standard non-parametric estimators of unknown probability distribution functions, whether conditional or not, utilize the Gaussian kernel (Silverman (1982), Hardle and Linton (1994), and Pagan and Ullah (1999)). It is well known that there exists a boundary bias with a Gaussian kernel, which assigns non-zero density outside the support on the dependent variable, when smoothing near the boundary. Chen (1999) has proposed a beta kernel density estimator (BKDE) defined on the unit interval [0,1], having the appealing properties of flexible functional form, a bounded support, simplicity of estimation, non-negativity, and an optimal rate of convergence n^{-4/5} in finite samples. Furthermore, even if the true density is unbounded at the boundaries, the BKDE remains consistent [Boeuzmarni and Rolin (2001)], which is important in the context of LGD, as there are point masses (observation clustered at 0% and 100%) in empirical applications. Detailed derivation of this model can be found in Jacobs and Karagozoglu (2010). We extend the BKDE [Renault and Scalliet (2004)] to a generalized beta kernel conditional density estimator (GBKDE), in which the density is a function of several independent variables, which affect the smoothing through the dependency of the beta distribution parameters upon these variables.

Results of the model validation are shown in Table 3. We see that while all models perform decently out-of-sample in terms of rank ordering capability, FIMLE-SEM performs the best (median = 83.2%), the GBKDE the worst (median = 72.0%), and our 2FSM-SR&RD in the middle (median = 79.1%). It is also evident from the Table and figures that the better performing models are also less dispersed and exhibit less multi-modality. However, the structural model is closer in performance to the regression model by the distribution of the Pearson correlation, and indeed there is a lot of overlap in these. Unfortunately, the out-of-sample predictive accuracy is not as encouraging for any of the models, as in a sizable proportion of the runs we can reject adequacy of fit (i.e., p-values indicating rejection of the null of model fit at conventional levels). The rank ordering of model performance for Hoshmer-Lemeshow p-values of test statistics is the same as for the Pearson statistics: FIMLE-SEM performs the best (median = 24.8%), the GBKDE the worst (median = 72.0%), and our 2FSM-SR&RD in the middle (median = 39.1%); and the structural model developed herein is comparable in out-of-sample predictive accuracy to the high-dimensional regression model. We conclude that while all models are challenged in predicting cardinal levels of ultimate LGD out-of-sample, it is remarkable that a relatively parsimonious structural model of ultimate LGD can perform so closely to a highly parameterized econometric model.

Conclusions and directions for future research
In this study we have developed a theoretical model for ultimate loss-given-default, having many intuitive and realistic features, in the structural credit risk modeling framework. Our extension admits differential seniority within the capital structure, an independent process representing a source of undiversifiable recovery risk with a stochastic drift, and an optimal foreclosure threshold. We also analyzed the comparative statics of this model. In the empirical analysis we calibrated the model for ultimate
LGD on bonds and loans, having both trading prices at default and at resolution of default, utilizing an extensive sample of agency rated defaulted firms in the Moody’s URD™. These 800 defaults are largely representative of the U.S. large corporate loss experience, for which we have the complete capital structures, and can track the recoveries on all instruments to the time of default to the time of resolution. We demonstrated that parameter estimates vary significantly across recovery segments, finding that the estimated volatilities of the recovery rate processes and their random drifts are increasing in seniority; in particular, for 1st lien bank loans as compared to senior secured or unsecured bonds. Furthermore, we found that the proportion of recovery volatility attributable to the LGD-side (as opposed to the PD-side) systematic factors to be higher for more senior instruments. We argued that this reflects the inherently greater risk in the ultimate recovery for higher ranked instruments having lower expected loss severities. In an exercise highly relevant to requirements for the quantification of a downturn LGD for advanced IRB under Basel II, we analyzed the implications of our model for this purpose, finding the later to be declining for higher expected LGD, higher for lower ranked instruments, and increasing in the correlation between the process driving firm default and recovery on collateral. Finally, we validated our model in an out-of-sample bootstrapping exercise, comparing it to two alternatives, a high-dimensional regression model and a non-parametric benchmark, both based upon the same MURD data. We found our model to compare favorably in this exercise. We conclude that our model is worthy of consideration to risk managers, as well as supervisors concerned with advanced IRB under the Basel II capital accord. It can be a valuable benchmark for internally developed models for ultimate LGD, as this model can be calibrated to LGD observed at default (either market prices or model forecasts, if defaulted instruments are non-marketable) and to ultimate LGD measured from workout recoveries. Finally, risk managers can use our model as an input into internal credit capital models.

References

- Basel Committee on Banking Supervision, 2005, Guidance on paragraph 468 of the framework document, BIS, July
- Basel Committee on Banking Supervision, 2006, International convergence on capital measurement and capital standards: a revised framework. BIS, June
- Finger, C., 1999, “Conditional approaches for CreditMetrics™ portfolio distributions,” CreditMetrics™ Monitor, April
The Capco Institute Journal of Financial Transformation
Empirical Implementation of a 2-Factor Structural Model for Loss-Given-Default

• Frye, J., 2000a, "Collateral damage," Risk, April, 91-94
• Frye, J., 2000c, "A false sense of security," Risk, August, 53-57
• Geske, R., 1977, "The valuation of corporate liabilities as compound options," Journal of Financial and Quantitative Analysis, 12, 541-552,
• Keesman, D., and K. van de Casteele, 2000, "Suddenly structure mattered: insights into recoveries of defaulted debt," Corporate ratings, commentary, Standard and Poors
• Pikhrt, M., 2003, "Unexpected recovery risk," Risk, August, 74-78
• Vasicek, O.A., 1984, "Credit valuation," KMVä Corporation, March
Regulatory Reform: A New Paradigm for Wealth Management

Haney Saadah — Principal Consultant, Capco
Eduardo Diaz — Senior Consultant, Capco

Abstract
The aim of this paper is to assess some of the key implications of the current regulatory proposals. Whilst the scope and impact of these regulatory changes are global, we have focused on two of the most prominent markets, namely the U.K. and the U.S. Their status means that any significant changes may not only be amongst the most visible, but if implemented correctly, should provide a blueprint from which other financial centers can reference.

1 We would like to thank Christine Ciriani for help kind help and advice with this paper. The comments made in this article are solely the responsibility of the authors and in no way representative of the views of Capco or its partner organizations.
Whilst the financial crisis has undoubtedly pushed regulation to the forefront of banking once again, it is by no means the only motivation. Industry insiders have long suggested that private individuals have needed the benefit of reform; especially around distribution of advice. This long held belief in the need for greater transparency and consumer protection is driving regulatory reform towards the elimination of compensation structures which are believed to deter financial advisors from providing impartial advice. Hence, financial regulation appears to be most focused on going after commissions.

Regulatory reform cannot occur without widespread impact on the industry. For the most part, the extent of the proposed regulatory reform will embrace investment advisory firms which have until now generated their revenues from ‘commission’ payments. These firms are clearly poised to face structural reforms over the years to come. Other firms in the industry which provide advice on a transparent fee basis will be subject to less regulatory reform as their advice model already falls in line with what regulators are proposing.

Two of the largest and most active wealth management markets exposed to regulatory overhaul are the U.S. and the U.K. Regulators on both sides of the Atlantic are pushing ahead with reforms to reshape investment advisory and revamp investor protection. In the U.S., broker-dealers have taken the spotlight as they have been exempt from SEC registration, and have provided investment advice under light supervision. In the U.K., regulation is also targeting commission-based advisors in an effort to remove conflicts of interest between client needs and advisor’s compensation. In both cases, reform is targeting the same concern: are client interests compromised when investment advice is paid for in the form of commissions?

The often conflicting ideology between commissions versus fee-based advice has persisted for some time. Independent professional groups have consistently questioned the rationale behind industry practitioners who label their services as investment advice whilst earning on commission. These groups argue that such labels mislead investors and diminishes the true meaning of advice, namely an impartial advice driven service, transparent, client-centric and devoid of any product bias. Now that future regulatory proposals aim to restrict the right of advisors to earn commission, it is a foregone conclusion that many key industry players must significantly adapt their model in order to remain in the market.

As many firms undergo transformation towards a fee-based model and a stricter regulatory regime, there is no doubt that technology can become a major driving force behind this transformation. The current environment is already a challenging one for the industry: increasing cost pressures and falling revenues are already causing major concerns for senior managers in wealth management. This provides a dilemma that will be often revisited in this paper, more specifically that of the transition from a commission- to a fee-based model. Such transition may simply price many clients out, with the subsequent effect of driving many smaller firms out of the market.

**Financial reform is moving forward**

In the U.S., the move towards a uniform fiduciary standard cannot unfold without significant challenges for the wealth management industry. One such initiative has resulted in the creation of the Dodd-Frank Financial Reform Act, which was signed into law on the 21st of July, 2010. This Act is intended to bring stability and order to a financial system that is unable to withstand another round of excessive risk taking and bank bailouts.

Whilst the implications of this Act are far-reaching for the entire financial system, the direct impact on the wealth management industry may seem less obvious at first glance. One of the provisions targeting the wealth management industry is the call for the SEC to conduct a six-month study of uniform fiduciary duty for anyone providing personalized investment advice to clients. The study will formulate the basis for determining whether or not the investment community stands to benefit from a uniform fiduciary standard and how viable this option may be for the industry. Since independent investment advisors and financial planners already abide by the fiduciary duty and are required to register with the SEC, the adoption of a uniform fiduciary standard in the industry will primarily impact broker-dealers, which have until now been subject to a less stringent and lighter ‘suitability’ standard under the custody of the National Association of Securities Dealers (NASD). The immediate outcome will be to shift broker-dealers under the supervision of the SEC for fiduciary duty oversight, hence placing them on a leveled playing field along with independent investment advisors and financial planners.

**While Europe restructures, London looks to reassert itself by reinventing the model**

Europe is still in the midst of reorganization, compounded by macro-level uncertainty of its member states. Excessive budget deficits in the weaker states exposed the fragility of the Union and gave rise to the inevitable remedy of fiscal tightening. As markets cheer fiscal discipline, many individual investors remain distrustful of financial markets and have become concerned about an economy which has been undermined by austerity measures at a time of anemic growth. This has led the wealth management industry to experience an unprecedented transformation in the way investors approach investing and perceive the true value of advice. Being naturally conservative, European investors are much more interested in long-term planning than risk taking. Investors, for their part, are unquestionably one of the leading voices in reshaping the distribution of advice in this once conservative continent.

As international pressures mount, traditional safe havens such as Switzerland are beginning to experience a shakeup of their model. This is
leading safe havens to gradually lose their allure among investors from the developed world amid their governments' fiscal hunt for undeclared accounts, thereby placing them on a more leveled playing field with other financial centers. As competition intensifies, traditional safe havens are slowly reinventing themselves towards an integrated value offering led by an advice-driven model, and away from bank secrecy as the core proposition.

London, the other large wealth management center in Europe, is now at the centre stage of financial reform. The Financial Services Authority (FSA) in the U.K. is nearing the final stages of its proposed Retail Distribution Review (RDR). This is a far-reaching Act that aims to strengthen investor protection by redefining the distribution of advice and how it is remunerated. In essence, the RDR proposes to withdraw commissions. It is widely expected that the impact of such reforms will be predominantly borne by the mass-affluent segment. The high net worth clients, accustomed to fees for professional services, will, it is anticipated, revert to fee-based advice with limited aggravation.

Financial reforms on both sides of the Atlantic target commission business
Moving from a commission to a fee based model will mean that greater resources need to be allocated to the clients.

This, in turn, will require commission-based market players to reflect in their pricing and newly established fees the cost of such additional dedicated resources. While wealthier client segments may not lose out from these changes as they may easily afford the fees, the mass-affluent segment could clearly be priced out. The outcome will likely place traditional advice out of the reach of mass-affluent clients, as the cost of advice relative to their investable assets may erode any additional return the client may perceive from a customized investment policy.

The question, therefore, is: where do mass-affluent investors go? As this segment becomes slowly priced out of investment advice, the obvious route for them will be to turn to either self-directed investments or the traditional retail banks. Whether this is in their best interests or not will depend on client needs. The commission allows the investor to pay for advice indirectly, but what remains open for debate is the quality of this advice and whether investors are better off by renouncing altogether from commission-based advice. As is the case in the U.S., where financial reform is evaluating the broker-dealer business, and in the U.K., where the FSA wants to force investment advisors to charge fees, the objective is to provide clarity to the investor, and remove product bias by the firm.

The implications of moving away from commissions
Commission-based business does not bode well for fiduciary duty
In the U.S., the proposed reform for broker-dealers is a positive step forward towards addressing quality in the investment profession. In the minds of investors the term ‘investment advice’ does not carry a precise definition. When broker-dealers claim to provide investment advice, they should do so in a holistic approach to client advisory. Investment advice cannot be associated with individual product sales, without taking into account the overall client view. The same principle applies to any financial advisor who mainly earns his compensation via commissions, as it is often the case in the U.K. The problem of earning through commissions in the investment profession lies in the advisor’s alignment. Whilst the client always comes first in any advisory relationship, the commission compensation model may induce the advisor to look after the wrong side of the relationship, that is, the employer or third party provider. Furthermore, it could be argued that commission compensation does nothing more than encourage short-termism. Consequently, to safeguard the concept of advice, compensation models must align the clients’ interests with those of the advisor, so that the advice is impartial and responds to the principle of fiduciary duty.

The unfolding landscape in the wealth management industry is still uncertain. In the U.S., the impact of regulation cannot be fully measured until the SEC discloses how it plans to regulate broker-dealers. Similarly, the FSA in the U.K. is fine tuning final provisions in its ‘retail distribution review.’ Regardless of the specificities of the final outcome, recent reform initiatives demonstrate that regulators are becoming increasingly concerned about commission-based advice. Reform will bring about change, and hence the wealth management industry is poised to evolve over the years to come.

What does this mean for commission-based businesses? Regulation can lead to two possible outcomes for financial advisors whose compensation is based on commission. On the one hand, financial advisors and broker-dealers who earn commissions may be obliged to stop providing investment advice and instead focus on transactional services for self-directed investors. This is because this group of firms could not assert
A uniform fiduciary duty standard will alter the competitive landscape

Once regulation becomes effective, the natural path for commission-driven advice will be to gradually shift to a fee-based model. To continue to draw revenues from commissions whilst providing ‘financial advice’ will clash with forthcoming regulation. First, keeping their current operating model will most certainly give way to increased litigation, as advisors will naturally attempt to maximize commission revenues at the possible expense of the client. Such a model will only serve to dissuade the advisor from providing the best advice. Second, the post-Lehman financial crisis left a pronounced footprint on trust; the founding principle of financial advice. To win clients’ trust back, advisors who have until now earned on commission will need to put the client center stage once again. This means advice will no longer be tailored to maximize commission revenues, but rather designed to respond to clients’ financial planning needs.

Slowly but surely, this will lead to a fiercer competitive landscape as market players search for the right service offering based on key principles defined around proposition, driven by cost and value.

- Brokerage – institutions offering brokerage services normally operate as a business segment of a large financial group; although in the current regulatory environment independent brokerage businesses coexist among larger players. Their service offering focuses on the distribution of individual financial products, and for the most part revenue is generated through commissions. Brokerage firms commonly target the retail/mass-affluent end of the market. In the U.S., the classic broker-dealers exist, whether independent or part of a large financial institution. In the U.K. brokers provide, through advisory or execution-only, a commission-based service. The traditional single product provider life assurance salesperson also exists through local or bank networks offering products in return for a commission payment. We anticipate that many of the changes will come to this market because it is dependent on commission as the main driver, with a customer base unlikely to accept a fee based approach.

- Wealth/investment advisory – wealth investment advisors work either independently or for a financial institution. They tend to target the higher end affluent part of the market and receive compensation for giving advice on investing in stocks, bonds, mutual funds, exchange traded funds, or alternative investments based on a portfolio management approach. Their compensation is normally based on a flat fee or on a percentage of assets under management. Therefore, it does not necessarily vary in response to portfolio turnover or type of product sold, although in some cases, as in when the advisor belongs to a large financial institution, advice may be paid for in the form of both fees and commissions. In the U.S., investment advisors are subject to fiduciary duties, hence advice is perceived to be more independent. In the U.K., private client investment management firms provide individuals with classic discretionary or advisory portfolio management services. These services may often be provided by either a boutique/independent investment manager or a large bank offering this type of service. They may often work in tandem with financial advisors who may provide the initial introductions along with financial planning, structure, and relationship support. This model should be fully supported through a fee-based approach charging for investment advice.

- Open architecture/whole of market – this service offering targets a higher end segment and should assess every facet of the client’s needs, including, but not limited to full financial planning, investments, taxes, savings, insurance, and estate planning. They are largely compensated on a fee basis. At present the U.S., U.K., and Europe have offered this type of service through high end private banks and boutiques. They tend to focus on the wealthiest clients who seek and prefer the fee-based model. The quality of advice should be high to reflect the breadth of offering and the sophistication of potential clients. While regulation could essentially alter business and operational practices of this market, we see this type of advice driving the new wealth management client that seeks the optimum service in lieu of managing their own affairs.

Up until now commission-based advice allowed firms to serve a wider client segment based on a presumably ‘more affordable’ value proposition (Figure 1). As looming regulation and battered investors’ trust advocate change for commission-based advice, some of these firms will start leaping closer to the traditional investment advisory business model (Figure 2). In fact, regulation is setting the stage for a uniform set of rules which will do nothing more than intensify competition in the upper client segments, whilst potentially pricing the mass-affluent segment out. Consequently, as competition intensifies in this area, differentiation factors which have long persisted in the industry could end up commoditizing. Service offering commoditization in the wealth management space will give entry to a new set of rules which will fuel transformation in the quest for the right value proposition.
The reassessment of the value proposition to the client is leading to changes in client segmentation and service offerings. Stricter client segmentation will be inevitable as enhanced regulation and intensifying competition place pressures on profits. In such cases, the segments prone to greater suffering are those unable to cope with price hikes, which in this case happen to be the segments at the bottom of the pyramid: retail and mass-affluent. As these segments become priced out of investment advice, the remaining segments will be the target of choice by firms chasing higher margins. Success in this new competitive landscape can only be accomplished through revamped and tailored service offerings to create a superior value proposition.

As pressure mounts in the competitive landscape and the quality of investment advisory services improve, there will be some advisors which lack the depth and breadth to rival more established firms. These advisors may be able to run a profitable business today with predominantly retail and mass-affluent clients. However, in the absence of commission revenues, these advisors may not be able to switch to a fee-based and personalized service scheme as their existing client base may not be able, and/or willing, to put up with fees. As mentioned earlier, the cost of advice for the lower segments may just erode the additional return the client may expect to obtain from upgraded service. The path will not be easy unless the right formula is developed in order to provide personalized investment advice for a lower cost. The question here can also be expanded to ask how much traditional propositions will be affected by regulation. In the U.K. it is widely anticipated that the mass affluent market will revert to a fund-based approach. This type of offering is a natural fit for mass affluent clients who often seek straightforward, conservative solutions to their needs. The logic here is simple, a variety of well chosen funds, levying an annual charge, that are held in a client portfolio long term. Funds, as a form of advice, have often been sold with the key benefit of ‘buying expertise.’ Fund-based businesses also allow the client advisor to focus on relationship management as a core driver, as opposed to consistent maintenance of an advisory portfolio of varying assets. Fees will be collected via annual management charges (AMC). This model suits large banks, or indeed any large firm which could manufacture/sell funds and collect fees for maintenance and client reporting on a large scale. Traditional fund-based sales have enjoyed commission incentives being paid to the advisor for selling the fund. New regulation looks set to challenge this by abolishing kickbacks from fund managers in favor of charging a single, straightforward overall AMC for providing the service of buying, selling, researching, and rebalancing these funds on behalf of the clients.

At the high net worth end, the space is open for ‘open architecture’ to take hold. A fee offered by clients in exchange for impartial service where the right product is offered, dependent solely on client need. Without commission as a driver the fee tariff remains consistent regardless of product recommended. Ultimately, the fee must be of sufficient premium to the advisor to provide appropriate margin. Investment professionals operating at the higher end of the market should find this transition to be easier, since their clients are traditionally more accustomed to paying fees for a service.

Whilst traditional asset classes revolve around aforementioned funds and also stocks and bonds, the expanding alternatives universe presents a useful addition to an open architecture model, in particular ETF’s. Where traditional funds may be structured as a unit trust or open ended investment companies, an ETF offers share-like liquidity with the option of covering an entire index or market, such as FTSE 100 or oil. Consequently, you achieve the breadth of coverage that a unit trust/mutual fund offers with the accessibility and convenience of a share. The concept has
proven highly popular in the U.S, with growing awareness in the U.K, and we anticipate a consistent increase in the utilization of this product in the years to come.

Regulatory reform will unquestionably also drive transformation. Part of this transformation will involve, among other things, upgrading of technology and reshaping change in the way financial advisors approach and advise clients. By switching clients to a fee-based contract, it is likely that the lower end segment may simply be priced out of investment advisory, whilst the higher segments may shift away searching for better value propositions or agree to pay the fees. Retail clients will most likely be driven to a self-directed investment model. The affluent segment, meanwhile, will remain in a grey zone (Figure 3). The challenges for the industry can, therefore, be of internal and external nature. Internally, alignment with fiduciary duties will call for business, technological, and cultural change. Externally, the additional costs emanating from this alignment may just simply price clients out.

Technology as the driving force of transformation

Technology at the cornerstone of the industry

Transformation will undeniably become a common theme for the wealth management industry over the years to come. On the one hand, recent market turmoil did not only leave a significant footprint in the financial markets, but it also altered clients’ behavioral responses to investing. On the other hand, regulation is evolving towards creating a fair, transparent, and efficient market place. These two forces will lead to change, and together with it, technology will emerge as a key driver for wealth management. Consequently, it is expected that technology will enable the industry to generate gains in quality, productivity, and compliance.

While it is widely anticipated that large global organizations will increase their technology expenditures as a matter of course, the challenge for smaller market participants is how much resources they will have to allocate to this area. This remains a current area of concern for some wealth managers especially those which have fallen behind the ‘technology curve’ and now face industry transformation with outdated systems that are beyond the scope of modern demands. The lack of priority in technology expenditures for some of these firms during the ‘expansionary’ years seems at odds with current management ethos. The need to aggressively increase technology investments to enhance client and advisor experience, as well as regulatory compliance, is widely acknowledged. The estimated technology investments should be relative to the scale and size of the organization, however, larger organizations will benefit from size synergies, thereby placing smaller players at a cost disadvantage. The depth and breadth of required investments may just force smaller competitors out. Overall, nearly all market participants expect that technology will represent a considerable portion of their medium-term costs.

We anticipate that niche technologies and processes in areas such as hedge funds, derivatives, risk management, and client reporting will be a top priority for the global players in the coming years (Figure 4). This will be followed by a more conservative move towards targeted operating models designed to encompass major regulatory changes such as the RDR. A third stage will be to evaluate and analyze more long-term cost-effective solutions, such as cloud computing. In principle, cloud computing offers the smaller end of the market considerable technology saving potential, as servers and software would be handled and upgraded by a third party. Issues around data protection rule out ‘public clouds’ for the larger banks. However, ‘private cloud’ offering cloud usability with enhanced security, such as on site servers, are being reviewed by large banks, seeking accessibility, efficiency, and cost savings for their front office. Overall, however, those firms seeking long term competitiveness must commit sustainable resources to ongoing technology transformation in order to remain both ‘cutting edge’ and compliant.

Figure 3 – Wealth management segmentation needs and profitability impact under proposed regulation

Figure 4 – Anticipated wealth management technology spend priority over the next five years
Technology can help drive transformation and allow for effective client segmentation

While technology investments often serve to help either increase revenues or reduce costs, or both, regulatory concerns will in no doubt play a crucial role in this technology transformation. Increased demands will be placed on processing, client reporting, and advice distribution. None of these can be effectively delivered without the right technology processing. In fact, firms which invest heavily in technology infrastructure will not only comply with regulatory requirements, but will also build the capability to fine tune their offerings to a wider spectrum of client segments. With the proper client segmentation and selection of distribution channels, firms which largely depend on commission revenues may well be able to serve their clients profitably, at fee schedules adapted accordingly to each segment. Consequently, for client segments to remain profitable, it is imperative for firms to formulate and dedicate the right mix of distribution channels to each segment. For example, whilst high net worth clients require integrated and personalized financial planning, affluent clients have life cycle needs, which may be effectively provided through mass-customized financial advice. In both cases, technology can provide the right level of support to ensure:

- **Quality** – technology can allow for the integration of investment advisory processes, order management, and reporting, hence bolstering the quality of advice, order processing, and performance measurement. Firms can disseminate rules, research, and investment strategies used in the investment advisory process in an efficient and quality-driven manner.

- **Productivity** – technology can enable the roll out of investment advisory processes designed to aid the advisor in formulating an investment policy for the client. The investment advisory process is supported by quantitative models, market data, investment strategy, portfolio rules, order processing, and portfolio performance measurement; thus allowing the advisor to run advanced risk profiling aimed at responding to the financial needs of the client. With the aid of technology, the advisor can dedicate more time to financial dialogue and relationship building.

- **Compliance** – technology can allow the firm to comply with regulatory requirements. By maximizing quality and discipline in the investment advisory process, the firm already demonstrates transparency in the investment process whilst ensuring that the investment proposal effectively responds to client’s objectives. This will include risk tolerance, horizon, preferences, and special circumstance, such as investment preferences and specific client exclusions (i.e., stock selection or geography).

The benefits cannot but support the case for investment in technology. Firms which decide to confront financial reform without significant investments will lose out in the new competitive landscape. In the particular case of investment advisory, where fiduciary duty gains relevance in the U.S. and impartiality through the elimination of commission in the U.K., firms will not be able to serve and maintain all client segments at current pricing levels. When and if proposed financial reform becomes effective, the impact will fall upon affluent clients, which can swiftly turn into an unprofitable segment for many of these firms. However, technology can help mass-customize the distribution of advice, so that loss-making segments can turn profitable. It is important to recognize that the financial needs of this segment are mainly driven by life-cycle needs, which can be easily fulfilled through ample, but standardized, investment solutions offering some degree of flexibility. Whilst some large organizations have already opted in for this option, there are still a number of firms which do not have the channel capability to serve different client segments, and therefore adjust the offering accordingly. Figure 5 highlights the different client segments with respect to behavioral traits, and how this may lead to choosing the right mix of channels and service offerings.

The key trend here is that wealthier clients will undoubtedly be the target of advice-driven service. They will be the clients most likely to pay the fees for the quality service. Broad range wealth managers catering for the complete client segmentation market will need technology that can adapt to the broad mix of clients. Product driven mass affluent clients will generally need a smaller scope of products, but with a higher concentration of technology processing as the numbers are greater.

The Capco Institute Journal of Financial Transformation

Figure 5 – Choosing the right channel mix for distributing financial advice to different client segments
The higher net worth clients require a broader scope and range of products, but the focus will be on service. There are clearly less of these clients globally as compared to retail. Products will need to be adaptable, flexible, and more unique. This contrasts sharply with the more homogeneous product offered at the retail end of the market.

The vast rise in wealth across the globe combined with the clear need for wealth managers to update systems has meant technology vendors are growing both in number and capability. It is also clear that financial planning services at the retail/affluent end will see some of the greatest technology spend in the near future. Product driven services, automated in approach and with higher transactions levels that technology must control effectively.

The higher net worth market will find greater demand for technology to enhance their portfolio management services. Quality front-end systems offering the complete portfolio optimization and reporting function.

Conclusion

While no one can deny the implicit importance of regulation in ensuring investor confidence and institution prudence, we must also be aware that regulation alone will not prevent crises. It is our contention that one of the single and most influential factors in determining market prudence should be client demand. At present it is clear clients are reluctant to support wealth managers unless they prove their value. The wealth managers holding the key advantage will be those clearly demonstrating an ‘impartial’ service, with a tangible value, that would encourage clients away from self-managing. To achieve this, firms need the right balance of quality relationship managers, technology, infrastructure, and training. Where regulation seeks to clarify some of these areas through oversight and legislation, it cannot compel wealth managers to provide service excellence. This can only be achieved through market advantage, client acquisition, retention, and satisfaction. This generates the profitability that will be the engine for change. For many clients the recent financial crisis was a step too far, and client inertia has been replaced by a marked sensibility that self-managing is the best approach, unless the firm can demonstrate real value. This has benefitted the boutiques banks immensely as they have traditionally worked on long standing client relationships and investment conservatism. The challenge has been regaining the ‘trusted advisor’ status for the big banks. However, regulation along with prudent internal reorganization and investment should see big banks regain this status in time. The concern lays more with some of the boutiques and many of the small one-man bands; already struggling with current regulatory costs.

Indeed, as we have already stated, the industry is handling significant levels of regulation, which, if increased, will force those smaller market participants to leave the market altogether. Two of the biggest financial markets, the U.K. and U.S., have undoubtedly been the recipients of most of the new wave of private client regulations. This is natural, of course, as they have the most to lose if it goes wrong. Overseas clients have traditionally favored these two financial centers due to their perceived security and sophistication. Factors such as infrastructure, safety, and regulation have been pivotal in attracting new funds. To damage this trust is to negate the rationalization for investing in the first place. However, diversity of the market and a strong competitive environment are also vital factors. Innovation and progress are all signs of a vibrant financial market. Small firms need to compete alongside large players to give the diversity that has made the City of London and Wall Street such vital parts of the global financial community. Whilst lessons learnt from the past highlight regulatory failings, and therefore the need for change, the warning signs for the future show us too much regulation may provide the final straw. For many market players, they may now become ‘too small to survive.’ Our challenge is to develop ways to help as many of these firms survive as possible. It is certainly a challenge worth undertaking, as many of these firms are renowned for their expertise in client servicing and innovation, and we all have a lot to learn from them.
Abstract
In this paper I discuss modern banks’ risk profile in the aftermath of the 2008 financial crisis in light of regulators’ responses, markets evolution, and stakeholders’ reactions. I then suggest risk mapping as a key analysis tool for top management to overcome the shortcomings of traditional risk assessment highlighted by the crisis. Within this approach, I present an alternative way to map risks according to severity of impact and potential management actions. Finally, I develop a risk reporting structure geared towards providing top management with a holistic picture of their company’s risk profile based on a combination of historical data, key risk indicators, expert opinion, and creative hypotheses.

Sergio Scandizzo — Head of the Operational Risk Unit, European Investment Bank

1 The views expressed in this article are those of the author and do not necessarily reflect those of the European Investment Bank.
“Risk ain’t what it used to be,” Yogi Berra would probably say if he worked in risk management. The worst financial crisis since 1929, a new, and still changing regulatory environment, a wave of public distrust in the financial system coupled, within the financial world itself, with a dwindling faith in the traditional models and tools that had looked so successful only a few years back have left risk managers feel very much like Alice in the Pool of Tears, wondering if they are the same as yesterday or if instead they have been changed during the night.

When credit risk management became an established corporate function it started as a kind of middle office providing the line with a second pair of eyes, an ‘independent’ view of each transaction. It then evolved into an assessment function that also rated and ranked credits. In 1988, the first Basel Accord (Basel I) came about linking capital structure to borrowers’ categories and, with the invention of VaR by JP Morgan’s Risk Metrics group, trading market risk was recognized as another key piece of the puzzle, one for which capital could be statistically (scientifically?) computed. This was promptly accepted by regulators in a 1996 amendment to Basel I. A few years more of growth and financial innovations, but also of scandals and high-profile failures, and it became clear that even with market and credit VaR (yes, meanwhile the scope of the magic technique had been extended to credit risk) the picture of what could go wrong was still incomplete. So in the 2004 second Basel Accord (Basel II) operational risk (“the management of risk management” according to Michael Power’s inspired definition) was added to the picture and, together with credit risk, was given a regulatory-endorsed quantitative framework (aka Basel II) intended to, ideally, dovetail with market VaR and provide an overall measure of risk and capital adequacy. The holy grail of enterprise risk management was just, so to speak, a matter of implementation.

As we all know, it did not quite work out. Rivers of ink have been spent to explain the how and why of the 2008 financial crisis and even just rehearsing the most succinct accounts would be beyond the scope of this work. Here I would rather mention two weaknesses of the current financial risk management practices that in my view are not only at the roots of the shortcomings highlighted by the crisis, but, if ignored, would hinder our ability to understand the landscape of banking risks that is now emerging.

In what is arguably one of the most important books of the XVII century, La logique ou l’art de penser, published for the first time in 1662, we can find the following statement: “Fear of harm ought to be proportional not merely to the gravity of the harm, but also to the probability of the event.” The seemingly incontrovertible logic of this thesis lies at the heart of all modern approaches to risk management, from the naïve, albeit hyper prescriptive COSO Framework to the latest sophisticated technologies allowed under the umbrella of the Second Basel Accord or of the Solvency II framework. However, while it is difficult to argue with the general principle, the consequences of taking this thesis at face value are far from trivial. Probably the most long lasting and widespread of them is the identification between risk and volatility, which, although comforting from a statistical point of view, accounts for the nearly exclusive use of VaR for measurement and reporting purposes as well as for the general disregard, when not outright blindness, manifested by financial firms towards highly improbable events (or so believed) and their disastrous consequences. Additionally, this approach carries with it an intrinsic bias in the accuracy of the risk estimates, as we tend to give the same weight to the two quantities (probabilities and severities) while in practice the estimate of probability is much less reliable.

Most importantly, various authors have discussed how risk management can help performance and have concluded that hedging can help reduce taxes, bankruptcy costs, payment to stakeholders, and need for capital as well as be a source of comparative advantage in risk taking. What most authors also agree upon, however, is that the single most relevant contribution to the value of the firm is the prevention of financial distress, or, otherwise put, the reduction in the probability of lower-tail outcomes. But the central role given to probability inevitably focuses risk management actions on the reduction of volatility and therefore on the management of financial performance, so to speak, at the margin, in order to ensure that it remains within certain boundaries. This further distances management’s attention from the prevention of financial distress and from the understanding that risk, rather than volatility, is making mistakes, be they bad investments, improperly managed or incorrectly retained exposures, wrong strategies, or outright operational failures. Not surprisingly then, the vast majority of the resources available have been devoted to activities which, although undoubtedly contributing to the stabilization of performance, do not really focus on preventing distress or bankruptcy.

The other weakness in the still prevailing paradigm underlying current risk management frameworks lies in the rigid and linear classification of risks and related assessment and management. In other words, risk managers tend to produce very long lists of exposures, classified by product, by business unit, and by type of risk, to be then matched by equity capital, hedging instruments, provisions, or other mitigating tools. At no point, however, such taxonomy comes together in an overall picture other than as a sum of its individual parts (a.k.a. the bottom-up approach). The result does not just lack a holistic view of a company’s risk, but, most importantly, the ability to envision, and thus prepare for, events or scenarios

that result from a combination of individual exposures across that static classification.⁶

Risks, furthermore, have a tendency to transform themselves, to disappear from view and reappear in other places and other forms. They cannot be analyzed in accordance with a fixed pattern and then forgotten about. Like the roads and buildings of a city or the dunes of the desert they form a landscape that is forever shifting and that need to be continuously re-mapped if we want to be able to find our way through them. That is why it is dangerous to classify risk into fixed categories, even those that seem established on old and sound foundations, as the result is a rigid paradigm where changes either in the external or in the internal environment may have trouble fitting. On the other hand, changes in the nature of already identified risks as well as the morphing of old risks into new ones happen more often than not precisely at times when our ability to make sense of them is at its lowest, like in the middle of major financial crises. If we consider for instance the risks embedded in the management of a new product, we will find that in many cases it is existing products, well-known and ‘old’ by any other standard, that, because of a sudden change in the market, a new law or regulation, an internal change in systems or personnel, develop unique or unknown risks and become, in this respect ‘new.’ Furthermore, relationships amongst different risks are complex and may become evident only after a peculiar set of circumstances has brought them to the attention of practitioners and analysts.

The history of financial risk management is largely the history of the emergence of certain risks out of changes in the markets, their regulation, and in the economy at large. The risk of exchange rates after the abandonment of the Bretton Woods framework, the risk of interest rates throughout the volatility of the seventies and the eighties (the U.S. prime rate reached its historical high of 21.50% on December 19, 1980), and the risk of oil prices (and of other commodities) after the colossal rises of 1973 and 1978. In some cases, it was the tools and techniques devised to deal with increasing and new risks that generated further risks of their own, like in the case of trading credit (or specific) risk as a consequence of derivatives trading, or model risk as a consequence of mathematical models being used to manage very complex exposures, or like the risks stemming from the success of CDSs as means of hedging credit risk.

Sometimes it takes an unprecedented, and often catastrophic, event to highlight a risk nobody had ever considered as such and sometimes it takes such an event to show how an otherwise well known risk can manifest itself through completely unexpected ways. As an example of the former, take the infamous case of Nick Leesom and Barings Bank. That the combination of lax control standards with the, otherwise understandable, reluctance to probe too much into what appeared to be an extremely successful trading strategy could cause one of the most ancient institutions in the world to go bankrupt had clearly occurred neither to Barings top management nor to British regulatory authorities. All of a sudden the world of finance awoke to the nightmare of ‘rogue trading’ and operational risk became a hot topic of discourse amongst regulators, practitioners, and academics alike.

As an example of the latter, consider the collapse of Long Term Capital management in 1998 when, as a consequence of, amongst other things, the Russian government’s default on its bonds, investors fled from those instruments perceived as less liquid (like off-the-run or ‘older’) T-bills and flocked to buy the more liquid, on-the-run ones (those issued more recently). The strategy of buying one kind of bills and simultaneously shorting the other (a ‘convergence’ trade as it was called) suddenly revealed itself as an open position on market liquidity, not entirely unlike the way Leesom’s strategy of selling straddles on the Nikkei had unwittingly resulted in a huge open position on the volatility of such index.

In more recent years the sudden fall of Lehman Brothers and the subsequent mayhem in global markets had given more than one banker a whole new understanding of the word ‘settlement risk’ and of its implications as they realized the complexities of managing multiple exposures in extreme market conditions. Similarly, albeit in different contexts, the risk of being wiped out from the market because of a failure to give proper consideration to the interest of key stakeholders has also brought home the necessity to revisit not just risk management frameworks, but the overall approach to corporate governance.

A new risk landscape

Three main forces are shaping the new landscape of banking risks: the evolution of the regulatory framework in a direction that is both more demanding and more comprehensive, the emergence of governance and reputation as the central concerns of both bank managers and bank stakeholders, and a realignment of exposures within the traditional domains of credit, market, and operational risks. Let us examine these forces in more detail.

A major overhaul in the structure of financial regulation and supervision was all to be expected, especially as it became clear how the 2008 financial crisis had arisen from a catastrophic underestimation of certain risks, and hence overestimation of capital and liquidity reserves, on one hand, and from a systematic lack of transparency in certain financial transactions, hence the misunderstanding of the risks involved, on the other.

The most important changes are, as expected, in regulatory capital where significantly higher requirements should reduce risk taking while at the same time putting pressure on the bank’s ability to distribute dividends.

and remunerate managers. The countercyclical buffer, an additional capital charge to be set aside during periods of economic expansion, should help keep banks from overextending themselves in good times while providing additional financial cushion for rainy days. More rigorous measures on counterparty credit risk may also help managing undue build ups in over-the-counter exposures while at the same time shifting a substantial burden, in terms of settlement and other operational risks, on clearinghouses, whose role in the global payment system will become even more central and will require very close monitoring. Coupled with stricter liquidity requirements, these provisions will create a more challenging environment for the systematic generation of the ever increasing profits banks had become used to before the crisis. Eventual withdrawal of funding and liquidity support from progressively more indebted governments may add strain to banks’ profit and loss account and, paradoxically, contribute to push banks to ‘search for yield’ through more exotic financial solutions and hence towards the taking of lesser known risks.

One area of the risk space that has not traditionally been part of the risk management practice and that most definitely would, along the aforementioned argument, be classified as ‘consequential’ is what has become known as environmental, social and corporate governance (ESG), a catchphrase denoting the criteria applied in socially responsible investment. The idea that ESG factors (ranging from environmental to social concerns to the quality of corporate governance) are not only issues that special kinds of investors may be considering in their decisions, but constitute a fundamental concern of a company’s management finds its justification in the stakeholder theory of the corporation.7 This theory argues, amongst other things, that management should pay simultaneous attention to the legitimate interests of all appropriate stakeholders and that safeguards should be put in place to balance the related information asymmetry. In a previous work8 I have argued that the risk linked to the distribution of information is embedded in the very nature of the firm because the main advantage in organizing resources within a firm is informational in nature: rather than having to obtain costly information from the market relevant to each transaction, the entrepreneur reduces and internalizes the information needed thereby reducing its marketing costs. In other words, it is in the nature of the firm to subtract information from the market in order to better exploit it internally. One key risk management task is, therefore, to bridge the information gap between the management and other stakeholders, both internal (board of directors, audit committee, employees) and external (regulators, investors, and other groups, through the appropriate reporting channels).

Finally, however, a risk manager does not need any normative justification to take ESG issues seriously. Regardless of whether or not all stakeholders’ interests have intrinsic value, such interests have bearings on a company’s management because their neglect can negatively impact bottom line and shareholders’ value. If stakeholders feel their interest are ignored or threatened by management decisions, they can take actions that may damage the company. These actions can take the form of labor strikes, consumer boycotts, and damaging media coverage, but also much more concretely, as we have seen in the aftermath of the 2007-08 financial crisis, regulatory intervention and ad hoc law making. Traditional risk management tends to view these kind of events as essentially external, events on which the corporation has no control, pretty much like flood or earthquakes. But it is clear, for instance, that the restrictive rules on management bonus enacted after the fall of Lehman and the AIG fiasco or the ‘exemplary’ nature of certain regulatory fines on large banks are the direct result of financial institutions disregarding some key stakeholders’ interests (in various measures those of investors, consumers, and consensus-hungry politicians). Even the now recurrent issue of the ‘too-big-to-fail’ bank can be seen, in an ESG perspective, as the issue of considering the interests of external stakeholders (in this case of the financial system as a whole). Outside of the financial world, the dramatic case of the BP Louisiana oil spill can be interpreted as a ‘too-big-to-fail’ company that has indeed (in ESG terms) ‘failed.’

Taking a look through the more traditional lenses of credit, market, and operational risk, gives us a picture of still substantial credit exposures compounded with the potential for losses driven both by market movements and by operational complexities. The continuing economic recession, coupled with still substantial portions of loan portfolios in commercial properties due for refinancing over the next couple of years, may force banks around the world to take further write-offs on their books while sovereign debt crisis may cause even larger losses and liquidity problems to the many institutions holding large amounts of government bonds.

The collapse of Lehman Brothers has also alerted financial institutions to a set of risks surrounding the actual ‘plumbing’ of the financial markets; that is, the combination of services, from custody to fiscal agency, from clearing to prime brokerage, that front line bankers usually take for granted. The dangerous complexities of post-trade operational risk were painfully evident after KfW Bankengruppe’s unwitting £275m swap payment to Lehman right before it declared bankruptcy. Other insolvency-related scenarios became suddenly very plausible and should be considered in any serious risk management strategy. For instance, the default of a correspondent bank or a fiscal or paying agent may severely hinder the ability of a financial institution to execute payments while payments received on nostro accounts may be captured in the bankruptcy. Potentially heavy financial losses aside, the institution would be exposed to severe reputation risk. Similarly, the default of a prime broker or of a custodian may

---

pose a nightmare scenario in terms of proper segregation and accounting of clients’ assets. And finally, as demonstrated by the Madoff scandal, fraud may not only be extremely difficult to detect, but also result in the outright disappearance of assets in gigantic amounts, with little or no recourse left for investors and other parties.

**A map of banking risks**

If you were appointed to the board of directors of a large, internationally active bank, what would you really want to know about that bank’s risks? One thing you certainly would not want is being overwhelmed with details and having to sift through dozens of pages and hundreds of numbers in order to find any relevant piece of information. Another is to be lectured about obscure quantitative concepts which you may have neither the training for nor the time needed even for a partial understanding thereof. What you probably want, and should actually demand is a document – a summary, a picture, a chart, a number of formats will do – representing all the risks faced by the bank and their interrelations. Like the navigator in a rally race, you are not in the driving seat, you may not even be particularly good at driving, but you need to be able to tell the driver when to slow down, when to turn and at what speed, and when to go faster ahead, knowing that if she makes the wrong turn and crashes into a tree it would be (also) your fault.

What you need, in other words, is a map. A map that tells you what the risks are, where they are, and that can be used to identify those that need your attention and possibly your action. Like any other map it should be drawn on flat media (meaning that a one hundred page document will not do), in the appropriate scale (meaning that it should not provide a level of detail beyond either your understanding or your interest), and that it should highlight the relevant features of the territory (i.e., the key risks in the ensemble of the firm’s activities) allowing you to make decisions as to what actions, if any, are required. Last, but not least, it should also be up to date, lest you find yourself ignoring new avenues that have yet to be charted or, worse still, wandering into roads that are no longer open to traffic. So far the usual solution to this seemingly straightforward problem (straightforward, that is, to formulate) has mainly taken the form of a three column table, delivered in alternative or simultaneously to a three color chart, covering credit, market, and operational risk. In the remainder of this section and in the next I shall try to develop a more articulated solution by reformulating some key principles of cartography in risk-related terms.

**A map should identify all the risks**

As any mariner could tell you, or at least could have told you prior to the ubiquitous availability of GPS devices, a partial map is a very dangerous tool. It may either give you a false sense of security or make you focus on what appear to be a major danger, only to be confronted with much more important ones when it is too late to take action. The issue of completeness in risk identification used to be confined to operational risk, while one could rest assured that credit and market risk had been fully considered to the extent that all portfolios and transactions had been listed and analyzed. But in this period of transformation, realignment, and scrutiny of banking risks, when enormous losses have been suffered because ‘risks had been ignored,’ the task of identification has become as important as, if not more than, assessment itself.

It is true that the map is a representation of the territory and not the territory itself, and that its value depends on what the cartographer chooses to represent and what she chooses to leave outside. But it is also true that what you leave outside may make all the difference in the world. UBS famously left its highly profitable equity derivative desk outside of its risk measurement system (on the basis that it used a different computer system from the rest of the firm) only to see it incurring very high losses a while later. There is no silver bullet that will ensure complete coverage of exposures and appropriate intelligence on their nature, but I will nevertheless mention three key practices that should always be part of an effective risk identification process.

**Process mapping** – is the key preliminary step to risk mapping as it is obviously hard to identify risks in a vacuum without reference to tasks or activities. It is true that it has been mainly used to identify operational risks (especially those related to transaction processing and IT systems), but analyzing all the key processes is a task that should always be performed prior to a risk assessment as it may highlight exposures in many other areas. For instance an analysis of the risk management process may point out deficiencies in the identification of connected exposures, leading thereby to the underestimation of credit risk, or identify problems in the documentations of credit support mechanisms, leading to higher than expected losses in case of default. Similarly an analysis of the trading process may help determine whether practices like repo 105 or 108 are used or if there are hidden liquidity or settlement risks.

**Expert opinions** – the managers and employees who actually run a business are the most important source of intelligence about that business’ risks, first because they know it better than any auditor or risk manager and second because they know what is really happening at the moment, not how the procedure manual says it should be happening. The opinion of the experts is not only a fundamental complement to historical data and statistical indicators, but also a way to recognize that there is always a key difference between the map and the territory and that very few people do actually know every inch of the latter.

---

9 From Wikipedia.org, “Repo 105 is an accounting maneuver where a short-term loan is classified as a sale. The cash obtained through this ‘sale’ is then used to pay down debt, allowing the company to appear to reduce its leverage by temporarily paying down liabilities – just long enough to reflect on the company’s published balance sheet. After the company’s financial reports are published, the company borrows cash and repurchases its original assets.”
Scenario analysis – is another very important tool in risk identification as it allows us to go from an abstract description of generic events to concrete examples of what can go wrong and how, thereby providing useful insights for risk management and decision-making. Furthermore, by stretching the analysis to situations far from the business-as-usual, it may make certain exposures emerge that might otherwise have remained overlooked. For instance, the formulation of a liquidity crisis scenario may highlight the fact that the replacement cost of certain swaps might be much higher than anticipated, with consequent underestimation of specific risk. Or a carefully constructed rogue trading scenario may help identify a concentration or a volatility exposure in the derivative portfolio.

A map should allow us to rank risks in order to identify the most important ones

There are two problems in performing a bank wide assessment of risks. One is the difficulty in finding a common metric and the other is the intrinsic unreliability of virtually all the metrics available. The first issue is well known to financial practitioners and the yearning to resolve it has largely been responsible for the almost universal adoption of VaR-based methodologies. VaR, however, like other statistically-based risk measures, like the ‘expected shortfall,’ is difficult to apply when the potential losses cannot, either analytically or empirically, be easily modeled through a probability distribution over a given time horizon. That is why there is no such thing as liquidity VaR or maturity transformation VaR. Furthermore, if one really wants to map risks like reputation or strategy, pretending to quantify them in terms of ‘losses that can only be exceeded 0.01% of the times’ is likely to sound preposterous.

Risks have also traditionally been assessed on the basis of ad hoc 5-point scales, or ‘mapped’ on two-dimensional charts according to probability and impact, but neither of these approaches can credibly be applied to the totality of banking exposures and is likely to be used only in the very initial stage of an exposure’s assessment, when no analytical models have been developed, historical data are unavailable, and there is no consensus on risk policy or methodology. On top of that, even if we were allowed the luxury of a common metric by which we could rank all the key exposures, we would still be a long way from being able to tell which risk deserves which action, which is the ultimate goal of the risk management process.

A map should identify gaps in management and suggest actions

That is why I suggest to map, and possibly rank, risks in a different way, by using maximum loss severity as one dimension, and management actions on the other. In this context the term ‘management actions’ refers to management ability to take, hedge, or avoid risks. As argued by Stultz, companies’ comparative advantage in risk taking depends on the availability of specialized information and on their ability to withstand large losses.

It follows that for each kind of exposure, management decision to retain, hedge, or eliminate should be based on whether they have superior knowledge of that exposure and if the capital structure is robust enough.

To exemplify, consider the example of a well-capitalized mortgage lender in a small country whose main exposures are residential and commercial property loans as well as foreign exchange rates in its treasury operations. The lender, as a specialist within a small market will know very well both the real estate market and the quality of the borrowers. This superior information can be put to good use by accurately pricing the related risk and optimizing the allocation of resources. On the other hand, the limited size of the foreign exchange operations will not allow any informational advantage through the observation of order flows and market making, hence the need to hedge this kind of risk lest management be caught holding positions they do not have any special ability to foretell. Figure 1 is a variation on the classical probability/severity plot and provides, in my view, a better and more informative mapping of the risks.

Severity of loss, like other quantitative measures, VaR included, is only informative in relation to a bank’s ability to withstand it. That is why I have also highlighted income, capital, and assets (not necessarily in scale), so as to indicate how close a given loss would bring the bank to financial distress or to bankruptcy. The Y axis represents the ability to retain or self-insure the risk, a more qualitative, but equally relevant indication. Figure 1 shows how large exposures can be retained when the bank is comparatively better (that is, has privileged information) at handling them while even relatively smaller ones are better to be hedged when it has no superior ability in their regard.

By contrast, if we now look at a different kind of bank, namely at Lehman Brothers in the last quarter of 2007, less than a year before becoming insolvent, we see a radically different picture. Lehman’s main exposures, with potential losses far exceeding its capital as its leverage was at the time in excess of 30 to 1, were either directly or indirectly in residential and commercial real estate, a business about which it had no privileged knowledge.
with respect to the other institutions. Furthermore these investments were highly illiquid and very difficult to hedge. So not only was most of the risk of the worst kind, with severity potentially exceeding the capital base and in areas where they had no special advantage in risk taking, but hedging or reducing those risks was especially challenging, and even more so during a financial crisis. A more nightmarish risk profile is hard to imagine.

The message from these pictures is that the most critical exposures are those that are at the same time dangerously close to, when not larger than, a company’s ability to bear the potential losses and for which the company itself has little or no advantage in knowledge and understanding. Those are the risks a board of directors should be alerted to; those are the risks top management should take swift decisions about, knowing that there are cases where such decisions may be very painful indeed.

A map should highlight current and potential relationships amongst risks and hence help identify complex scenarios

In 1940, shortly before his premature death, German philosopher and historian Walter Benjamin wrote a short essay15 in which he criticized the traditional view of history as a sequence of events linked by cause-effect relationships that follow each other “like the beads of a rosary” and where “procedure is additive: it musters a mass of data to fill the homogeneous, empty time.” By contrast, he argued that past and present form “constellations” where the tensions of history take place. While reading Benjamin’s work, I cannot help being reminded of the way risk management is routinely practiced in financial institutions and of how sorely we need to identify those “constellations” of risk leading to scenarios that could make the difference between variability in performance and unmitigated disaster. One might say that risk mapping, rather than the charting of the earth, resembles more the mapping of constellations, where relationships have to be imagined and links drawn amongst stars that are both distant and different in size.

If we want to get an idea of how high real losses could be and how they could come about, we must think in terms of events, or scenarios, even if, for coherence’s sake, we still prefer to stick to the usual risk categories. We must envision in what ways a certain kind of loss may be brought about which is in the order of magnitude of the company’s equity, so that we can judge whether possible preventative measures have been taken and if any remedial actions could be foreseen.

The link between VaR figures and loss scenarios is hard to establish as it is, to a certain extent, arbitrary. In principle, of course, anything can happen and a great number of possible outcomes can all credibly result from the same risk profile. However, this needs to be neither an exercise in future telling nor a dull list of merely theoretical possibilities.

Let us take a look at Figure 3. The biggest exposures are credit risks in commercial and retail banking (big surprise!) and, if the assessment has been done in line with Basel II requirements, those two amounts represent roughly the economic capital to be set aside for those two risks. But how can that kind of loss occur, say, in commercial banking? Imagine the following scenario. The bank advises a major client in the energy industry

The sale of unsuitable hedging products, although potentially damaging to the bank’s reputation, would normally result in a settlement or in punitive damages to be paid. Regulatory fines are incurred all the time for non-compliance and the lending departments routinely overlook the potential for damage to third parties or to the environment, as this is not, strictly speaking, the bank’s responsibility. But in this case the combined impact of regulatory intervention, client and third party litigation generates a catastrophic loss of reputation forcing senior management to bow out. The subsequent loss of confidence in the market allows a rival to perform a successful takeover. Figure 4 summarizes the key elements of this scenario.

**Talking to the top**

In this section I would like to suggest an approach to risk assessment and reporting that can convey all the features just discussed: comprehensiveness of risk identification, risk assessment along different metrics, identification of actions, and analysis of complex scenarios. Such an approach does rely on historical data, but also on information about the current status of the firm and the market as well as on hypotheses and analyses on what might happen in the future. It also extends to credit and market risk the key principle of the Advanced Measurement Approach in the second Basel Accord, namely that risk assessment should be based on a combination of historical data, management assessment, and information on the specific business and control environment. Another way to express the same principle, as shown in Figure 5, is that risk assessment should look at the past in producing statistical estimates, at the present in monitoring key risk indicators and other environment-driven information, and at the future, by stressing statistical results and formulating scenarios about how exposures could result in future losses.

Let us now discuss a more articulated application of these ideas in the form of the risk report structure displayed in Figure 6.

The risk types in the first column of Figure 6 are the usual suspects, with a special mention for liquidity and ESG. However, nothing would prevent a bank, to the extent it had the necessary intelligence about it, from including other risks, like reputation or strategy. The second column provides the classical quantitative assessment, based on historical information and, whenever possible, elaborated statistically as appropriate. The third column looks at the current situation as depicted by key risk

![Figure 3 - A traditional risk report](image)

![Figure 4 - A disaster scenario](image)

![Figure 5 - A risk assessment framework](image)
The Capco Institute Journal of Financial Transformation

The Map and the Territory: The Shifting Landscape of Banking Risk

<table>
<thead>
<tr>
<th>Risk types</th>
<th>Backward-looking historical data</th>
<th>Present-looking business environment</th>
<th>Forward-looking scenario analysis</th>
<th>Key catastrophic scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit</td>
<td>VaR, ES, or other</td>
<td>Indicators and trends</td>
<td>Increase in PD and LGDs</td>
<td>Stress-driven</td>
</tr>
<tr>
<td>Market</td>
<td>VaR, ES, or other</td>
<td>Indicators and trends</td>
<td>Yield curve shift</td>
<td>Liquidity crisis</td>
</tr>
<tr>
<td>Specific</td>
<td>VaR, ES, or other</td>
<td>Indicators and trends</td>
<td>Major move in indexes</td>
<td></td>
</tr>
<tr>
<td>Liquidy</td>
<td>Impact on economic value of assets</td>
<td>Values and trends</td>
<td>Major move in Forex</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>VaR, ES, or other</td>
<td>Key risk indicators</td>
<td>Major increase in cost</td>
<td></td>
</tr>
<tr>
<td>ESG</td>
<td>Historical information (if available)</td>
<td>ESG ratings by</td>
<td>Major move in relevant liquidity</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 – Structure of a modern risk report

indicators and their trends while the fourth looks at the future and in particular at a future of low probability and high severity, far from business as usual. This is where losses should be identified that can truly threaten the achievement of the bank’s objective and destroy its profitability. But it is in the fourth column that things get really ugly: in those combinations of adverse scenarios that can generate catastrophic outcomes, with losses that can wipe out the capital of the bank. The objective of this kind of report is neither to mix up incommensurable quantities nor to throw in every possible bit of information in the hope of getting something right. The idea is rather to accomplish three simple things.

- To provide a clear distinction between hard data, managers’ opinions, and hypotheses about the future; between what is certain, what is not certain, but based on experience and what is creative hypothesis-making and, to an extent, guessing. As a great American philosopher once wrote,16 this is in the end the only way through which we can make progress in science. I would modestly add that this is also the only way we can make any non-trivial statement in risk analysis.
- To show how the above pieces of information are not disconnected, but add up to identify the key vulnerabilities that make up a bank’s risk profile.
- To tell the boards of directors and executives what they really should lose sleep over, those unthinkable, but definitely possible combinations of events that could wipe out a company’s capital, goodwill, and reputation.

Finally, the report should specify what, if any, mitigating actions the bank is taking in order to manage those extreme risks and whether they are deemed sufficient and why. This would put the recipients of the report in a position of knowing what decisions may be requested from them as well as of forming their own opinion as to the appropriateness of the actions to be taken.

Conclusion

In this paper I have discussed the shifting landscape of risk management post financial crisis, the changes in the risk profile of banks, including the increasingly relevant role taken by risks not traditionally considered by risk management departments, the importance of building a map of such risks and the practical challenges in providing boards of directors, top management, and other stakeholders with a comprehensive view of all the risks faced. Within this context I have tried to show how we need to replace the traditional, naïve view of separate risk categories with a more modern one in which exposures continually transform into one another and call for a framework of analysis that resembles more the charting of complex paths through partially unknown territories than financial accounting. Finally, I have suggested a reporting structure that puts together the key types of information that are always required to make effective decisions: historical data, up to date indicators, expert opinions, and creative hypotheses.

---

16 Peirce, C. S., 1903, Harvard lectures on pragmatism, collected papers, Harvard University Press, Cambridge. “All the ideas of science come to it by the way of Abduction. Abduction consists in studying facts and devising a theory to explain them. Its only justification is that if we are ever to understand things at all, it must be in that way.”
Towards Implementation of Capital Adequacy (Pillar 2) Guidelines

Kosrow Dehnad — Adjunct Professor, IEOR Department, Columbia University
Mani Shabrang — AIMS Consulting

Abstract
This article describes an approach that can be used to discover market conditions that could have major negative implications on the financial health and viability of financial institutions. The approach can be used in the implementation of capital adequacy (pillar 2) guidelines. The main advantage of the approach is that it is simple and scalable because it uses two or three levels for each market parameter and these levels cover extreme market conditions that should be cause for concern even though such conditions might be economically improbable tail events. A simpler version of the methodology uses Orthogonal Arrays (OA) and selects a smaller yet ‘balanced’ and ‘representative’ subset of the above scenarios as test cases. These approaches, which are applied widely in manufacturing, can be used by regulators and management to determine capital adequacy and corrective measures that should be taken to make financial institutions more robust in cases of market stress.
The recent financial crisis has revealed certain weaknesses of the existing risk management systems. Most of these systems are based on continuous time finance that attempt to describe the markets and their evolutions under ‘normal’ conditions. The explicit and implicit assumptions of continuous time finance imply a negative feedback that tends to stabilize the markets and prevent them from experiencing sharp price movements. It assumes that if prices fall, an army of arbitrageurs and market savvy investors are ready to jump in and bid up the price. These assumptions, however, ignore the fact that markets are composed of people with their own fears and greed who during times of market stress tend to look to exit the market at the same time, thus creating a positive feedback which in turn tends to destabilize the markets and push prices to extremes.

For example, a market move might force a large hedge-fund to liquidate some of its positions that will in turn trigger liquidation by other hedge funds with similar strategies. This will also be the case if a large number of investors have similar positions and are forced to liquidate them because of, say, margin requirements. During such times, the magnitude of the moves can be many times their historical standard deviations and they can take place in a short period of time. The combination of the two would imply an extremely high volatility.

One reason that some of the highly elegant mathematical models do not work in these cases is because the time and volatility are assumed to be homogenous. For example, when volatility of JPY is said to be 13%, this value does not depend on the time of the day. Though this number can be a reasonable average over an extended period, when markets are under stress, the price action can be quite fast and violent and much more than that implied by the 13% annual volatility. Clearly, one could expand the models and include this feature; however, extensions of this nature would require estimation of additional parameters.

When markets are in equilibrium, the linear approximation of the value of a position as a function of its constituents – the linear component of Taylor expansion – is a reasonable representation of the evolution of the value of that position. This approximation is closely related to delta hedging.

\[ f(a) + \frac{\partial f(a)}{\partial x} (x-a) + \frac{\partial^2 f(a)}{\partial x^2} \frac{(x-a)^2}{2!} + \frac{\partial^3 f(a)}{\partial x^3} \frac{(x-a)^3}{3!} + \ldots \]

or,

\[ \sum_{n=0}^{\infty} \frac{\partial^n f(a)}{\partial x^n} \frac{(x-a)^n}{n!} \]

Under normal market condition, when market moves are small and approximately continuous one could argue that \( F(X+\Delta X) \approx F(X) + \Delta \partial F(X) \).

This is the basic idea of delta hedging that a continuous smooth path can be approximated with a series of straight line segments. However, when markets have sharp moves and so-called gaps, the second order effects, such as convexity, gamma, and cross gamma, become important and the risks associated with them manifest themselves. This could, for example, happen when one big player or a number of smaller ones with similar positions are forced to liquidate their positions for, say, collateral purposes.

\[ F(X+\Delta X) = F(X) + \Delta x \partial F(X) + \Delta x \partial^2 F(X) \Delta x + \ldots \]

\[ \partial F(X) = \begin{bmatrix} \frac{\partial F_1(x_1)}{\partial x_1} & \ldots & \frac{\partial F_n(x_n)}{\partial x_n} \end{bmatrix} \]

This indicates that if we expand up to the second term then pair-wise correlation and second order, convexity in the case of \( \partial^2 F(X) \Delta x \) – when both derivatives are with respect to the same variable – becomes important and significantly impacts the valuation of the position. The full scale approach discussed in this paper aims to determine the impact of correlation and convexity when markets are under stress.

Let us recall that models and approaches such as VAR often do not incorporate these situations and most risk management systems are simulation-based that could take quite some time to run. The question of interest before regulators and risks managers is the conditions under which banks would lack adequate capital and might be under distress. In other words, when the difference between their assets and liabilities falls below a critical level. Should this happen, there is a chance that lenders will be reluctant to provide short term liquidity to the bank thus increasing the institution’s cost of fund. This will, in turn, exacerbate the situation and will be destabilizing – the so-called liquidity risk. To determine the above market scenarios for banks with simple balance sheets, it might be sufficient to consider their largest classes of assets and liabilities.

For example, if most of the assets of the bank are in real estate loans that are funded through short term borrowings, then one could argue that the bank would most probably be under stress if the credit spread of its borrowers widens and loan losses and cost of funds of the bank increase. Similarly, if the bank has long term assets that are funded through short term borrowing, any increase in cost of funds will negatively impact the bank’s capital unless the curve inverts and the asset values increase sufficiently.

However, in the case of banks with more complicated balance sheets the question of detecting scenarios that will negatively impact the bank could be more complex and one cannot simply rely on the size of an exposure as the main driver of the test. For example, consider an insurance company that funds its long term liabilities with long term assets. In this case the shape of the yield curve might not be that important to the solvency of the company. Specifically, suppose a financial institution has...
invested in a large life settlement portfolio. In this case the combination of increases in life expectancy and high interest rates will be more important than the shape of the yield curve. In short, when the balance sheet is more complicated, the approach based on full factorial or OA provide a more efficient way for exposing areas of potential risk and concern; particularly when considering a portfolio of banks and their potential impact on FDIC resources. After using these techniques to determine market scenarios that could result in financial distress, one can use historical data and macroeconomic models to assign a probability to each of those scenarios.

Using extreme values of market factors also simplifies valuation of structures with embedded options because in most cases such options will be either deep in-the-money or deep out-of-the-money and essentially all calculations become those based on forwards. The combinations of these features plus use of the concept of nearest neighbor make it easy to detect market scenarios that will cause a financial institution to run into capital adequacy difficulties, thus requiring remedial action by regulators.

Discussion
The following simple numerical example illustrates the workings of the approach. Consider a very simple balance sheet where the assets minus liabilities are a function of the yield curve and default. For default, we assume two levels of very high (1) and very low (2) and for the yield curve we assume flat and high (1), flat and low (2), steep (3), and inverted (4). Since each factor has only two levels we use 2x2, which is four to represent the shape of yield curve. The following is the full factorial approach to the problem.

<table>
<thead>
<tr>
<th>Credit</th>
<th>YC</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>L</td>
<td>I</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
</tr>
<tr>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>H</td>
</tr>
</tbody>
</table>

We have two levels for default, low and high (L,H), and four levels for the shape of the curve, flat low (L), flat high (H), steep (S), and inverted (I). We use the following full factorial to assign shape of curve:

<table>
<thead>
<tr>
<th>Credit</th>
<th>YC</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>L</td>
<td>I</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
</tr>
<tr>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>H</td>
</tr>
</tbody>
</table>

In the above table each level of factors and each pair of variables appear the same number of times – the main properties of OA, a balanced and to some extent comprehensive set of market conditions that does not favor any particular scenario. Since we want to ensure that the case of high default rate (H) and inverted yield curve (I) are included in these scenarios, we adjust the assignments of yield curve scenarios and credit as follows:

<table>
<thead>
<tr>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
</tr>
<tr>
<td>H</td>
</tr>
</tbody>
</table>

Based on the above we realize that the most risky scenario corresponds to the case represented by 7th row, where defaults are high and the yield curve is inverted. Using historical data we conclude that the chance of this scenario happening is not remote, hence the smaller and more sparse combinations i.e., OA, must include this market condition.

The OA combination is as follows:

<table>
<thead>
<tr>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
</tr>
<tr>
<td>H</td>
</tr>
</tbody>
</table>

In the above table each level of factors and each pair of variables appear the same number of times – the main properties of OA, a balanced and to some extent comprehensive set of market conditions that does not favor any particular scenario. Since we want to ensure that the case of high default rate (H) and inverted yield curve (I) are included in these scenarios, we adjust the assignments of yield curve scenarios and credit as follows:

<table>
<thead>
<tr>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
</tr>
<tr>
<td>H</td>
</tr>
</tbody>
</table>
And the corresponding OA will be:

<table>
<thead>
<tr>
<th>Row</th>
<th>Credit</th>
<th>YC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>3</td>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>4</td>
<td>L</td>
<td>L</td>
</tr>
</tbody>
</table>

If the financial institution shows lack of sufficient capital under the above extreme and tail events, the regulators could ask the bank management to take remedial actions. By using this analysis for all the banks, the regulators could determine market conditions that could put the whole banking sector on national or regional levels under stress.

**Conclusion**

A simple procedure is proposed whereby significant financial parameters that impact the solvency of a bank are factored from a Pillar II perspective. It is shown that an exhaustive list of financial parameters can be reduced to a manageable set represented by an orthogonal array without losing the main solvency drivers, thereby gaining efficiency and simplicity.

**References:**

- Belmont, D. P., 2004, Value added risk management, Wiley Finance Series
The Failure of Financial Econometrics: Estimation of the Hedge Ratio as an Illustration

Imad Moosa — Professor of Finance, School of Economics, Finance and Marketing, Royal Melbourne Institute of Technology

Abstract

This paper demonstrates how the econometric modeling of the hedge ratio has no value added whatsoever to the improvement of hedging effectiveness and that using the so-called naïve model (a hedge ratio of one) produces similar results to those obtained from elaborate model specifications and ‘sophisticated’ estimation methods. The exercise involves the estimation of the hedge ratios for a position on the Singapore dollar when the base currency is the New Zealand dollar. The results, based on monthly data covering the period 1998:5-2009:9, show that the effectiveness of money market and cross currency hedging does not depend on model specification or the estimation method.
The recent global financial crisis has taught some of us a lesson that all of us should have learned a long time ago. The lesson is that we must not put too much faith in econometric models, particularly in the field of finance where a significant amount of other people’s money (and livelihood) may be exposed to risk. This is to say that we should not trust the tools of financial econometrics. Unfortunately, it seems that it is still business as usual for some true believers in the power and usefulness of financial econometrics.

Financial econometricians tell us that financial prices are integrated variables – they follow a random walk and consequently, by definition, are not forecastable – yet they endeavor to come up with models that allegedly can be used to forecast financial prices. Even worse, some of the models used in financial risk management are dangerous because they instill complacency. When a bank executive believes in the prediction of a model that the bank has adequate capital to protect it from insolvency with a confidence level of 99.9 percent, the tendency would be to do nothing about managing risk properly. The disaster that hit AIG and cost U.S. taxpayers U.S.$170 billion was a product of greed and reliance on a copula-based model that precluded the possibility of mass, nationwide default on mortgages and that of house prices falling in all parts of the U.S. simultaneously. The development (by the so-called “JP Morgan Mafia”) of the toxic assets that blew up the world financial system and destroyed Iceland and Greece was based on models that predicted that what actually happened could only happen once every 15 billion years. It is rather unfortunate that some of us believe that models can forecast the unforecastable, Black Swans – that is, low-frequency, high severity loss events.

Most of the work done by econometricians (financial and otherwise) bears no relevance to reality. Some elaborate models have been developed to represent and forecast financial variables, including ARMA, ARIMA, ARFIMA, TAR and SETAR models. We also have neural networks, wavelet analysis, and multi-chain Markov switching models. The Nobel Prize was awarded to Robert Engle for inventing ARCH models, which can supposedly explain and predict financial volatility, but things did not stop there. The objective of this paper is limited to demonstrating that model sophistication (with respect to specification or estimation method) is irrelevant to achieving a practical objective, which in this exercise is effective hedging of foreign exchange risk. There seem to be mixed views on this issue: some believe that econometric sophistication does matter for hedging effectiveness, while others (a minority) do not share this view. This paper provides evidence that pertains directly to the issue and supports the minority view.

The effectiveness of financial hedging

Financial hedging of exposure to foreign exchange risk, resulting from unanticipated changes in the exchange rate, entails taking an opposite position on a financial asset whose price is correlated with the price of the unhedged position. Money market hedging involves borrowing and lending with the objective of creating a synthetic forward contract whose price is the interest parity forward rate. In cross currency hedging, a position is taken on another currency whose exchange rate against the base currency is correlated with the exchange rate between the base currency and the exposure currency. Let x, y, and z be the base currency, currency of exposure, and third currency, respectively. Let \( i_x \) and \( i_y \) be the interest rates on currencies x and y, respectively. The price of the unhedged position is, therefore, the exchange rate between x and y, or \( S(x/y) \). Foreign exchange risk arises from fluctuations in this rate. The price of the hedging asset is the interest parity forward rate, which is consistent with covered interest parity. It is calculated as \( F(x/y) = S(x/y)[1+ i_x]/[1+ i_y] \).

In the case of cross currency hedging, the price of the hedging asset is \( S(x/z) \). What is important for the effectiveness of money market hedging is correlation between \( S(x/y) \) and \( F(x/y) \) and for forward hedging it is the correlation between \( S(x/y) \) and \( S(x/z) \) (more precisely, correlation between the percentage changes in these rates, known commonly as rates of return).

The effectiveness of hedging exposure to foreign exchange risk can be measured by the variance of the rate of return on the unhedged position relative to the variance of the rate of return on the hedged position. Thus we test the equality of the variances of the rates of return on the unhedged and hedged positions. The null hypothesis is \( H_0: \sigma^2(R_u) = \sigma^2(R_h) \) against the alternative \( H_1: \sigma^2(R_u) > \sigma^2(R_h) \), where \( R_H = \Delta p_H \)}
The Capco Institute Journal of Financial Transformation
The Failure of Financial Econometrics: Estimation of the Hedge Ratio as an Illustration

\[= R_U - hR_A \quad \text{and} \quad R_U = \Delta p_{U,0} \quad \text{is the first log difference of the price of the hedged position,} \]
\[\Delta p_{U,1} \quad \text{is the first log difference of the unhedged position, and} \]
\[R_A = \Delta p_{A,0} \quad \text{is the first log difference of the price of the hedging asset. The null hypothesis is rejected if} \]
\[VR = \frac{\sigma^2(R_{U,0})}{\sigma^2(R_{U,1})} > F(n-1, n-1) \quad (4), \]
\[\text{where} \ VR \ \text{is the variance ratio and} \ n \ \text{is the sample size. This test can be complemented by calculating variance reduction:} \]
\[VD = 1 - 1/VR = 1 - \frac{\sigma^2(R_{U,0})}{\sigma^2(R_{U,1})} \quad (6). \]

The variance ratio test can be conducted to compare the effectiveness of two hedging positions resulting from the use of different hedge ratios or different hedging instruments. In this case, the null hypothesis becomes \(H_0: \sigma^2(R_{U,1}) = \sigma^2(R_{U,2})\), where \(\sigma^2(R_{U,1})\) and \(\sigma^2(R_{U,2})\) are the rates of return on the hedged positions resulting from hedge number one and hedge number two, respectively.

**The econometrics of the hedge ratio**

When it comes to the estimation of the hedge ratio, the easiest thing to do is use a hedge ratio of one, which implies the covering of the whole exposed position. In the literature, this is known as the naïve model. There is also the implied model, which allows the estimation of the conditional covariance by employing the implied volatilities derived from currency options. And there is the random walk model, which assumes that the most appropriate forecast of future variance and covariance is the variance and covariance observed today.

Financial econometricians have been on a quest to develop increasingly sophisticated models to estimate the hedge ratio, warning continuously that failure to use an appropriate model (and we do not know what that is) will result in under-hedging or over-hedging. The starting point is the conventional first difference model, also called the simple model and the historical model. This model amounts to estimating the hedge ratio from historical data by employing a linear OLS regression of the form \(\Delta p_{U,t} = \alpha + h \Delta p_{A,t} + \beta_1 + \varepsilon_t \quad (7)\), in which \(h\) is the hedge ratio and the \(R^2\) of the regression measures hedging effectiveness. Sometimes, the regression is written in levels rather than in first differences to give \(p_{U,t} = \alpha + hp_{A,t} + \epsilon_t \quad (7)\).

It is argued that one problem with the conventional model is that equation \(7\) ignores short-run dynamics, whereas equation \(6\) ignores the long-run relation as represented by \(7\). Specifically, if \(p_U\) and \(p_A\) are cointegrated such that \(\gamma = 0\), then equation \(6\) is misspecified, and the correctly specified model is an error correction model of the form \(\Delta p_{U,t} = \alpha + \sum_{m=1}^{\gamma} \beta_m \Delta p_{U,t-m} + h \Delta p_{A,t} + \sum_{m=0}^{\gamma} \gamma_m \Delta p_{A,t-m} + \beta_1 + \varepsilon_t \quad (8)\), where \(\beta_1\) is the coefficient on the error correction term, which should be significantly negative for the model to be valid. This coefficient measures the speed of adjustment to the long-run value of \(p_{U,t}\), as implied by equation \(7\). In other words, it is a measure of the speed at which deviations from the long-run value are eliminated.

Lien (1996) argues that the estimation of the hedge ratio and hedging effectiveness may change significantly when the possibility of cointegration between prices is ignored. In Lien and Luo (1994) it is shown that although GARCH may characterize the price behavior, the cointegrating relation is the only truly indispensable component when comparing ex-post performance of various hedging strategies. Ghosh (1993) concluded that a smaller than optimal futures position is undertaken when the cointegrating relation is unduly ignored, attributing the under-hedge results to model misspecification. Lien (1996) provides a theoretical analysis of this proposition, concluding that an errant hedger who mistakenly omits the cointegrating relation always undertakes a smaller than optimal position on the hedging instrument. While Lien’s proof is rather elegant, the empirical results derived from an error correction model are typically not that different from those derived from a simple first difference model (for example, Moosa (2003)).

But there is more in the econometricians’ bag of tricks. Kroner and Sultan (1993) used a bivariate GARCH error correction model to account for both nonstationarity and time-varying moments. Broll et al. (2001) suggested that the hedge ratio should be estimated from a nonlinear model, which can be written in first differences as \(\Delta p_{U,t} = \alpha + h \Delta p_{A,t} + \gamma \Delta p^2_{A,t} + \epsilon_t \quad (9)\).

Nonlinear error correction models have also been suggested (not necessarily for estimating the hedge ratio) by Escribano (1987), and the procedure is applied to a model of the demand for money in Hendry and Ericson (1991). Nonlinearity in this case is captured by a polynomial in the error correction term. Thus the nonlinear error correction model corresponding to equation \(8\) is \(\Delta p_{U,t} = A(L) \Delta p_{A,t} + B(L) \Delta p_{A,t} + \sum_{l=1}^{m} \gamma_m \Delta p_{A,t-l} + \varepsilon_t \quad (10)\), where \(A(L)\) and \(B(L)\) are lag polynomials. Hendry and Ericson (1991) suggest that a polynomial of degree three in the error correction term is sufficient to capture the adjustment process.

Yet another procedure to estimate the hedge ratio is to use an autoregressive distributed lag (ARDL) model of the form \(\Delta p_{U,t} = \sum_{m=1}^{\gamma} \alpha_m p_{U,t-m} + \sum_{m=0}^{\gamma} \beta_m p_{A,t-m} + \varepsilon_t \quad (11)\). In which case the hedge ratio may be defined as the coefficient on \(\Delta p_{A,t}\) (\(h = \beta_1\)) or as the long-term coefficient, which is calculated as \(h = [\sum_{l=0}^{\gamma} \beta_l] - \frac{\gamma \beta_1 \gamma \Delta p_{A,t-m} + \varepsilon_t} \quad (12)\). The empirical results presented in this paper are based on equations \(6\), \(8\), \(9\), \(10\), \(11\), and \(12\).

**Data and empirical results**

While any currency combination can be used to conduct this empirical exercise, we picked (at random) the following currency combination: the base currency is the New Zealand dollar, the currency of exposure is the Singapore dollar, and the third currency used for cross currency hedging is the Hong Kong dollar. Monthly data on exchange and interest rates are used, covering the period 1998:5-2009:9. The data were obtained from Bloomberg.
In this exercise we estimate the hedge ratio from nine combinations of model specifications and estimation methods. These procedures are listed in Table 1. The objective is to find out whether the estimation method or model specification makes any difference for hedging effectiveness. This will be applied to money market hedging and cross currency hedging. If $s_1 = \log(S(x/y))$, then $\Delta_pU = \Delta s_1$. If $f = \log(F(x/y))$ and $s_2 = \log(S(x/z))$, then $\Delta_pA = \Delta f$ for money market hedging and $\Delta_pA = \Delta s_2$ for cross currency hedging. A large number of studies used GARCH models to estimate the hedge ratio (for example, Scarpa and Manera (2006)), but in this paper we try other models and methods that have not been used extensively. This is not to say that these models and methods have not been used before. For example, Coffey et al. (2000) used the Cochrane-Orcutt method, which is also known as GLS, whereas Scholes and Williams (1977) used instrumental variables. Other models and methods used in the literature include BEKK, EWMA, VAR, VECM, and EGARCH.

The estimation results are presented in Table 2, which reports the estimated value of the hedge ratio, its t statistic, and the coefficient of determination. Also reported in Table 2 are the variance ratio and variance reduction. Consider money market hedging first. No matter which procedure is used, the hedge is highly effective – in reality a perfect hedge as the variance of the rate of return on the unhedged position is reduced to almost zero (by over 99 percent). The variance ratio is statistically significant in all cases. What is also interesting is that the so-called naïve model (choosing a hedge ratio of one) produces similar results, a variance reduction of 99 percent.

Consider now cross currency hedging. In all cases, the hedge is effective, reducing the variance of the rate of return on the unhedged position by 80 percent, irrespective of the procedure used to estimate the hedge ratio. In this case, the naïve model also produces an effective hedge that reduces the variance by 75 percent. What explains the difference between the results obtained under money market hedging and cross currency hedging is correlation. The spot and interest parity forward rates are almost perfectly correlated, so they produce an almost perfect hedge, irrespective of the underlying econometrics. However, the two spot rates involved in cross currency hedging are not as highly correlated. This is why cross currency hedging is less effective than money market hedging. This is also why a hedge ratio of one produces a slightly less effective cross currency hedge. However, by choosing a hedge ratio that is equal to the correlation coefficient between the rates of change of the two spot rates, the hedge becomes as effective as any of those based on the more elaborate models and estimation techniques. It is noteworthy that the use of GARCH models would not change this conclusion. For example, Casillo (2004) found that a multivariate GARCH model is ‘marginally better’ than other models (‘marginally’ does not imply statistical significance). These results corroborate the findings of Moosa (2003) and Maharaj et al. (2009).

<table>
<thead>
<tr>
<th>Specification</th>
<th>Estimation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First difference OLS</td>
</tr>
<tr>
<td>2</td>
<td>First difference The Cochrane-Orcutt method with an AR(2) process in the residuals</td>
</tr>
<tr>
<td>3</td>
<td>First difference Maximum likelihood with an MA(2) process in the residuals</td>
</tr>
<tr>
<td>4</td>
<td>First difference Instrumental variables with an AR(2) process in the residuals</td>
</tr>
<tr>
<td>5</td>
<td>Quadratic first difference OLS</td>
</tr>
<tr>
<td>6</td>
<td>Linear error correction OLS</td>
</tr>
<tr>
<td>7</td>
<td>Nonlinear error correction OLS</td>
</tr>
<tr>
<td>8</td>
<td>Autoregressive distributed lag model in first differences OLS (the hedge ratio is the coefficient on the contemporaneous explanatory variable)</td>
</tr>
<tr>
<td>9</td>
<td>Autoregressive distributed lag model in first differences OLS (the hedge ratio is the long-run coefficient calculated from the impact coefficients)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Hedge ratio</th>
<th>t-statistic</th>
<th>$R^2$</th>
<th>VR</th>
<th>VD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Money market</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0006</td>
<td>669.33</td>
<td>0.99</td>
<td>2461.8</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>1.0006</td>
<td>669.07</td>
<td>0.99</td>
<td>2461.8</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>1.0005</td>
<td>623.43</td>
<td>0.99</td>
<td>2463.8</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.9950</td>
<td>111.26</td>
<td>0.99</td>
<td>2377.2</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>1.0003</td>
<td>654.15</td>
<td>0.99</td>
<td>2466.8</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>1.0009</td>
<td>648.59</td>
<td>0.99</td>
<td>2455.8</td>
<td>0.99</td>
</tr>
<tr>
<td>7</td>
<td>1.0006</td>
<td>636.08</td>
<td>0.99</td>
<td>2461.8</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>0.9997</td>
<td>661.18</td>
<td>0.99</td>
<td>2473.5</td>
<td>0.99</td>
</tr>
<tr>
<td>9</td>
<td>0.9699</td>
<td>13.46</td>
<td>0.99</td>
<td>795.5</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Cross currency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.754</td>
<td>26.22</td>
<td>0.84</td>
<td>5.07</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>0.840</td>
<td>25.88</td>
<td>0.84</td>
<td>5.08</td>
<td>0.80</td>
</tr>
<tr>
<td>3</td>
<td>0.756</td>
<td>26.14</td>
<td>0.84</td>
<td>5.08</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>0.760</td>
<td>7.36</td>
<td>0.73</td>
<td>5.09</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>0.758</td>
<td>25.96</td>
<td>0.84</td>
<td>5.09</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>0.757</td>
<td>25.62</td>
<td>0.84</td>
<td>5.08</td>
<td>0.80</td>
</tr>
<tr>
<td>7</td>
<td>0.759</td>
<td>24.49</td>
<td>0.84</td>
<td>5.08</td>
<td>0.80</td>
</tr>
<tr>
<td>8</td>
<td>0.756</td>
<td>24.78</td>
<td>0.84</td>
<td>5.12</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>0.778</td>
<td>16.27</td>
<td>0.84</td>
<td>5.12</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 1 – Model specifications and estimation methods

Table 2 – Estimation results
Conclusion
The econometric models used in finance for forecasting and measurement (what constitutes financial econometrics) are useless at best and dangerous at worst. This paper illustrates how the econometric models of the hedge ratio are useless. Unlike the models used in risk management for calculating economic capital, models of the hedge ratio are not dangerous. However, they are still not worthy of the tremendous brain power required to develop them and the computer power required to run them.

References

Systemic Risk Seen from the Perspective of Physics

Udo Milkau — Head of Strategy, Market Development and Controlling, Transaction Banking, DZ BANK AG, and part-time lecturer, Goethe University Frankfurt, House of Finance

Abstract
One of the lessons learned from the recent financial crisis is the need to understand the systemic risk in the financial system – i.e., the risk (or probability) for the financial system ‘as a whole’ to turn unstable. As there are analogous concepts in the physics of complex systems, three key aspects of the development of complex systems will be evaluated here: (i) non-linearity, unpredictability, and deterministic chaos; (ii) non-linearity, non-equilibrium, and patterns in space and time; and (iii) phase transitions as a result of ‘magnetic’ ordering. Although financial systems depend to a certain degree on human behavior, the perspective of physics can help to build up a better understanding of how systems ‘as a whole’ develop as a result of interactions between the participants.
One should always be cautious with analogies between economical and physical principles. As Paul A. Samuelson said 40 years ago [Samuelson (1970)]: “There is really nothing more pathetic than to have an economist or a retired engineer try to force analogies between the concepts of physics and the concepts of economics.”

Nevertheless, Paul A. Samuelson talked about ‘structural relations’ between physics and economics in the same lecture. Since the 1950s, the development of theories about complex systems has been one of the big success stories of physics. Thus, it should be worthwhile to examine some structural relations between financial systems with many participants and the physics of complex systems, which could be helpful for the discussion about systemic risk. Because physics is not the only outside-in perspective on systemic risk, it should be remarked that ‘cross-industry’ perspectives – from fire fighting to aviation – on system-wide risk have also been analyzed and the readers are referred to those publications [see e.g., WEF (2010)].

The system as a whole

In the aftermath of the financial crisis, which started with real estate bubbles and is still virulent as a result of the sovereign debt/government budget crisis, questions were raised how our financial systems could have been on the edge of instability and how certain or uncertain the future of our whole financial system is. A 2001 report of the Group of Ten provided a definition, which was sometimes criticized for its tautology (“Systemic financial risk is the risk”), contains an intriguing description about what systemic risk is: “increases in uncertainty about a substantial portion of the financial system.”

Since 2007, there have been a huge number of papers with different definitions of ‘systemic risk,’ and it is beyond the scope of this paper to review all the different approaches. For an overview of recent developments, see, for example, ECB (2009, 2010) and ECB-CFS (2010).

In a lecture in December 2009, President of the European Central Bank, Jean-Claude Trichet, said: “Systemic risk within the financial system relates to the risk that these inter-connections and similarities render emerging financial instability widespread in the system. Even if the original problem seems more contained, important amplification mechanisms can be at work... Systemic risk is about seeing the wood, and not only the trees.”

Before discussing the perspectives of physics about the development of systems as a whole more comprehensively, it is helpful to take a brief look at the developments of economic theory concerning ‘systemic’ perspectives over the past century. While in the first half of the twentieth century the ‘interaction’ between the participants of an economic system was prevalent (or at least part of the discussion), in the second half independent behavior of market participants was the leading paradigm.

First half of the twentieth century: Poincaré and Barchelier

In the first half of the twentieth century, many scientists viewed the ‘economy’ as a system of interacting participants, although the formalism was neither mathematical nor sophisticated. An example is Henri Poincaré, who wrote in his book, “Science et méthode” in 1908: “When men are close to each other, they no longer decide randomly and independently of each other, they each react to the others. Multiple causes come into play which trouble them and pull them from side to side, but there is one thing that these influences cannot destroy and that is their tendency to behave like Panurge’s sheep. And it is that which is preserved.”

Few years before Poincaré published this statement, he was one of the three professors who had to write the report on Louis Bachelier’s thesis about “Théorie de la spéculation” in 1900. Louis Bachelier was, according to common accords, the first to develop a model of stock prices based on the concept of probability. The impact of his pioneering work was discussed in contradictory ways, especially before and after the financial crisis; see, for example, papers by Davis (2006) and Courtnail et al. (2000) versus Kirman (2009). The starting point of Bachelier’s thesis was his first assumption of a general principle that (quote) “L’espérance mathématique du spéculateur est nulle.” In modern terminology, he assumed a market was made up of participants who acted rationally, randomly and independently. His second assumption was that price deviations would not be very large, that the probability of a deviation from the quoted price would not depend on the absolute value of this price, and therefore, the random variables should follow a Gaussian probability distribution. With his third assumption, that price movements do not depend on the past (i.e., lack of memory or Markov property), Bachelier derived a description of a price process which is called a random walk or Brownian motion today.

There is an interesting parallel in physics with the (physical) process of the Brownian motion identified in 1827 [Brown (1828)]. The (physical) Brownian motion was explained by Einstein (1905) and Smoluchowski (1906) based on the assumption of Gaussian distribution of particle velocities. For about a century, nobody challenged this assumption, and Brownian motion and Gaussian distribution were used as synonyms. By chance, but exactly during the financial crisis, Wang et al. (2009) published a

---

1 The original quote in French is: “Quand des hommes sont rapprochés, ils ne se décident plus au hasard et indépendamment les uns des autres ; ils râglissent les uns sur les autres. Des causes multiples entrent en action, elles troublent les hommes, les entraînent à droite et à gauche, mais il y a une chose qu’elles ne peuvent détruire, ce sont leurs habitudes de moutons de Panurge. Et c’est cela qui se conserve.”

2 These different perspectives on Louis Bachelier’s work show that the theoretical perception of economics is quite dependent on the existing economic situation.

3 The idea of a random walk was not new at the time Bachelier wrote his thesis, since 20 years earlier the Danish astronomer and mathematician T.N. Thiele (1880) had derived the idea of Brownian motion with independent and normally distributed increments.
paper on "Anomalous yet Brownian" motion. They delivered the experimental proof that there are certain types of (physical) Brownian motion in which the distributions of displacement are not Gaussian but exponential in some complex liquid systems. The basic assumption was not totally wrong – but limited.

The viewpoints of Barchelier attracted no attention in economics (but some in mathematics) at his time, and he was nearly forgotten for about half a century. However, the concept of ‘interaction’ between individual market participants continued to exist, and Friedrich A. Hayek wrote in 1945: “The whole acts as one market, not because any of its members survey the whole field, but because their limited individual fields of vision sufficiently overlap ...”

This understanding of markets or financial systems was ‘non-mathematical’ but rather invariantly referring to interacting participants. However, in the second half of the twentieth century, the perspective changed, thus resembling Barchelier’s point of view, and focused on independent participants.

**Second half of the twentieth century: Markowitz and Mandelbrot**

In 1952, Harry Markowitz developed his theory of portfolio allocation under uncertainty using the assumption that the changes in returns on assets show a Gaussian distribution. A later concept was presented by Benoit Mandelbrot in 1963 with three hypotheses including random walk and efficient markets, but replacing Gaussian distribution with more general ones. This allows ‘fat tails’ of price change distributions and is a better fit for seldom but extremely large changes of prices. While the type of the distribution selected has a big effect on time series of individual prices, from a ‘systemic’ perspective, it makes no fundamental difference which distribution will be selected, because both concepts start with the assumption of fair game models for financial markets, i.e., with efficient markets [Fama (1970)].

It took until 1987 when a small group of scientists and economists met at the Santa Fe Institute to discuss ‘the economy as an evolving, complex system.’ The result was comprehensively summarized by W. Brian Arthur as a conclusion in a paper about “Complexity in economic and financial markets” [Brian (1995)]: “An economy of course, does indeed consist of technologies, actions, markets, financial institutions and factories – all real and tangible. But behind these, guiding them and being guided by them on a sub-particle level are beliefs: the subjective expectations, multiple hypotheses, and half-hoped anticipations held by real human beings. ... When beliefs form an ocean of interacting, competing, arising and decaying entities, occasionally they simplify into a simple, homogeneous equilibrium set. More often they produce complex, ever-changing patterns. Within the most significant parts of the economy, interacting, non-equilibrium beliefs are unavoidable, and with these so is a world of complexity.”

Although there was a continuous discussion about ‘fat tails’ in the 1990s, the mainstream paradigm of independent participants in efficient markets was not challenged until the financial crisis dramatically changed everything. In the second Banque de France/Bundesbank conference on “The macroeconomy and financial systems in normal times and in times of stress,” Jean-Pierre Landau said in his introductory remarks about “Complexity and the financial crisis” [Landau (2009)]: “... Finally and most importantly, complexity resulted in an increase in overall uncertainty. – Complex systems exhibit well-known features: non-linearity and discontinuities (a good example being liquidity freezes); path dependency; sensitivity to initial conditions. Together, those characteristics make the system truly unpredictable and uncertain, in the Knightian sense. Hence the spectacular failure of models during the crisis: most, if not all, were constructed on the assumption that stable and predictable (usually normal) distribution probabilities could be used ...”

The de Larosière report (2009) argued pointedly about the causes of the financial crisis: “too much attention was paid to each individual firm and too little to the impact of general developments on sectors or markets as a whole.”

After one hundred years, the discussion came back to the point of Poincaré, and the development of a system as a whole was put in the focus again.

**Systemic importance versus systemic risk**

To analyze the financial system as a whole, physics can provide concepts of non-linearity, systems far from equilibrium, and phase transitions. Before elaborating on those approaches, it should be made clear what systemic risk (of a system) is as compared to systemic importance (of one participant of a system).

**Systemic importance** – describes the significance of a single participant, or a group of participants, for the function of a system of many participants. This can be measured by, for example, network analysis [Arewa (2010)], in which the importance of a ‘node’ is equivalent to the percentage of other nodes disconnected from the network if this specific node is destroyed. Similar approaches to measure the systemic importance of individual institutions – especially of systemically important financial institutions. ...
institutions (SIFIS) – are reviewed in ECB (2009) and ECB (2010). This idea of how the ‘rest of a system’ behaves after a critical event is complementary to the perspective of how a system ‘as a whole’ is developing over time. Consequently, ‘systemic importance’ will not be covered in this paper, and the reader is referred to the current discussion.

**Systemic risk** – is related to the behavior of the system as a whole and describes the risk (or the probability) that a system becomes unstable. This instability does not necessarily arise from one ‘big’ participant, but from the interactions in the system. Also small contributions can lead to instability (the ‘butterfly effect’).

In the following, three different approaches to ‘systemic risk’ from the perspective of physics will be discussed:

- **Non-linearity and deterministic chaos** with the example of the magnetic pendulum.
- **Systems with a tremendous number of participants** (of a magnitude 10^23) far from equilibrium as examples for patterns in space and time.
- **Phase transitions in many-body systems**, such as magnetism.

### Non-linearity, unpredictability, and deterministic chaos

As stated by Jean-Pierre Landau, ‘non-linearity’ is a typical feature of complex systems and the behavior of dynamical systems with a non-linear differential equation of motion is highly sensitive to the initial conditions.

By human intuition, deterministic systems (i.e., systems which are fully described by equations such as Newton’s law etc.) should possess the property that if the initial conditions are close, then the resulting solutions are close. Also in deterministic non-linear systems, the future behavior is fully determined by their equation of motion and the initial conditions, and a single trajectory of a particle can be calculated. Nevertheless, these systems are often unpredictable, as even the smallest differences in the initial conditions render the trajectories to diverge exponentially^5. This is called a deterministic chaos. Remarkably, an early proponent of this concept was Henri Poincaré [Poincaré (1890)].

An impressive example for deterministic chaos is the magnetic pendulum: a metallic mass attached at the end of a pendulum bob that can move in all directions in the magnetic field of three magnets placed at the end of a triangle with its centre under the pendulum. If you start the pendulum from a non-center position, over which magnet will it ultimately end up? For any starting position of the pendulum projected into the x-y-plane of the magnets, one can indicate the end position and visualize that by different ‘colors’ representing the three different magnets, over which the pendulum will come to rest at the end (Figure 1).

\[
\dot{x} + \frac{g}{l} x + \frac{y}{m} \sum_{i=1}^{3} a_{ij} (x-x_i)^2 + d_i (x-x_i) = 0
\]
\[
\dot{y} + \frac{g}{l} y + \frac{y}{m} \sum_{i=1}^{3} a_{ij} (y-y_i)^2 + d_i (y-y_i) = 0
\]

Below left: end-positions for each starting point. That is, all white starting points (light grey, dark grey) represent positions that are ultimately attracted to the magnet in the center of the white (light grey, dark grey) region for a specific set of parameter values (related to the mass, damping, and attracting force parameters). Below right: a zoom of the region indicated by the circle. In this visualization, the details are also dependent on the numerical solution of the non-linear differential equation systems by Runge-Kutta method of 4th order and by the size of the pixels.

Figure 1 – Non-linear differential equation system for the motion of the magnetic pendulum.

Without further discussion of the details, which can be found in physics textbooks, the following issues are important:

- In the inner region with starting points near the final (equilibrium) position over a near-by magnet, this system is quite predictable but rather trivial.
- In the outer region with starting point far from the end position, the behavior is unpredictable, as smallest perturbations to the initial condition can result in a totally different end position. The more one tries to get accuracy by ‘zooming’ into the picture, the more structures will show up (see insert in Figure 1, especially with the thinnest readable lines, which will lead to a new structure by more zooming).
- The higher the value of the damping parameter for the pendulum, the bigger is the region with predictability. Nevertheless, somewhere the chaos will start.

The magnetic pendulum is not the only example: In many mechanical systems – from an ordinary game of pinball to the orbits of planets in stellar systems – the non-linearity of the equation of motion results in the unpredictability of the (far) future of the system.

---

5 The sensitivity to initial conditions is $d_n = d_0 e^{\lambda n}$ with the so called Lyapunov exponent $\lambda$. 
Although it might be non-trivial to derive an ‘equation of motion’ for a financial system based on the interaction of the participants, the simple mechanical example leads to two basic questions:

- Why should a – generically complex – financial system be more predictable (or less non-linear) than a simple magnetic pendulum?
- Are we willing to accept that deterministic systems show Knightian ‘unmeasurable uncertainty’?

However, there is also good news. Since the work of Ott et al (1990) it is known that small manipulations of a chaotic system can control chaos, so that they can direct chaotic trajectories to desired locations. This control strategy has been implemented in a wide variety of situations from mechanical systems and lasers to cardiac tissue and complex chemical reactions such as the Belousov-Zhabotinsky Reaction [Petrov et al. (1994)], which will be discussed below.

The explanation behind the capability to control chaos is closely connected with the characterizations of chaos, i.e., the exponential sensitivity of a system to smallest perturbations. Looking to Figure 1, it is clear by intuition that movement in the areas with clear final states is quite ‘stable’ and it would require a lot of energy to change the trajectory from one final state to another. But in a region where many trajectories with different final states are close together, it requires only a tiny push to move from a trajectory with one final state to a trajectory with another final state. Consequently, the smallest correction\(^6\) can help to control the trajectories in the chaotic regions and bring them back to a desired path. Unfortunately, the control strategy for chaotic systems with ‘minimum effort’ cannot easily be applied to financial systems with human beings as participants. While in chaotic systems, there is generally enough time to bring the trajectories ‘back to track,’ in the financial system with human actors, time is often the limiting factor.

Additionally, it may not be possible to determine the ‘equation of motion’ of each participant in the financial system, as this system is composed of very different participants. Consequently, the next section focuses on systems with so many participants that from an outside-in perspective, only an average behavior can be observed, but not any individual feature.

### Patterns and cycles in space and time – far from equilibrium

While a system such as the magnetic pendulum with one movable body is one extreme situation, the other extreme is a system with so many particles that we can treat them as a gas or a fluid, where we do not ‘see’ the individual particles but only a homogeneous system. As long as those systems contain only identical particles which interact simply ‘mechanically,’ like solid balls, these systems are the textbook examples of statistical physics. The ideal gas is the prototype with equilibrium, Gaussian distribution and all the laws of thermodynamics.

However, there are more complex systems, such as the Belousov-Zhabotinsky reaction: an oscillating chemical reaction of metal ions and bromic acid in a homogeneous solution. Boris Pavlovitsch Belousov discovered the first reaction of this class more or less by chance in 1958. As in thermodynamic equilibrium, the principle of detailed balance forbids oscillations in homogeneous systems, the majority of chemists at the time of Belousov’s discovery interpreted the results as some undetermined heterogeneous processes or simply as technical errors. However, what was incorrect was the assumption of thermodynamic equilibrium, and not Belousov’s results. It took until the end of the 1960s to accept this fact, when A. M. Zhabotinsky together with V. A. Vavilin and A. N. Zaikin showed that it is indeed a homogeneous reaction.

The reaction mechanism that they analyzed consists of three linked reactions:

\[
\begin{align*}
\text{HBrO}_3 + \text{HBrO}_2 & \rightarrow \ 2 \text{BrO}_2^- + \text{H}_2\text{O} \\
\text{H}^+ + \text{BrO}_2^- + \text{Fe(phen)}_3^{2+} & \rightarrow \ \text{Fe(phen)}_3^{3+} + \text{HBrO}_2 \\
\text{HBrO}_2 + \text{H}^+ + \text{Br}^- & \rightarrow \ 2 \text{HOBr}
\end{align*}
\]

The Belousov-Zhabotinsky reaction makes it possible to observe the development of patterns in space and time in a homogeneous but non-linear, non-equilibrium system by the naked eye on a very convenient human time scale of dozens of seconds and space scale of several millimeters.

Without going into details once again, the mathematical description of this chemical process is given by a dimensionless, non-linear differential equation system:

\[
\begin{align*}
\frac{dx}{dt} &= (qy - xy + x - x^2) / \epsilon \\
\frac{dy}{dt} &= (-qy - xy + fz) / \epsilon' \\
\frac{dz}{dt} &= x - z
\end{align*}
\]

It is important to mention that non-linearity can lead to chaos but can also generate patterns in space and time, even when the starting point is a fully homogeneous system but far from equilibrium.

However, the question is, are there similarities in financial systems? Karl Whelan described quite an analogous example in his 2009 report...
“Containing systemic risk,” [similar to the model of Brunnermeier (2009) and Brunnermeier and Pedersen (2009)] with three banks, each represented by their balance sheets. The starting point is for bank A (with symmetric permutations for bank B and bank C):

<table>
<thead>
<tr>
<th>Bank A Assets</th>
<th>Liabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loan to customers 100</td>
<td>Retail deposits 130</td>
</tr>
<tr>
<td>Loan to bank B 30</td>
<td>Borrowings from B 30</td>
</tr>
<tr>
<td>Loan to bank C 30</td>
<td>Borrowings from C 30</td>
</tr>
<tr>
<td>Other securities 40</td>
<td>Equity capital 10</td>
</tr>
<tr>
<td>Total 200</td>
<td>Total 200</td>
</tr>
<tr>
<td></td>
<td>Leverage ratio 200/10 = 20</td>
</tr>
</tbody>
</table>

The ‘chemical reactions’ in this system were described by Karl Whelan (Figure 2):

- Bank A makes a loss of 5 in an arbitrary currency on its loan book – equivalent to only 5/30 of the whole equity capital of the simple system.
- The loss is halving bank A’s equity capital to 5 with the resulting increase in its leverage ratio, most likely putting it close to or below its capital adequacy requirement. This forces bank A to fire sale some of its securities to the market with discount. According to accounting standards (mark-to-market or mark-to-model), the remaining assets also have to be reevaluated, and this in turn reduces bank A’s equity capital to 1.
- Banks B and C now also have to reevaluate their security holdings, reducing their equity capital to 6. Needing to shrink their balance sheets and worried about bank A’s solvency, they decide not to roll over their loans to bank A.
- Bank A now needs to come up with the liquidity to pay off the other banks, but with its equity almost zero and the market value of its securities falling, it fails to do so. Banks B and C now need to write off their loans to bank A (or bank bonds of bank A, etc.) and this, combined with the losses on their securities, probably reduces their equity capital also below capital adequacy requirement.

All in all, a loss of only 5 results in a collapse of the whole system, although the overall equity capital is 30, because of interlinking and certain reaction mechanisms such as cyclic capital adequacy requirement, accounting standards, and the difference between available liquidity and non-mature assets.

The non-linearity in this example comes from (i) the thresholds in capital adequacy requirement and (ii) the reevaluation of asset prices, due to IFRS also for banks B and C, although they want to hold them, or (iii) it could also be introduced by risk weight assumptions, which could ‘channel’ investments to no/lower weighted sovereign debt.

It will be the challenge for future research to find a way to ‘translate’ the schematic reaction, as shown in Figure 2, in the equivalent parameters of a non-linear differential equation system:

\[ \frac{dx}{dt} = f(x, \text{assets}_i, \text{Market}, a_{ij}, l_{ij}, \text{assets}_j, \text{Market}) \]

\[ x_i = \text{Cap.Ratio(bank}_i) \]

\[ a_{ij} = \text{assets of bank}_i \] with bank_j,

\[ l_{ij} = \text{liabilities of bank}_i \] to bank_j.

Examples like the Belousov-Zhabotinsky reaction show how ‘non-linearity + non-equilibrium’ can produce patterns in space and time. This is no contradiction to what was said in the previous section about deterministic chaos. Although we cannot predict the long-term future development of a non-linear deterministic chaotic system, this does not mean that this system will not show clear – and sometimes quite beautiful – patterns and structures. We simply cannot predict them ex-ante, but of course will see them ex-post.

**More patterns and cycles**

There is a second type of pattern in a non-linear system of interlinked participants, which is related to the distribution of energy between the states of the system. In classical statistical thermodynamics, systems develop over time into equilibrium with an equal distribution of the energy. In May 1955, the Los Alamos Scientific Laboratory published a technical report LA-1940 “Studies of nonlinear problems,” which can be seen as the birth of the FPU-problem named after the authors Enrico Fermi, John Pasta, and Stanislaw Ulam [Fermi et al. (1940)]. The FPU-problem
is related to the evolution of non-linear systems in time (i.e., a system of solid masses linked by non-linear springs). Fermi et al. expected that such a system, once excited into motion in a defined (lowest) oscillation mode, will distribute the energy equally to all possible oscillation modes of the system after some time. Solving the problem by one of the earliest computer simulations, they were astonished that the energy was indeed distributed after the start, but after some time up to 97% of the energy was once again concentrated in the initial state. This finding was verified in 1972 by James L. Tuck and Mary Tsingou and was called ‘recurrence’ or ‘super-recurrence’ on very long timescales.

In the FPU-problem, two issues, referred to as the FPU paradoxon, are important:

■ In such non-linear systems, energy is not distributed to all possible states – and not even Gaussian distributed – but the systems tend to ‘come back’ to a single state.

■ Such non-linear systems will show patterns in time with recurrence cycles on sometimes very long timescales.

The interested reader will find more details in recent reviews and books [Gallavotti (2007) and Dauxois and Peyrard (2006)]. In brief and with very much of physics in it, the explanation of the FPU paradoxon of recurrence was found in the existence of quasi-particles called ‘solitons,’ which are generated in such a finite system with periodic boundary conditions and, from time to time, come back to the positions they initially had, restoring the initial condition.

One should respect the warning of Paul A. Samuelson and be very skeptic about whether such physical systems have anything to do with financial systems, as we are far away from those simple models. Without any proof that there is a link between the FPU-problem and the following examples, two datasets from very complex systems indicate similar long-term periodic behavior (Figure 3):

■ The extinction density over hundreds of millions of years in the past (i.e., percentage of species on earth, which become extinct in a certain geological period of time) [Rohde and Muller (2005)].

■ The percentage of countries in default or restructuring (over the last 200 years) [Reinhart and Rogoff (2008)]

If those similarities could be ascribed to recurrence cycles due to non-linearity by further research, this would provide a starting point to explain the strange recurrence of situations in financial and ecological systems in contradiction to the assumption that these systems develop into equilibrium over a long time.

**Phase transitions or ‘magnetism’ in finance**

After looking at the non-linear behavior of ‘one particle’ as an example for deterministic chaos, and after looking at very big ‘macroscopic’ non-linear systems as examples for patterns and cycles, the third example will focus on many-body systems with an intermediate number of interaction participants (comparable to the financial system with hundreds or thousands of participants).

One characteristic phenomenon in many-body systems is the collective behavior, i.e., an alignment of the behavior of the participants as a result of interactions between them. Collective behavior can be found in different fields such as in nuclear many-body systems of protons and neutrons or solid state many-body systems like magnetic materials. A simple model used in statistical mechanics to describe those systems is the Ising model. The Ising model tries to imitate the behavior in which individual participants (i.e., atoms as elementary magnets) modify their behavior so as to conform to the behavior of other participants in their vicinity (nearest-neighbor interactions). This model was proposed 1925 by Ernst Ising. He tried to explain certain empirically observed facts about ferromagnetic materials using an approach proposed by his teacher W. Lenz in 1920.

In a general Ising model, the energy of a system is defined as

\[ H = -\sum_i H_i S_i - \sum_{i<j} H_{ij} S_i S_j - \sum_{k<j} H_{ikj} S_i S_j S_k + \ldots \]

in which the first term is the coupling to an (external) field, the second...
temperature \( T_c \). Figuratively, when we turn down the temperature, the interaction energy of this combination. And the sum of the spin orientations of the different atoms) of participant \( i \) or, respectively, \( j \).

When one sticks to the interaction of two participants, the typical term for the energy of the system – to be minimal – in the Ising model is:

\[
\sum J_{i,j} S_i S_j
\]

Here \( S_i \) and \( S_j \) are the orientations of the elementary magnets (i.e., the spin orientations of the different atoms) of participant \( i \) or, respectively, \( j \). \( J_{i,j} \) is the interaction energy of this combination. And the sum \( \Sigma \) runs over the nearest-neighbors \( j \) for a given \( i \). As a consequence of this interaction, phase transitions can occur in dependence of the temperature:

- From a phase without magnetic ordering (‘paramagnetic’) with an overall integrated magnetization of zero at high temperatures to a phase with a macroscopic magnetization (‘ferromagnetic’), i.e., alignment of the elementary magnets due to the interaction between the neighbors. The critical temperature \( T_c \) at which the magnetism vanishes is called Curie temperature \( T_C \).
- Figuratively, when we turn down the temperature, the critical point \( T_c \) is the state of the Ising model where the system minimizes its energy. Below this temperature, the system will order magnetically, meaning that the spins of the elementary magnets align with each other. The order parameter, which measures the deviation from the paramagnetic state, increases as the temperature decreases.

Concerning the financial crisis, this could – as a hypothesis – be translated as a phase transition from an ‘efficient market’ phase into a highly correlated phase, in which all the actors do the same as all the others they are interacting with.

The idea to use the Ising model approach to describe some circumstances of the financial crisis has three interesting features. First, the concept of nearest-neighbor interactions can reduce the complexity of hundreds of participants in the financial systems because only the nearest-neighbor interactions are counted. Second, nearest-neighbor interactions can consist of both financial interactions like in Karl Whelan’s simple example or behavioral interdependencies of the actors. And third, the notion of a phase transition could provide some insight into the question of why the efficient market model worked very well for a long time but collapsed in rather a short time into a ‘behavior driven’ world. Further research will be needed to test this idea and to compare predictions by the phase transition model with the reality.

### Conclusion

What is the benefit that can come from the perspective of physics on complex systems, when looking to the question of systemic risk in the financial world? Non-linear systems are by their very nature – following the title of a paper of Troy Shinbrot [Shinbrot (1993)] – ‘unpredictable yet controllable.’ Contrary to this understanding, many discussions about the financial crisis seem to focus on hopes that the financial system should be predictable.

When long-term stability is non-existing, a fundamental understanding of the development of a system is even more required to be able to apply means of control. Unfortunately, only setting other ‘initial condition’ can change the systems but might not result in principal changes in the non-linear chaotic behavior. Non-linear systems far from equilibrium show patterns in space and time, including recurrence cycles on sometimes very long timescales. The structural relations between financial systems of many participants and the development of complex systems could be helpful for the discussion about systemic risk and standard methodologies such as Monte-Carlo simulations could also provide insights into the development of such non-linear systems. However, the Belousov-Zhabotinsky reaction and the FPU-problem alone reveal two important features. First, equilibrium is often assumed but may be wrong – and the assumption of equilibrium is even not required to describe those complex systems. Second, recurrence and cycles are closely linked with non-linear interaction in finite many-body systems.

Phase transitions are rather typical for physical many-body systems; and the financial crisis is depicted by many as a change from an efficient
market (with independent actors) to herd behavior (with general alignment in trust or in mistrust). Models which only deal with the perspective of one participant (or a group of) can a priori not explain these transitions. Even with quite simple ‘nearest-neighbor’ interactions, a large number of features of many-body systems can be analyzed. Consequently, ‘magnetism in finance’ can open a window to study the structural changes of a market as a whole.’

In his work ‘Lombard Street,’ Walter Bagehot (1873) compared different banking systems. His elaboration can be summarized as follows: the social benefits of a developed banking system go hand in hand with proportional risk to be covered by the state itself. Either banks keep liquidity in cash (and are not able to hand out liquidity as loans to the economy), or the banking system is in need of a lender of last resort. In modern terminology, we can say that if banks interact with the participants of the economic systems (including other banks), there will be a risk of the system ‘as a whole.’ The perspective of physics on complex systems with interacting participants can be helpful for future research to get more insight into this essential problem of today’s financial systems.

References

- Arthur, W.B., S. Durlauf, and D. Lane, 1997, The economy as an evolving complex system II, Series in the Sciences of Complexity
- Brown, R., 1828 “A brief account of microscopic observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies,” Private Publication and various reprints, such as Philos Magazine, Ann. Philos, new series, 4, 161-173
- Knight, H.F., 1921, Risk, uncertainty, and profit, Harper, New York
- Landau, L.-P., 2009, Introductory remarks at the at the conference on “The macroeconomy and financial systems in normal times and in times of stress”, jointly organized by the Bank of France and the Deutsche Bundesbank, Gouvieux-Charilly, 8 June
- Thiet, T.J., 2009 “Systemic risk,” Clare Distinguished Lecture in Economics and Public Policy, University of Cambridge, 10 December
- Zhabotinsky Reaction,” Nature 361, 240-243
- Zhabotinsky Reaction,” Nature 361, 240-243
Abstract
The article analyzes the role of international supply chains as transmission channels of a financial shock. In these production networks, individual firms rely on each other, either as supplier of intermediate goods or client for their own production. An exogenous financial shock affecting a single firm, such as the termination of a line of credit, reverberates through the productive chain, with potential disruption effects. A resonance effect amplifies the back and forth interaction between real and monetary circuits when banks operate at the limit of their institutional capacity, defined by the capital adequacy ratio, and their assets are priced to market.

The transmission of the initial financial shock through real channels is tracked by modeling supply-driven international input-output interactions. The paper applies the proposed methodology on an illustrative set of interconnected economies: the U.S. and nine developed and developing Asian countries.

---

1 This article revises and updates a WTO Staff Working Paper published in 2009 (ERSD-2009-06). The views expressed in this document are those of the authors and do not represent a position, official or unofficial, of the WTO.
For the past 20 years, trade in tasks has been progressively competing with trade in final goods as the major driver of globalization. Global value chains have emerged as a dominant business model, based on the geographical dissociation of consumption and production and the fragmentation of the production processes within networks of firms. Nowadays, specific industrial operations, from the conception to the assembly of final products, are no longer undertaken by a single establishment but are increasingly outsourced within these global supply chains. It is becoming common practice for firms to process unfinished goods through affiliate or non affiliate firms. Indeed, most of the enormous growth in trade recorded in the last 20 years has consisted of relatively similar goods (manufactures) between relatively similar countries; moreover, a high proportion of trade takes place within industries rather than between them [Neary (2009)].

Economic theory predicts that if the production process of a final good can be segmented, then opportunities for economies of scale or scope may exist. In such a case, slicing the value chain into smaller segments leads to more efficient production, especially when done in an international context, due to wider differences in factor endowments and comparative advantages. However, the greater supply interconnection has also provided greater and faster channels of propagation of adverse external shocks. Because production is internationally diversified, adverse external shocks may affect firms not only through final demand (a sudden decline in exports), but also through a disruption of the flow of inputs received from their suppliers. Optimal chain length is, therefore, determined by the trade-off between the gains to specialization and the higher failure rate associated with longer chain length [Levine (2010)]. Indeed, the large drop in trade registered in 2008-2009 is attributed to the leverage effect induced by the geographical fragmentation of production [Tanaka (2009), Yi (2009)], albeit the estimation of trade elasticity in times of crisis is a complex matter [Escaith et al. (2010)].

While the financial and macroeconomic channels of shock transmission have received much attention, the role of industrial linkages as vectors of contagion remains to be thoroughly investigated. The disruptive potential of a failure in the international supply chain is becoming larger with time: trade in manufactures represented a quarter of the world industrial output in 2000, this proportion doubled in only five years. Almost 40% of this trade relates to the exchange of intermediate inputs and goods for processing, either traded between establishments pertaining to the same multinational enterprise, or exported to contracting parties for processing, then re-imported.

The objective of the article is to focus on the real transmission channels of financial shocks. The paper formalizes, from a combination of macroeconomic and sectorial perspectives, how financial restrictions – for example, a credit crunch initiating in a particular country – can disrupt a series of productive activities, affect worldwide production processes, and lead to self-sustained debt deflation. To do so, the paper develops an approach that builds on two interrelated concepts (input-output analysis and the monetary circuit) to model the sequence of financial and productive interactions along the international supply chain. Because firms rely on suppliers in carrying out part of the production process (outsourcing or off-shoring), and/or because they sell their production to other firms, the smooth realization of production plans from initial investment to final sales depends on the availability of credit at all stages of the production chain. An initial exogenous monetary shock, occurring, for example, when a bank shuts down an existing line of credit, will, therefore, replicate through the production chain. Through this ‘real transmission channel,’ the initial financial shock will propagate itself along the chain, affecting all firms in the supply network. Modeling how these supply-driven impulses propagate through open economies and feedback into the monetary circuit are the main objectives of this paper. An application is made on the U.S.-Asian case in 2000 and 2008. A section of conclusions presents the main findings.

The conceptual building blocks: monetary circuit and international input-output models

The turmoil that led to the 2008-2009 trade collapse was initially a financial crisis, which translated into a series of real effects. Modeling the parallel transmission of real and financial shocks along inter-industry channels is important to understand the systemic nature of the crisis. The approach used in this paper builds on two interrelated concepts inherited from the Physiocrats: the monetary circuit and input-output analysis.

The monetary circuit

Resolute believers that credits made deposits, Physiocrats viewed money as a medium of exchange, a mere ‘signe représentatif’ (token money). In this ‘entrepreneur’ economy, money circulates as a counterpart to goods and services. Physiocrats attributed to bankers the double responsibility of validating productive projects, extending the credit which would allow firms to produce those products. By the same stroke of a pen, they would also create the money which would allow their consumption by households. The monetary counterpart of production begins with credit granted by the banks to the producers, and it ends when the goods that were produced are sold and the initial loan is reimbursed (money is ‘destroyed’ at the end of the circuit).

The monetary circuit closely matches the production process. This process is divided into a finite number of stages, so that the output of one stage constitutes the input of the next – with the final stage yielding consumable output. All firms depend on credit to finance the current production costs (wages, intermediate consumption, and use of capital goods). The ‘temps du circuit,’ the elapsed time between money creation and its destruction, is closely related with the production time.
This monetary concept went through several years of relative neglect during the 1980s, when the focus of macroeconomics was controlling inflation. Controlling the quantity of an (exogenous) stock of money was central to monetarist policies. Since then, the instability of monetary aggregates and the contagion of financial crisis since the late 1990s called for alternative paradigms. It is certainly not a coincidence if endogenous money is making a ‘come back’ in the recent macroeconomic literature, like the New Institutional Economics [Stiglitz and Greenwald (2003)] and the Post Keynesians [Godley and Lavoie (2007)]. From the practitioner’s perspective, which guides this essay, the monetary circuit and its close links to production are attractive features when analyzing the actual functioning of an economy. Indeed, the concept of endogenous money is at the core of financial regulations such as minimum reserve requirements, which aim at controlling the capacity of banks to extend new loans and is closely associated with the establishment of international standards as those of Basel.

A canonical version of the monetary model starts with a request for credit by a firm to a bank in order to start a production process. Using the borrowed money, the firm purchases inputs and pays workers to produce the merchandises. The goods produced are sold to consumers or to other firms (if the firm produces investment or intermediate goods).

When the firm is paid, it uses the money resulting from the sale to repay its debt to the bank (plus interest rate), and its own suppliers if they had extended payment facilities. In order to simplify the model, without modifying the reasoning, it will be assumed that all profits made by the firm are redistributed to owners, so that there are no retained earnings. All the value-added created in the production process goes back to households as wages or distributed profits.

Since this is credit money, any money injected in the circuit is balanced by a debt obligation. The repayment of outstanding loans not only ‘destroys money,’ but also allows the bank to extend new credit within the limit of the prudential loans/assets ratio. The record of debt and its ratio with respect to the bank’s assets (gold in the Physiocrats’ perspective, or any asset considered as secure by the regulatory authorities in a contemporary context) is a key feature of our model. It acts as a bridge between flows and stocks, between real and monetary shocks, and between micro and macro effects. Because the price of the asset is also linked to the macroeconomic conjuncture (business cycle), the regulatory process is pro-cyclical: in phases of boom, asset prices go up, increasing the lending capacity of banks; when the business cycle is downward oriented, asset prices go down and banks have to cut on their credits in order to respect the prudential ratio.

The simple model is shown in Table 1. The first column is simply a book entry that tracks the net asset situation of the banking system. Each loan weighs on the adequacy ratio (a stock variable in this flow model) according to the risk attached to it. In turn, this risk, while specific to each firm (its own financial situation and that of its clients and key suppliers), depends also on the macroeconomic situation and the particular sensitivity of the sector to downturns.

<table>
<thead>
<tr>
<th>Time</th>
<th>Capital Adequacy Ratio</th>
<th>Flow of funds Bank account</th>
<th>Flow of funds Firm account</th>
<th>Flow of funds Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Initial situation</td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1. Credit</td>
<td>A/L</td>
<td>0/L-W+X(1)</td>
<td>W-X(1)</td>
<td></td>
</tr>
<tr>
<td>2. Production</td>
<td>A/L</td>
<td>0</td>
<td>L-W</td>
<td>W</td>
</tr>
<tr>
<td>3. First round of sales</td>
<td>A/L</td>
<td>0</td>
<td>L-W+X(1)</td>
<td>W-X(1)</td>
</tr>
<tr>
<td>4. Firm pays interest and distributes profit</td>
<td>A/L</td>
<td>0/L</td>
<td>L+X(1)-W-L-nX</td>
<td>W-X+πX+rL</td>
</tr>
<tr>
<td>5. Bank pays employees and distributes profits</td>
<td>A/L</td>
<td>0</td>
<td>L-X+X-W-L-nX</td>
<td>W-X+πX+rL</td>
</tr>
<tr>
<td>6. Second round of sales</td>
<td>A/L</td>
<td>0</td>
<td>L+X-W-L-nX</td>
<td>W-X-πX+rL</td>
</tr>
<tr>
<td>7. Firm repays loan</td>
<td>A</td>
<td>0</td>
<td>X-W-L-πX</td>
<td>W-X-πX+rL</td>
</tr>
</tbody>
</table>

Notes:
A: Initial assets of the bank, L: Loan from the bank to the firm to finance the production costs, c: risk weight attached to the loan (credit rating), W: wages needed to produce X, X: value of merchandises produced by the project; X(1) is the value sold during the first round (corresponding to wages paid by firms), X(2) is the amount that is sold later after profits are distributed and employees from the financial sector are paid. There is no savings, all profits are distributed and X(1)=X(2); r: interest rate, π: rate of profit after wages and operating costs.

Credit money is created ex nihilo when a loan is granted to a firm and its account is credited with a sum (L) that the firm will be able to use in order to pay for the goods and services it needs. The system is sustainable as long as (i) the production plan pays its costs and remunerates the stakeholders (salaries and distributed profits), i.e., X-W-L-nX is positive or nil, and (ii) the banking system does not exceed its adequacy ratio.

Because all profit is distributed to households and all income is consumed, the last position (No.7) is equivalent to the initial one (0), closing the circuit from a dynamic perspective. All credit money has been destroyed when the loan is repaid in full, and net flows sum up to zero. For
money to be destroyed all real transactions should take place as planned, i.e., there is no unsold final or intermediary goods. This characteristic of the monetary circuit provides an insight on an important property of the system: any stock of goods remaining in the ‘real’ system (where it is reported as gross investment in national accounts) has a counterpart in outstanding credit money in the financial circuit.

Outstanding inventories can be voluntary, when firms wish to smooth production and sales (i.e., to protect themselves from disruption in their production chain, or to be able to face a surge in demand). But stocks can be undesired when they correspond to negative shocks or when production plans based on ex-ante previsions prove to be too optimistic when confronted with the ex-post situation. Any accumulation of inventories, either desired or undesired, must be financed out of retained profits or bank credit. In practice, because firms have a structural saving gap, any increase in their inventories (assimilated to gross investment in national bank credit. In practice, because firms have a structural saving gap, any increase in their inventories (assimilated to gross investment in national accounts) will increase their net demand for credit and the quantity of outstanding credit money to be destroyed all real transactions should take place as planned, i.e., there is no unsold final or intermediary goods. This characteristic of the monetary circuit provides an insight on an important property of the system: any stock of goods remaining in the ‘real’ system (where it is reported as gross investment in national accounts) has a counterpart in outstanding credit money in the financial circuit.

Outstanding inventories can be voluntary, when firms wish to smooth production and sales (i.e., to protect themselves from disruption in their production chain, or to be able to face a surge in demand). But stocks can be undesired when they correspond to negative shocks or when production plans based on ex-ante previsions prove to be too optimistic when confronted with the ex-post situation. Any accumulation of inventories, either desired or undesired, must be financed out of retained profits or bank credit. In practice, because firms have a structural saving gap, any increase in their inventories (assimilated to gross investment in national accounts) will increase their net demand for credit and the quantity of outstanding credit money.

The procyclical nature of prudential ratios is a central feature of the model, and the object of much debate. In many reports on the implications of minimum capital-requirements, the potential restrictions are often qualified by mentioning that most banks hold capital in excess of the regulatory minima or are able to circumvent the binding constraints. According to Repullo and Suarez (2008), this ‘benign neglect’ of the potential procyclical effect is due to a series of misconceptions. The 2008-2009 crisis showed that larger than expected market swings, with deteriorating balance-sheet quality, severely limit access to equity and financial markets. As mentioned by Krugman (2008), in time of crisis, the core problem is capital rather than liquidity.

This very simple model grossly underestimates the complexity of the actual circuit. In reality, a multiplicity of simultaneous production plans are in place and the closure of the system (the sale of the production) does not depend on the wages and profits distributed by the producer, but on a stream of activities going on in the rest of the economy. In the same way, firms are not homogeneous: some produce mainly final goods, others investment or intermediate goods and the productive process can be fragmented among various establishments. Both the productive and monetary circuits are longer, and the elapsed time between initial and final positions (the ‘temps du circuit’) is increased. The longer the circuit, the larger the number of individual firms participating in the supply chain, the higher the probability of outstanding credit money.

The open monetary circuit

When the economy is open to international trade, domestic production competes with imports, but it can also be exported. Additionally, domestic production of final goods may include imported intermediate inputs, increasing the complexity of the process and the length of the circuit. Disregarding any differences in exchange and interest rates, a very simplified circuit involving two firms and two countries (a firm in home country producing a final good, and its supplier located in a foreign country) would look as in Table 2.

When the system is open to the rest of the world, a series of complications arises. Part of the purchasing power created during the production process is distributed in the foreign country while the final goods are sold in the home country. If X is not exported to the rest of the world, then the quantity produced will be greater than the quantity sold domestically (X > X(1)+X(2); even if there are no savings in the home country and all profits are distributed. Unless these final goods are exported, undesired stocks of finished products (X – [X(1)+X(2)]) will accumulate in the home country, associated to outstanding credit, while foreign households will accumulate savings for the amount of wages and profits created when processing the intermediate goods (Ww+rLw). Contrary to the case of a closed economy, the situation described by the final row is not identical to the initial one. In terms of national accounts, this appears as a trade deficit in the balance of payments of the home country (and a surplus for the rest of the world).7

The funds borrowed to finance production are used to purchase intermediate goods and services from other firms that may be located in different countries. In the same way, the production process depends on the capacity of the respective supplier firms to obtain credit from their own banks and deliver in time their intermediate inputs.

---

5 In a modern industrial system, firms cannot finance investment and production costs on their accumulated assets (initial capital plus retained earnings) and have to attract funding. For most firms, funding comes from loans rather than by issuing bonds or equities. Due to imperfect and asymmetric information, the Modigliani-Miller theorem does not hold and when firms are denied bank credit, they usually do not wish, nor are able, to raise capital by issuing new equity (Stiglitz and Greenwald (2003)).

6 Godley and Lavoie (2007) offer a detailed presentation of a complete stock-flow representation of the circuit. Although their approach is clearly built from a Post-Keynesian perspective regarding the capacity of banks to modulate their supply of credit, their description can be adapted to many other non-Walrasian theoretical settings, such as the loanable funds theory that competed with Keynesian theory since the 1930s or the Austrian school. Indeed, it is the flexibility of the monetary circuit in adapting to a number of theoretical settings that makes it very attractive from the practitioner’s perspective.

7 Opening the monetary circuit to cover balance of payments operations involves a series of complex interactions that are not treated in this very simple model. See Godley and Lavoie (2007) for an example.
The real circuit
One prominent and often discussed new element in contemporaneous business models is the emergence of global value chains, that built on outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages.

As a result, both exporters and importers face some uncertainty within their trading relationship. The higher the complexity of the goods being traded, the higher the uncertainty. To reduce those risks in times of crises, leading firms may extend short-term trade finance to their suppliers and extend payment facilities to their customers. This reduction in the risk of supply chain disruption is compensated by higher financial risks for the leading firm.

In addition, since the intermediate goods produced are not commodities, but are specific to the client’s need, it is usually not easy — and it is certainly costly — to shift to another supplier in case of disruption. The technological dimension of complexity related with the imperfect substitutability of inputs and the associated search costs on international markets is reviewed in Altomonte and Békés (2010). Firms dealing with very specific, low-substitutability goods that require particular production processes or specialized channels face higher trade complexity. As a corollary, the failure of any single supplier will affect the entire production chain in the short- and medium-term. At best, as a result of this supply shock, the client-firm will suffer an increase in costs of production when shifting to an alternative supplier; at worse, it will have to stop its production.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.

The greater industrial interconnection of the global economy has also created newer and faster channels for the propagation of adverse external shocks. Referring to the 2008-2009 breakdown, some authors have pointed out that these productive chains may explain the abrupt decrease of outsourcing and offshoring opportunities to develop comparative advantages. Among the structural changes that have impacted the way trade has been conducted since the last decade of the twentieth century is the geographical slicing up of the value chain into core and support activities, and the emergence of ‘trade in tasks.’ An early appraisal of the extent of outsourcing can be found in Feenstra (1998) who compares several measures of outsourcing and argues that all have risen since the 1970s. An illustrative example of a globalized value chain can be found in Linden et al. (2007), who study the case of Apple’s iPod.
At the microeconomic level, only anecdotic information on intra-firm interdependence is available, usually based on case studies. Fortunately, at macroeconomic and sectoral levels, it is possible to exploit the information provided by national accounts, linking them to form international input-output tables (IIO). Because IIOs track the inter-industry flows of intermediate goods in an international context, they can be used to model at industry level the real transmission channels of such a financial shock occurring at any stage of the production chains across the countries. In an IIO framework, the ‘real channel impact’ from a country of origin to a country of destination is proportional to (i) the foreign final demand for exported consumer goods and services, and (ii) the volume of trade in intermediate goods linking industries of both countries.

Supply-driven shocks are simulated through forward linkages, with IIO tables adapted into what is known as the Ghosh matrix (see Appendix 1). Ignoring final demand effects (the usual ‘Leontief approach’ based on backward linkages), the intensity of inter-country transmission of financial shocks following a credit crunch affecting the industrial sectors will differ according to the degree of vertical integration, as measured by the strength and depth of forward industrial linkages. Accordingly, we define the imported real supply-driven impact coefficient (IRSIC) as:

\[ \text{IRSIC} = \frac{\%\Delta P/P}{\Delta X(I-B)^{-1}} \]  

Where:  
\( X \) a row vector of initial sectoral output,  
\( \Delta X \) a row vector of supply-driven shocks,  
\( \%\Delta P/P \) a vector of price shocks in percentage,  
\( (I-B)^{-1} \) The Ghosh Inverse Matrix,  
\( \bullet \) denotes the Hadamard product.\(^{10}\)

The Ghosh multipliers computed on international I-O tables simulate the transmission of the higher production costs caused by the initial shock through the entire supply chain. By factoring-in the direct and indirect cost effects, it provides the analyst with an adequate ‘tracking methodology’ which incorporates transnational impacts. As trade in intermediate goods can be either imports or exports, similarly, the IIOs can be used to measure either the vulnerability to imported exogenous shocks, or the disruptive potential of exporting national industries’ problems to foreign industries.

As mentioned in the appendix, only non-disruptive supply-driven shocks – translating into an increase in production prices – can be captured by the indicator. The price shock hypothesis is obviously well-suited for segmented markets, with semi-monolithic characteristics, but is also compatible with standard neoclassical hypotheses, as long as marginal costs are increasing in the short term. It implies that client firms are always able to find a substitute supplier instantly, but at a higher cost. In addition, the distance of the initial shock to the final demand is taken into account by the indicator, because the Ghosh matrix ponders the suite of technical coefficients according to the proximity of each round to the initial demand. The closer the shock, the larger the impact.\(^{11}\)

These additional production costs are related to the elasticity of technical substitution in production, including elasticity of substitution between domestic and imported inputs (Armington elasticity). When Armington elasticities are low, large price changes are required to accommodate small changes in quantities. This is typically the case when manufactured intermediate inputs are differentiated and client-specific, at the difference of commodities like oil and minerals. Substitution in a supply constrained situation is difficult and takes time in the long run. In the short-run framework of this essay, it takes money and has a big impact on supply chain management: buying critical components on the spot market, paying premium air or sea fare rates to get material, and over-time compensations to subcontractors [Stadtler and Kilger (2008)].

**Unifying the two circuits: banks’ assessment of business risks and credit transition**

The previous two sections presented sequentially, from the monetary and supply chain perspectives, the linkages that exist within and between globalized industrial systems. The following section will formalize the connections between both financial and real circuits, through the modeling of the ‘\( \alpha \)’ parameter in the capital adequacy ratio that appears originally as a simple accounting device in the monetary circuit (see Table 1 and Table 2). This institutional ratio constrains the total amount of credit that a bank may issue. International standards, issued by the Bank for International Settlements (BIS) under Basel II, set a 8% threshold for total risk-weighted assets.

The monetary circuit starts with a request for credit by a firm to a bank in order to finance a production process. The bank’s decisions of granting the loan is based on a mix of (i) microeconomic considerations, directly related to the financial situation of the firm and the quality of its management; (ii) sectoral specificities, such as the cyclical nature of the business in which the firm operates; (iii) macroeconomic considerations, such as the probability of expansion or recession, and (iv) the institutional capacity of the bank to extend new credit within the limit of its loans/assets adequacy ratio.

The microeconomic component of the bank’s decision-making process is based, inter alia, on the direct supply-use connections of the firm requesting a loan. In a vertically integrated production chain, the default

---

\(^{10}\) Entry-wise product: \( A \times B = (a_1 b_1, a_2 b_2, ..., a_n b_n) \) on two matrices or vectors of same dimensions.

\(^{11}\) Because technical coefficients are less than unity, \( A^n \) (the impact of initial demands at the \( n^{th} \) stage of the production chain) rapidly tends to 0 when \( n \) increases. The incidence of the average length of the supply chains is analyzed in a following section.
of a client can cause distress to its suppliers, and the difficulties of a key supplier can jeopardize the viability of a production plan. Business cycles, beside their macroeconomic impact, have also industry-specific effects. Some industrial sectors (like construction or automobiles) are more ‘procyclical’ than others. According to the strength of its backward and forward linkage, the credit worthiness of a procyclical firm will reflect, through the microeconomic channel, on its direct clients and suppliers, even when they do not operate in the same sector.

Also, because risks and the market value of assets are strongly related, the position in the business cycle has a greatly inflated procyclical effect on the banks’ propensity to extend new loans. When a firm’s request for credit is turned down, it must scale down production; it affects in turn its suppliers and even its clients through the supply chain, and influence final demand through lower household income (wages and profits). For the most vulnerable sectors, that are both vertically and cyclically integrated, the conjunction of two waves of shocks, supply- and demand-driven, can lead to a first resonance effect. The total effect of the shock is then a multiple of each component taken in isolation.

This multiplier effect can be greatly amplified by a second resonance effect, when banks operate close to their lending limits. According to the monetary circuit described in Table 2, the shock will affect both flow variables (additional demand for credit money to cope with the shock) and stock variables (the credit rating of individual firms requesting additional loans, and the capital adequacy ratio of the banks extending loans). When a downward phase develops into a recession (i.e., when systemic risks materialize), the credit rating of many firms is downgraded and asset prices go down. The capital adequacy ratio is doubly affected and may reach a critical value, forcing banks to stop any new credit activity, canceling existing credit arrangements to reduce risk exposure irrespective of the merit of investment projects and firms creditworthiness. This situation defines what is called a ‘credit crunch.’

In a recessive cycle, the conjunction of real supply and demand shock, on the one hand, and of stock-flow financial shocks, on the other hand, may therefore have large systemic effects, even when the initial shocks are rather limited in scope. Thus, even if a financial shock is initially exogenous, its effects on the industrial chain will cause this shock to reverberate through the real circuit, affecting in turn the monetary circuit through the default risks, as perceived by the financial sector. The size of the multiplier effect depends on the initial balance sheets of the financial intermediaries.

Actually, modeling the weights ‘v_i’ (the financial rating of productive sectors) in the capital adequacy ratio (A_i/Σ Li) synthesizes most of the real-monetary dynamics embedded in the present approach, including stock-flow interactions:

- From a ‘flow’ perspective, the ‘v_i’ are the result of a credit rating process that considers both financial and productive aspects, associated with (i) the firm itself and its management (what we called the microeconomic dimension), (ii) its mode of insertion in the productive economy and the related risks of supply-driven shocks transmitted through the supply chain, and (iii) its exposure to the macroeconomic business cycle, as captured by the demand-driven shocks.

- From a ‘stock’ perspective, the real shocks translate into the accumulation of undesired stocks and extend the life expectancy of credit money, with the related accumulation of liquidity, because loans are not reimbursed in full. Not only money is not destroyed as expected, but on the financial side, the accumulation of bad debts deteriorates the banks’ balance sheet (loss provisions) and its capital adequacy ratio (A_i/Σ Li).

The ‘stock’ effect on the capital adequacy ratio is not limited to the domestic industries. In a globalized economy such as described in our model, the national financial sectors are also closely integrated and all economies now share ‘leveraged common creditors.’ In such a context, balance sheet contagion becomes pervasive [Krugman (2008)].

To resume, the chain of causalities, from a short-term perspective, is as follows:

a. The shock initiates in the monetary circuit (i.e., an existing line of credit is unexpectedly shut down), and it affects the production plans of a firm that are inserted at some point in a larger production chain. Because both its clients and suppliers cannot shift to other producers immediately and at no cost, the discontinuity in the production flow will reverberate through higher costs across the system represented by the IIO matrix. The real shock, once it has fully reverberated through the entire supply chains across industries and across countries, is measured by IRSIC and is proportional to the Ghosh coefficients that factor – in direct and indirect forward effects.

b. Exogenous supply-driven shocks also affect demand because increases in production costs reflect into higher prices and lead to lower demand. The resulting negative demand-driven secondary shocks can be modeled individually using the traditional Leontief model (final demand impulses), or by capturing the backward (demand) sectoral effects of supply-driven multipliers [Papadas and Dahl (1999)].

c. The real (demand and supply) shocks feed back into the monetary circuit through (i) the building-up of undesired stocks of finished and intermediate goods through the supply chain, leading to the accumulation of outstanding credit-money in the circuit; and (ii) the contagion of financial risks affecting the rating of firms. In a recessive cycle,

---

12 As mentioned, the core problem in the 2008/2009 financial crisis was capital, not liquidity.
<table>
<thead>
<tr>
<th>Origin of the shock, year:</th>
<th>From all manufacturing sectors, 2008</th>
<th>All other countries</th>
<th>Change 2000-2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>From China to:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>China 4.6 Indonesia 0.2 Japan 0.4</td>
<td>Thailand 0.7 U.S. 0.4</td>
<td>0.5 177</td>
</tr>
<tr>
<td>Mining</td>
<td>China 4.6 Indonesia 0.1 Japan 0.6</td>
<td>Thailand 0.4 U.S. 0.2</td>
<td>0.5 244</td>
</tr>
<tr>
<td>Agroindustries</td>
<td>China 0.2 Indonesia 0.5 Japan 0.8</td>
<td>Thailand 0.4 U.S. 0.7</td>
<td>0.6 82</td>
</tr>
<tr>
<td>Textile and clothing</td>
<td>China 1.2 Indonesia 4.4 Japan 3.3</td>
<td>Thailand 1.1 U.S. 1.2</td>
<td>2.5 85</td>
</tr>
<tr>
<td>Industrial machinery</td>
<td>China 3.8 Indonesia 1.8 Japan 3.7</td>
<td>Thailand 1.7 U.S. 3.6</td>
<td>2.7 254</td>
</tr>
<tr>
<td>Computers and electronic equipment</td>
<td>China 2.2 Indonesia 2.5 Japan 4.6</td>
<td>Thailand 4.6 U.S. 2.4</td>
<td>4.2 316</td>
</tr>
<tr>
<td>Other electrical equipment</td>
<td>China 2.0 Indonesia 2.2 Japan 4.4</td>
<td>Thailand 1.7 U.S. 6.8</td>
<td>3.7 240</td>
</tr>
<tr>
<td>Transport equipment</td>
<td>China 1.0 Indonesia 1.8 Japan 3.6</td>
<td>Thailand 2.4 U.S. 1.4</td>
<td>2.1 183</td>
</tr>
<tr>
<td>Other products</td>
<td>China 0.6 Indonesia 0.9 Japan 2.2</td>
<td>Thailand 0.8 U.S. 0.8</td>
<td>1.2 107</td>
</tr>
<tr>
<td>Utilities (water, gas, elect.)</td>
<td>China 6.7 Indonesia 0.2 Japan 0.4</td>
<td>Thailand 0.5 U.S. 0.7</td>
<td>0.3 57</td>
</tr>
<tr>
<td>Construction</td>
<td>China 19.3 Indonesia 1.1 Japan 2.4</td>
<td>Thailand 1.9 U.S. 1.1</td>
<td>1.6 221</td>
</tr>
<tr>
<td>Trade and transport services</td>
<td>China 8.8 Indonesia 0.3 Japan 0.2</td>
<td>Thailand 0.5 U.S. 0.2</td>
<td>0.3 178</td>
</tr>
<tr>
<td>Other services</td>
<td>China 7.8 Indonesia 0.3 Japan 0.4</td>
<td>Thailand 0.5 U.S. 0.5</td>
<td>0.3 185</td>
</tr>
<tr>
<td>Total</td>
<td>China 30.5 Indonesia 0.7 Japan 1.8</td>
<td>Thailand 1.6 U.S. 0.6</td>
<td>1.3 200</td>
</tr>
<tr>
<td>From Japan to:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>China 0.2 Indonesia 7.7 Japan 0.5</td>
<td>Thailand 0.4 U.S. 1.3</td>
<td>0.5 7</td>
</tr>
<tr>
<td>Mining</td>
<td>China 0.2 Indonesia 10.3 Japan 0.3</td>
<td>Thailand 0.7 U.S. 0.4</td>
<td>0.4 -22</td>
</tr>
<tr>
<td>Agroindustries</td>
<td>China 0.2 Indonesia 0.2 Japan 0.7</td>
<td>Thailand 0.4 U.S. 1.0</td>
<td>0.6 -8</td>
</tr>
<tr>
<td>Textile and clothing</td>
<td>China 0.5 Indonesia 0.4 Japan 1.2</td>
<td>Thailand 1.6 U.S. 0.4</td>
<td>1.0 -40</td>
</tr>
<tr>
<td>Industrial machinery</td>
<td>China 1.4 Indonesia 4.9 Japan 2.9</td>
<td>Thailand 2.3 U.S. 7.5</td>
<td>3.5 -15</td>
</tr>
<tr>
<td>Computers and electronic equipment</td>
<td>China 3.6 Indonesia 1.5 Japan 3.0</td>
<td>Thailand 4.3 U.S. 7.4</td>
<td>5.8 -21</td>
</tr>
<tr>
<td>Other electrical equipment</td>
<td>China 2.3 Indonesia 1.4 Japan 3.0</td>
<td>Thailand 4.3 U.S. 5.2</td>
<td>3.1 -29</td>
</tr>
<tr>
<td>Transport equipment</td>
<td>China 1.4 Indonesia 1.6 Japan 2.9</td>
<td>Thailand 3.8 U.S. 2.1</td>
<td>2.7 -46</td>
</tr>
<tr>
<td>Other products</td>
<td>China 0.7 Indonesia 0.6 Japan 1.8</td>
<td>Thailand 2.1 U.S. 1.2</td>
<td>1.2 -36</td>
</tr>
<tr>
<td>Utilities (water, gas, elect.)</td>
<td>China 0.4 Indonesia 0.2 Japan 1.4</td>
<td>Thailand 0.8 U.S. 0.6</td>
<td>0.3 -43</td>
</tr>
<tr>
<td>Construction</td>
<td>China 0.9 Indonesia 12.0 Japan 1.7</td>
<td>Thailand 2.6 U.S. 2.4</td>
<td>1.6 -13</td>
</tr>
<tr>
<td>Trade and transport services</td>
<td>China 0.4 Indonesia 3.6 Japan 0.3</td>
<td>Thailand 0.6 U.S. 0.5</td>
<td>0.4 -21</td>
</tr>
<tr>
<td>Other services</td>
<td>China 0.4 Indonesia 4.2 Japan 0.4</td>
<td>Thailand 0.7 U.S. 0.6</td>
<td>0.5 -10</td>
</tr>
<tr>
<td>Total</td>
<td>China 0.8 Indonesia 19.4 Japan 1.3</td>
<td>Thailand 2.1 U.S. 1.4</td>
<td>1.3 -22</td>
</tr>
<tr>
<td>From Korea to:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>China 0.1 Indonesia 0.1 Japan 8.0</td>
<td>Thailand 0.3 U.S. 0.3</td>
<td>0.2 2</td>
</tr>
<tr>
<td>Mining</td>
<td>China 0.2 Indonesia 0.0 Japan 3.9</td>
<td>Thailand 0.3 U.S. 0.5</td>
<td>0.2 10</td>
</tr>
<tr>
<td>Agroindustries</td>
<td>China 0.2 Indonesia 0.1 Japan 42.7</td>
<td>Thailand 0.4 U.S. 0.3</td>
<td>0.2 1</td>
</tr>
<tr>
<td>Textile and clothing</td>
<td>China 0.5 Indonesia 0.8 Japan 0.3</td>
<td>Thailand 0.7 U.S. 0.5</td>
<td>0.5 -55</td>
</tr>
<tr>
<td>Industrial machinery</td>
<td>China 1.0 Indonesia 0.8 Japan 0.5</td>
<td>Thailand 0.9 U.S. 1.6</td>
<td>0.9 11</td>
</tr>
<tr>
<td>Computers and electronic equipment</td>
<td>China 2.9 Indonesia 0.7 Japan 0.6</td>
<td>Thailand 2.6 U.S. 1.6</td>
<td>1.7 6</td>
</tr>
<tr>
<td>Other electrical equipment</td>
<td>China 1.8 Indonesia 0.7 Japan 0.5</td>
<td>Thailand 2.1 U.S. 1.8</td>
<td>1.2 3</td>
</tr>
<tr>
<td>Transport equipment</td>
<td>China 0.8 Indonesia 0.4 Japan 1.8</td>
<td>Thailand 0.9 U.S. 1.1</td>
<td>0.7 -11</td>
</tr>
<tr>
<td>Other products</td>
<td>China 0.6 Indonesia 0.3 Japan 0.3</td>
<td>Thailand 0.8 U.S. 0.5</td>
<td>0.5 -31</td>
</tr>
<tr>
<td>Utilities (water, gas, elect.)</td>
<td>China 0.3 Indonesia 0.1 Japan 0.0</td>
<td>Thailand 0.3 U.S. 0.6</td>
<td>0.2 -21</td>
</tr>
<tr>
<td>Construction</td>
<td>China 0.7 Indonesia 0.5 Japan 11.0</td>
<td>Thailand 0.9 U.S. 0.6</td>
<td>0.5 9</td>
</tr>
<tr>
<td>Trade and transport services</td>
<td>China 0.3 Indonesia 0.2 Japan 4.5</td>
<td>Thailand 0.2 U.S. 0.4</td>
<td>0.2 -7</td>
</tr>
<tr>
<td>Other services</td>
<td>China 0.4 Indonesia 0.1 Japan 4.6</td>
<td>Thailand 0.3 U.S. 0.3</td>
<td>0.2 21</td>
</tr>
<tr>
<td>Total</td>
<td>China 0.7 Indonesia 0.3 Japan 24.6</td>
<td>Thailand 0.9 U.S. 0.6</td>
<td>0.5 -2</td>
</tr>
<tr>
<td>From Malaysia to:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>China 0.0 Indonesia 0.1 Japan 0.0</td>
<td>Thailand 0.4 U.S. 0.1</td>
<td>0.1 31</td>
</tr>
<tr>
<td>Mining</td>
<td>China 0.0 Indonesia 0.1 Japan 0.0</td>
<td>Thailand 0.2 U.S. 0.2</td>
<td>0.0 15</td>
</tr>
<tr>
<td>Agroindustries</td>
<td>China 0.0 Indonesia 0.0 Japan 0.1</td>
<td>Thailand 0.2 U.S. 0.2</td>
<td>0.0 -13</td>
</tr>
<tr>
<td>Textile and clothing</td>
<td>China 0.1 Indonesia 0.3 Japan 0.1</td>
<td>Thailand 0.2 U.S. 0.2</td>
<td>0.1 -12</td>
</tr>
<tr>
<td>Industrial machinery</td>
<td>China 0.2 Indonesia 0.6 Japan 0.1</td>
<td>Thailand 0.4 U.S. 0.3</td>
<td>0.1 -13</td>
</tr>
<tr>
<td>Computers and electronic equipment</td>
<td>China 1.3 Indonesia 0.5 Japan 0.3</td>
<td>Thailand 1.3 U.S. 2.5</td>
<td>1.0 23</td>
</tr>
<tr>
<td>Other electrical equipment</td>
<td>China 0.6 Indonesia 0.5 Japan 0.2</td>
<td>Thailand 0.6 U.S. 2.3</td>
<td>0.7 9</td>
</tr>
<tr>
<td>Transport equipment</td>
<td>China 0.1 Indonesia 0.3 Japan 0.2</td>
<td>Thailand 0.3 U.S. 0.8</td>
<td>0.1 7</td>
</tr>
<tr>
<td>Other products</td>
<td>China 0.1 Indonesia 0.3 Japan 0.1</td>
<td>Thailand 0.2 U.S. 0.3</td>
<td>0.1 -30</td>
</tr>
<tr>
<td>Utilities (water, gas, elect.)</td>
<td>China 0.1 Indonesia 0.2 Japan 0.1</td>
<td>Thailand 0.3 U.S. 0.3</td>
<td>0.0 -22</td>
</tr>
<tr>
<td>Construction</td>
<td>China 0.1 Indonesia 0.4 Japan 0.1</td>
<td>Thailand 0.2 U.S. 0.6</td>
<td>0.1 -1</td>
</tr>
<tr>
<td>Trade and transport services</td>
<td>China 0.1 Indonesia 0.2 Japan 0.0</td>
<td>Thailand 0.2 U.S. 0.1</td>
<td>0.1 20</td>
</tr>
<tr>
<td>Other services</td>
<td>China 0.1 Indonesia 0.1 Japan 0.0</td>
<td>Thailand 0.2 U.S. 0.2</td>
<td>0.1 39</td>
</tr>
<tr>
<td>Total</td>
<td>China 0.2 Indonesia 0.2 Japan 0.1</td>
<td>Thailand 0.3 U.S. 0.5</td>
<td>0.1 20</td>
</tr>
</tbody>
</table>
both value-to-market asset pricing and the accumulation of bad loans affect banks’ capital adequacy ratio.

d. Because the capacity of banks to create new money is limited by their capital adequacy ratio, their capacity to extend new credit is severely constrained, initiating a vicious circle. Credit crunch affects the most vulnerable firms (those more closely connected to the sector of activity affected by the initial shock), but has also a systemic impact through the supply chain effect that affect other sectors of activity and reduces the banks’ systemic capacity to extend new credit, regardless of the individual merits of the investment programs.13

A simulation on the international U.S. - Asia production compact

This reduced model of contagion through the supply chain only requires two variables to simulate and track the systemic implications of an exogenous financial shock: one flow variable (coefficient IRSIC, constructed on the real circuit, and possibly augmented for secondary demand-driven effects) and one stock variable (capital adequacy ratio, derived from the monetary circuit). Because the stock-variable is partially dependent on the flow-variable, the only strategic variable to be measured in order to evaluate the risk of contagion from supply shocks is IRSIC.

We build our case study on ten economies (China, Indonesia, Japan, Korea, Malaysia, Chinese Taipei, Philippines, Singapore, Thailand, and the U.S.). All these economies are key international or regional traders at different stages of industrial development and with strong specificities in terms of their insertion in the global economy. We use a subset of the Asian international input-output tables (AIO tables) developed by the Institute of

<table>
<thead>
<tr>
<th>Origin of the shock, year:</th>
<th>From all manufacturing sectors, 2008 a</th>
<th>All other countries a</th>
<th>Change 2000-2008 b</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Thailand to:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>China 0.0, Indonesia 0.1, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.4, Taipei 0.1, Philippines 0.2, Thailand 9.1, U.S. 0.0</td>
<td>0.1, Average 80</td>
</tr>
<tr>
<td>Mining</td>
<td>China 0.0, Indonesia 0.1, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.3, Taipei 0.0, Philippines 0.2, Thailand 2.5, U.S. 0.0</td>
<td>0.1, Average 122</td>
</tr>
<tr>
<td>Agroindustries</td>
<td>China 0.0, Indonesia 0.1, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.9, Taipei 0.2, Philippines 0.2, Thailand 0.0, U.S. 0.0</td>
<td>0.2, Average 34</td>
</tr>
<tr>
<td>Textile and clothing</td>
<td>China 0.1, Indonesia 0.2, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.7, Taipei 0.2, Philippines 0.2, Thailand 0.1, U.S. 0.1</td>
<td>0.2, Average 39</td>
</tr>
<tr>
<td>Industrial machinery</td>
<td>China 0.1, Indonesia 1.5, Japan 0.2, Korea 0.1</td>
<td>Malaysia 0.7, Taipei 0.2, Philippines 0.3, Thailand 0.1, U.S. 0.4</td>
<td>0.4, Average 151</td>
</tr>
<tr>
<td>Computers and electronic equipment</td>
<td>China 0.9, Indonesia 0.3, Japan 0.2, Korea 0.3</td>
<td>Malaysia 0.9, Taipei 0.5, Philippines 0.6, Thailand 0.2, U.S. 0.1</td>
<td>0.5, Average 2</td>
</tr>
<tr>
<td>Other electrical equipment</td>
<td>China 0.4, Indonesia 0.4, Japan 0.2, Korea 0.2</td>
<td>Malaysia 1.0, Taipei 0.3, Philippines 0.3, Thailand 0.1, U.S. 0.4</td>
<td>0.4, Average 26</td>
</tr>
<tr>
<td>Transport equipment</td>
<td>China 0.1, Indonesia 0.5, Japan 0.2, Korea 0.1</td>
<td>Malaysia 1.3, Taipei 0.2, Philippines 0.9, Thailand 0.1, U.S. 0.4</td>
<td>0.4, Average 106</td>
</tr>
<tr>
<td>Other products</td>
<td>China 0.1, Indonesia 0.2, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.7, Taipei 0.1, Philippines 0.2, Thailand 0.1, U.S. 0.2</td>
<td>0.2, Average 9</td>
</tr>
<tr>
<td>Utilities (water, gas, elect.)</td>
<td>China 0.0, Indonesia 0.1, Japan 0.0, Korea 0.0</td>
<td>Malaysia 0.2, Taipei 0.0, Philippines 0.2, Thailand 0.0, U.S. 0.1</td>
<td>0.1, Average 17</td>
</tr>
<tr>
<td>Construction</td>
<td>China 0.1, Indonesia 0.3, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.6, Taipei 0.1, Philippines 0.2, Thailand 6.9, U.S. 0.0</td>
<td>0.2, Average 58</td>
</tr>
<tr>
<td>Trade and transport services</td>
<td>China 0.1, Indonesia 0.2, Japan 0.0, Korea 0.0</td>
<td>Malaysia 0.3, Taipei 0.0, Philippines 0.2, Thailand 3.5, U.S. 0.0</td>
<td>0.1, Average 99</td>
</tr>
<tr>
<td>Other services</td>
<td>China 0.1, Indonesia 0.2, Japan 0.0, Korea 0.0</td>
<td>Malaysia 0.2, Taipei 0.0, Philippines 0.2, Thailand 5.5, U.S. 0.0</td>
<td>0.1, Average 59</td>
</tr>
<tr>
<td>Total</td>
<td>China 0.1, Indonesia 0.2, Japan 0.1, Korea 0.1</td>
<td>Malaysia 0.6, Taipei 0.1, Philippines 0.2, Thailand 24.9, U.S. 0.0</td>
<td>0.2, Average 16</td>
</tr>
</tbody>
</table>

| From U.S. to:             |                                       |                      |                   |
| Agriculture               | China 0.1, Indonesia 0.1, Japan 0.3, Korea 0.4 | Malaysia 0.4, Taipei 0.7, Philippines 0.3, Thailand 0.4, U.S. 8.8 | 0.3, Average -2 |
| Mining                    | China 0.1, Indonesia 0.1, Japan 0.3, Korea 0.2 | Malaysia 0.6, Taipei 0.2, Philippines 0.4, Thailand 0.1, U.S. 3.4 | 0.2, Average -18 |
| Agroindustries            | China 0.1, Indonesia 0.2, Japan 0.3, Korea 0.5 | Malaysia 0.8, Taipei 0.9, Philippines 0.3, Thailand 0.4, U.S. 0.4 | 0.4, Average -28 |
| Textile and clothing      | China 0.2, Indonesia 0.3, Japan 0.4, Korea 0.6 | Malaysia 1.1, Taipei 1.5, Philippines 0.2, Thailand 0.4, U.S. 0.4 | 0.6, Average -43 |
| Industrial machinery      | China 0.4, Indonesia 1.5, Japan 0.7, Korea 1.0 | Malaysia 1.5, Taipei 1.6, Philippines 0.6, Thailand 1.4, U.S. 1.1 | 1.1, Average -28 |
| Computers and electronic equipment | China 1.4, Indonesia 0.5, Japan 0.7, Korea 1.4 | Malaysia 4.0, Taipei 2.4, Philippines 5.6, Thailand 2.1, U.S. 2.3 | 2.3, Average -25 |
| Other electrical equipment | China 0.8, Indonesia 0.6, Japan 0.7, Korea 1.3 | Malaysia 3.5, Taipei 1.9, Philippines 1.0, Thailand 2.0, U.S. 1.5 | 1.5, Average -39 |
| Transport equipment       | China 0.5, Indonesia 0.4, Japan 0.9, Korea 1.2 | Malaysia 1.5, Taipei 1.1, Philippines 0.6, Thailand 1.2, U.S. 0.9 | 0.9, Average -28 |
| Other products            | China 0.3, Indonesia 0.3, Japan 0.4, Korea 0.6 | Malaysia 1.0, Taipei 0.9, Philippines 0.4, Thailand 0.4, U.S. 0.5 | 0.5, Average -45 |
| Utilities (water, gas, elect.) | China 0.1, Indonesia 0.1, Japan 0.1, Korea 0.1 | Malaysia 0.4, Taipei 0.0, Philippines 0.3, Thailand 0.0, U.S. 2.1 | 0.1, Average -47 |
| Construction              | China 0.3, Indonesia 0.3, Japan 0.3, Korea 0.5 | Malaysia 0.9, Taipei 0.8, Philippines 0.3, Thailand 0.5, U.S. 11.4 | 0.5, Average -17 |
| Trade and transport services | China 0.1, Indonesia 0.2, Japan 0.1, Korea 0.1 | Malaysia 0.3, Taipei 0.1, Philippines 0.4, Thailand 0.1, U.S. 3.7 | 0.2, Average -33 |
| Other services            | China 0.2, Indonesia 0.2, Japan 0.1, Korea 0.2 | Malaysia 0.4, Taipei 0.3, Philippines 0.4, Thailand 0.3, U.S. 3.8 | 0.2, Average -28 |
| Total                     | China 0.3, Indonesia 0.2, Japan 0.3, Korea 0.5 | Malaysia 1.5, Taipei 0.9, Philippines 0.9, Thailand 0.6, U.S. 14.3 | 0.7, Average -29 |

Notes: a/ direct and indirect impacts of a 30% increase in the price of inputs originating from the manufacturing sectors, in per cent of the respective sectorial production costs, weighted averages. b/ manufacturing sectors are “Agroindustries,” “textile and clothing,” “industrial machinery,” “computers,” “electrical equipment,” “transport equipment,” and “other products.” c/ simple average on partner countries, excluding domestic and “rest of world,” and percent change between 2000 and 2008.

Source: authors’ calculations

Table 3 Transmission of an initial 30% price shock from manufacturing sectors, 2000-2008 * (percentage)

13 When the adequacy ratio is reaching a critical limit, the banking sector turns down most loan requests and flies for safety by investing in good quality government bonds, especially the U.S. bonds. The flight for quality that followed the 2008 subprime crisis and the subsequent melting down of the international banking system illustrate this point and explain why the dollar appreciated despite the fact that the U.S. economy was at the core of the crisis.
Developing Economies, for year 2000 [DE-Jetro (2006)]. A 2008 estimate was derived from the AIO matrix, incorporating updated information on multilateral trade and national accounts aggregates in current US$.  

**Simulation results**

Supply-driven shocks occurring in the regional sourcing network are modeled as price shocks emanating from one of the ten economies linked in the international I-O table. An arbitrary value of 30% will be used for the size of the shock, and all manufacturing sectors are shocked simultaneously. The simulation computes the domestic impacts and its transmission to the other regional partners through the imported real supply-driven impact coefficient (IRSIC) as previously described. As IRSIC uses the Ghosh Inverse Matrix, the sectoral impacts include primary and secondary effects (i.e., the real transmission channels follow both direct and indirect forward linkages). 2000 and 2008 results are not directly comparable because of exchange rate movements and changes in the composition of trade with the rest of the world that occurred between these two years; in addition, 2008 is based on an estimate. Nevertheless, the evolution of the IRSIC values provides relevant information on the direction of changes.

As seen in Table 3, the largest secondary impacts from a price shock are felt domestically. The relative effect on the domestic economy depends negatively on its degree of openness and on the relative size of the originating sector in relation to the rest of the economy. As expected, manufacturing industries are more sensitive to imported shocks originating from foreign manufacturers, especially transport equipments. With the exception of Thailand (and Philippines, not shown in the table), the national impact of a shock originating in the domestic manufacturing sector tends to decrease between 2000 and 2008, indicating a greater openness to imported inputs and/or a greater participation of non-manufacture domestic inputs in the domestic content of manufacturing sectors.

Between 2000 and 2008, China emerged as a key element of the regional and international supply chains. This trend is particularly clear in the industry of computer equipment, where its IRSIC values for exported shocks increased fourfold. Thailand followed a similar trend, albeit with less strength, and its role as international supplier increased in all sectors, except in textile and clothing. Conversely, the role of others as key suppliers decreased. The influence of Malaysia (and Chinese Taipei, not shown in the table) lowered in industrial sectors, although it increased in other sectors; the relative role of Japan and the U.S. declined across the industrial board.

To visualize the changes in the relative position of each country between 2000 and 2008, Table 4 presents a summary of the results obtained for shocks initiating from their manufacturing sectors. Based on the existing intra and inter-industrial linkages, in 2000 and 2008, Japan is potentially the largest exporter of price shocks. Nevertheless, its dominant position as supplier of intermediate inputs has been eroded from 2000 to 2008 due to the rise of other competitors (from the region or from the rest of the world), China, which ranked 4th in 2000, gained two spots and became the second potential exporter of supply shocks through its industrial sectors. On the contrary, Korea and the U.S. retrograded at the 3rd and 4th position in 2008. The relative position of other economies remained unaffected.

Malaysia remains the first importer of shocks in both 2000 and 2008 because of the high degree of integration of its manufacturing sectors and reliance on imported inputs from other partners. Thailand became more reliant on imported parts and components and ranked second among importers in 2008, gaining two spots. At the other extreme of the spectrum, we found large economies with a relatively small external sector compared with their domestic activities. It is the case of the U.S. and Japan, two advanced economies where services dominate the GDP structure, but also of China and Indonesia.

**Further considerations on the simulation results**

**Disruptive shocks**

One should remember that IRSIC works only for non-disruptive supply impacts, maintaining constant production levels. It may sub-estimate the intensity of an imported supply-driven shock for two reasons:

- For the relatively less developed countries, if the shock originates from an industrialized country or from a domestic trade credit crunch, it might become disruptive since the affected national firm cannot shift easily to another domestic or external supplier.
- For the most advanced industrialized countries, there is always the possibility of substituting domestically an intermediate input produced in a less developed country. But the increase in production costs may be much higher than the standard 30% used in the simulation, due to, inter alia, the difference in the cost of factorial services (primary inputs).

As discussed in Appendix 1, the input-output framework is inappropriate to measure the first type of disruptive shocks because the combination of strict complementarities of inputs and forward linkages would progressively bring the economy to an almost complete halt. In a more realistic scenario, one can consider that the affected export-oriented activities would stop, generating a severe macroeconomic shock to the economy.

---

14 See Escaith and Gonguet (2009), Inomata and Uchida (2009), and Pula and Peltonen (2009) for examples.

15 This conservative option may underestimate the price impact of a supply shock for developed countries for labor intensive products, if the alternate suppliers have to be found in the domestic market (see next section).
For Japan and the U.S., the induced rise in domestic prices caused by a shutdown of their Asian suppliers of intermediate manufacture goods would be significant when differences in production costs are imputed on the basis of differences in wage costs.\textsuperscript{16} This is the case especially for the textile and clothing industries. The disruption of supply chains in the manufacturing sectors of the three developing Asian countries would lead to a 3.2% average increase in the average price of manufacture outputs in Japan, and 2.6% in the U.S. Textile and clothing are particularly vulnerable, suffering an increase in production cost of 13% in Japan and 8.5% in the U.S.

Considering that only a minority of firms engage in off-shoring, this average sectorial impact will fall disproportionately on a few firms – probably the most dynamic ones – with potentially large disruptive microeconomic impacts as their production costs will rise by a multiple of the average rate. Incidentally, the results of this simulation also show the potential gains and competitive advantage those vertically integrated firms were able to obtain in the first place, when outsourcing part of their production to emerging Asia.

**Average propagation length and the bullwhip effect**

The topology of a supply chain and the number of individual firms which participate have an impact on the disruptive intensity of a shock affecting any one of the participants. In a linear setting, where tasks are successively performed on goods in process of manufacturing, the disruption of one segment will affect all downward players in the real sphere. Failure of a single link may cause a ‘cascade’ of failures, with amplified effects. The scale of the damage depends on the length of the supply chain. As analyzed by Levine (2010), longer production chains are subject to a ‘weakest link’ effect, they are more fragile and more prone to failure. In addition, the more complex and specialized the production process, the more difficult it is to find alternative suppliers and the more disruptive the impact is. Even in simpler cases, the well-known ‘bullwhip’ effect [Stadtler and Kigler (2008)] amplifies the amplitude of the shock along the supply chain, as affected firms run down the inventories in the face of

\textsuperscript{16} The purchasing power parity ratio between developed and developing countries reflects mainly the cost of non-traded services, and is about 0.40. On this basis, factorial services imbedded in the total production costs, besides tradable intermediate inputs, are deemed 150% more expensive in the U.S. and Japan than in the developing Asian countries.
uncertainty. A slowdown in activity may transform itself into a complete standstill for the supplying firms that are located upstream because of the amplified fluctuations in ordering and inventory levels in the management of production-distribution systems.

In the monetary circuit, the costs already incurred by upward firms cannot be recovered nor credit reimbursed as long as the final good is not produced and sold. As money is not destroyed, outstanding loans accumulate in the circuit, their quality decreases, and systemic risks increase. Levine (2010) reflects on the similarity between financial interconnection and the interdependence revealed by the disruption of a long chain of production caused by the failure of a single producer. The greater the specialization of the supply chain, the greater the efficiency returns to specialization but also the higher the risks.

Although it is not possible to differentiate individual firms and supply chains in an input-output setting, a rough measure of the depth of supply chains can be given by the average propagation length (APL) of a shock. Based on the ability of an inverse Leontief or Ghosh matrix to trace both direct and indirect impacts, APL is formulated as a weighted average of the number of production stages through which the impact from industry j goes until it ultimately reaches industry i. At each iteration, the net impact is used as a weight; it tends to zero when the number of iterations increases. APL is closely related to the notion of vertical integration, as shown by Inomata (2008).

Compared to the specialized supply chains analyzed by Levine (2010) in the microeconomic referent, the aggregated IIO approach considers only undifferentiated shock among largely independent firms, where the amplitude of the shock falls rapidly with the length of the chain. In other words, APL estimated on IIOs will almost certainly underestimate the extent of the issue at hand, especially when considering that only a few firms participate in international outsourcing.

Based on Dietzenbacher and Romero (2007), we computed APL on Ghosh matrices for year 2008 (see appendix 2). Furthermore, we rescaled the resulting APLs in order to correct for the bias identified by these authors, obtaining APL$^\$ as per equation [3], in appendix 2. The simulation was done restricting the measure of the indicator to manufacturing sectors, and discarding domestically induced impacts. The results, shown in Figure 1, provide a measure of the economic relevance of the average propagation length, by sector.

Rescaled APL$^\$ for computers and electronic equipment show that it is by far the dominant sector, when the intensity and the length of the shock are jointly considered. The relative size of the sector ranking second, metals and metal product, represents only 40 percent of the former. The economic weight of computers and electronic equipment is due to both a relatively high APL and a large Ghosh coefficient, reflecting the fact that the production of electronic equipment is vertically integrated and is used as an input by many other sectors.

![Figure 1 – Index of weighted average propagation length by sector, country average (2008)](image-url)

Note: Simple average over ten countries of the sectorial average propagation length weighted by Ghosh matrix (G-I), non-domestic effects created initially by a shock on manufacturing sectors.

Index 100 = highest sectorial value.

Source: Authors calculations.

---

94
Conclusions
This study has analyzed the role of international supply chains as transmission channels of financial shocks from the monetary circuit into the real economy. As the initial monetary shock reverberates through the production chains and affects final demand, more and more firms face difficulties in completing their production plans or selling their output. Because individual firms are interdependent and rely on each other, either as supplier of intermediate goods or client for their own production, an exogenous financial shock affecting a single firm – such as the termination of a line of credit – reverberates through the production chain. The transmission of the shock through real channels can be tracked, at macroeconomic level, by modeling input-output interactions. In this respect, the article illustrates the methodology by devising and computing IRSIC, an indicator of supply-driven shocks based on forward linkages.

These disruptions that occur in the real economy eventually do feedback into the monetary circuit. The disruption of the production chain and the building-up of undesired stocks impede the expected destruction of money and determine the accumulation of outstanding loans as well as a further downgrading of the exposed firms. Since the downgrading of an indebted individual firm affects the capital adequacy ratio of its banker, both flows and stocks are affected in the monetary circuit and all firms see their access to credit potentially restricted.

This paper shows that if banks are operating at the limit of their institutional capacity, defined by the capital adequacy ratio, and if assets are priced to market, then a resonance effect amplifies the back and forth transmission between real and monetary circuits. The chaotic behavior of the international financial system at the end of 2008 and the dire consequences on the real economy observed in 2009, are examples of such resonance and amplification.

Using an international version of the input-output matrices, this paper illustrates the calculation of an indicator of supply-driven impact (IRSIC) on ten interconnected economies including the U.S. and 9 developed and developing Asian economies. Results indicate that the real transmission effects through the international supply chain linking firms among these economies were heterogeneous across industries and across countries. Based on the existing inter- and intra-industrial linkages, Japan was the largest exporter of potential price shock, while Malaysia and Thailand were the most vulnerable to such shocks, because of the high degree of vertical integration of their manufacturing sectors. Between 2000 and 2008, China registered a notable increase in both inter-country forward linkages and domestic backward linkages, which increased its influence as an exporter of price shocks, while its vulnerability to an imported shock remained relatively stable. When both the intensity of the shock and its average propagation length across countries and sectors are accounted for, computers and electronic equipment are the most noteworthy channels of transmission. Because IRSIC are industry averages based on national account data, they underestimate the intensity of the shock for the export-oriented firms, especially for small developing countries.

The synchronization of domestic business cycles after September 2008 was unprecedented, especially between advanced economies, even when compared to the oil shocks of the 1970s. The potential role of international supply chains as transmission channels during the 2008-2009 financial crises has been thoroughly investigated. A number of governments, due to their concerns that decline in trade finance would deepen the slowdown of world economy, took a number of initiatives through improved insurance schemes or specific credit lines. Focusing on the Japanese financial crisis, Amiti and Weinstein (2009) attribute about one-third of the aggregate drop in exports to the credit crunch. Nevertheless, as these authors recognize, a more general impact of credit crunch affecting trade finance and causing the disruption of specific value chains has so far not been identified, due to measurement and endogeneity issues.17 Available information based on firm-level data in developed countries seems to support the idea that the financial restrictions acted through a reduction on final demand rather than through micro-disruption, at least for firms involved in international trade. Yet, firms located in emerging countries might face stronger risks of propagation of shock [Menichini (2009)]. Be they of macro or micro in nature, the financial implications of the 2008 crisis provide evidence for the need to address problems of macro-prudential procyclicality in order to minimize the risks of boom and bust cycles initiating from the financial sector and reverberating through international supply chains.

17 Levchenko et al. (2010) do not find overwhelming support for the hypothesis that trade credit has been important for the 2008-2009 collapse. Bricongne et al. (2010), on the French case, and Behrens et al. (2010) for Belgium, show that the micro-economic impact of credit constraints on disrupting trade has been rather limited.
References

- Feenstra, R., 1998, “Integration of trade and disintegration of production in the global economy,” Journal of Economic Perspectives, Fall, 31-50
- Godley, W., and M. Lavoie, 2007, Monetary economics: an integrated approach to credit, money, income, production and wealth, Palgrave Macmillan
- Inomata, S., 2008, “A new measurement for international fragmentation of the production process: an international input-output approach,” IDE Discussion Papers No.175, December
- Inomata, S., and Y. Uchida, 2009 “Asia beyond the crisis – visions from international input-output analyses,” IDE-Jetmo, December
- Stadtler, H., and Ch. Kigler, 2008, Supply chain management and advanced planning, Springer-Verlag
- Tanaka, K., 2009, “Trade collapse and international supply chains: evidence from Japan” VoxEU.org, 7 May

Appendix 1 – Supply-driven input-output models

The well-known ‘demand-driven’ model was developed by Wassily Leontief in the 1930s. Two decades later, Ambica Ghosh adapted the I-O model to analyze supply shocks. The Ghosh approach states that each intermediate output is sold to a series of industrial sectors in fixed proportions. When the production of an intermediate product ‘i’ is exogenously altered, the primary effect is felt by those sectors that need ‘i’ as input. This will trigger forward, either direct (to the sectors requiring ‘i’ as input for their production) or secondary effects (sectors depending on intermediate goods that had required ‘i’ as input). As in the Leontief case, the iterative process dies down to reach another equilibrium.

The accumulation of impacts can be measured by the Ghosh inverse (I-B)-1. As in the Leontief case, the matrix B is built using the inter-sectoral transaction matrix, but the allocation coefficients are normalized in rows (destination of output) by the value of production, and not in columns as for technical coefficients (origin of productive factors used in the production).

The Leontief logic for backward linkage is based on standard economics: sectors do respond to changes in demand. The Ghoshian approach is much weaker, and its theoretical aspects are somewhat contentious. Indeed, the theoretical reservations about the Ghosh model led to its relative demise as a macroeconomic modeling tool in the quantity space. Nevertheless, the Ghosh approach is still useful in the price space, and can be used, within certain limits, to model the transmission of shocks through costs of production [Dietzenbacher (1989); Mesnard (2007)]. It is particularly true for short-term analysis, when firms have limited capacity for substituting the disrupted input by shifting to alternative and more expensive suppliers.

The mechanism is as follows: a quantity restriction on any single intermediate good used as input forces the client-firm to shift to other suppliers (foreign or domestic). While this is always possible in the model, it has a cost, as alternative suppliers will supply the needed quantities at a higher price. It should be noted that the dual of the Leontief model could also be used to model price effects, when the shocks originate in primary inputs (i.e., wages).

The final impact on production costs depends on a conjunction of quantitative and qualitative factors. The quantitative factor is proportional to the contribution of the disrupted input in the production function and is captured by the allocation coefficients of the I-O matrix. The qualitative factor is determined by the particular market structure for this product, in particular by the possibility of substitution (Armington elasticity). Estimating these elasticities has generated an abundant literature, especially in areas linked with international trade literature and ‘computable general equilibrium’ models. A review of existing results show that the elasticities
The Capco Institute Journal of Financial Transformation
International Supply Chains as Real Transmission Channels of Financial Shocks

(estimated using multilateral trade data) for the intermediate inputs industries tend to be higher than those for the final consumption goods industries [Saito (2004)].

Appendix 2 – Average propagation length

Average propagation length (APL) measures, in a forward-looking model proper to Ghosh matrices, the average number of steps it takes a cost-push in one sector to affect the output value of others. It should be noted than APLs are symmetric in the demand and supply domain, and both Leontief and Ghosh approaches yield the same result.

\[ \text{APL} = G_i(G-I)\left(\frac{1}{G'}\right) \]

With \( g'_{ij} = g_{ij} \) if \( i \neq j \) and \( g'_{ii} = g_{ii} - 1 \) \hfill (2)

With \( G = (I-B)^{-1} \)

\( (I-B)^{-1} \): The Ghosh inverse matrix,

\( \odot \) denotes the Hadamard product; in this case, \( H \odot (1/G') = (h_{ij}/g'_{ij}) \)

APLs are usually inversely correlated to the intensity of the linkage. As mentioned by Dietzenbacher and Romero (2007), many linkages with a large APL are often also almost irrelevant in terms of size. In order to limit this bias, we rescaled each of APLs coefficients using the strength of its respective forwards linkage:

\[ \text{APL}^\# = \text{APL} \odot (G-I) \] \hfill (3)
Part 2

Chinese Exchange Rates and Reserves from a Basic Monetary Approach Perspective

Asset Allocation: Mass Production or Mass Customization?

Practical Attribution Analysis in Asset Liability Management of a Bank

Hedge Funds Performance Ratios Adjusted to Market Liquidity Risk

Regulating Credit Ratings Agencies: Where to Now?

Insurer Anti-Fraud Programs: Contracts and Detection versus Norms and Prevention

Revisiting the Labor Hoarding Employment Demand Model: An Economic Order Quantity Approach


Indexation as Primary Target for Pension Funds: Implications for Portfolio Management
Chinese Exchange Rates and Reserves from a Basic Monetary Approach Perspective

Bluford H. Putnam — President, Bayesian Edge Technology & Solutions, Ltd.
Stephen Jay Silver — Professor of Economics, the Citadel
D. Sykes Wilford — Frank W. Hipp Distinguished Professor of Business, the Citadel

Abstract
As China exercises greater influence over global economics commensurate with the size of its economy, political discussions of the correct exchange rate for China will demand center stage. Questions about whether China should float the yuan, what the ‘competitive’ exchange rate might be, and so forth, will take a great deal of political time and energy of policymakers. Most discussions by U.S. and European political officials and central bankers heretofore have centered on unfair trade advantages accruing to China due to an ‘undervalued’ yuan, as if this is an absolute fact. Some focus on imbalances in bilateral trade accounts. Others focus on imbalances in current accounts, with accusations back and forth. Our belief is that most of these discussions miss the point, seeing only the trees and misunderstanding the size and complexity of the forest. That is, the exchange rate regime, exchange rate level, and the accompanying accumulation of foreign reserves by the Chinese government are a natural result of a set of carefully chosen domestic monetary policy decisions. Using a basic version of the monetary approach to the balance of payments model, one can develop a framework in which the Chinese monetary policy decisions can be better understood and interpreted. This type of analysis, we believe, provides a superior context for discussing the value of the yuan compared to the misguided focus on competitiveness and fair value. It also provides insight as to where problems may occur as well as what factors driving policy may be important in defining adjustments in either policy or economic outcomes involving exchange rates, inflation, wage demands, and equity prices.

1 All opinions are those of the authors and do not necessarily reflect those of their respective institutional affiliations.
As China’s role in the global economy expands, so does the importance of its exchange rate policy and its impressive accumulation of foreign reserves. As of the end of December 2010, the Chinese authorities reported holding U.S.$2.8 trillion in foreign reserves. Understanding the domestic economic forces and the subsequent monetary policy choices made by China can shed light on the global debate concerning the appropriate exchange rate regime for China and its appetite for accumulating foreign reserves.

As with any analytical task, choosing the relevant perspective and theory can make a huge difference in how one perceives the outcomes and understands the driving forces behind the issues. In the case of China, the monetary approach to the balance of payments offers an excellent theoretical framework within which to study empirical findings and to draw conclusions about the forces driving Chinese exchange rate and foreign reserve policy. The choice of our theoretic guide is not new. David Hume wrote extensively and with clear insights in the 18th century about the is-...
the flexible exchange rate model, and the currency substitution literature. For our research, we concentrate on the more basic aspects of the approach.\textsuperscript{2}

The thrust of the literature is that the balance of payments may be examined as a monetary phenomenon. This does not imply trade versus non-traded goods issues are unimportant; it simply notes that in the context of a global economy with a reserve currency (or a gold standard) and fixed exchange rates, the non-reserve currency countries will find that money supply adjusts to demand. And, the source of the liquidity (high powered money) that provides the input to satisfy that demand may come not from the central bank or the local government entity, but rather from external sources. The implications are accepted today by most economists, but often ignored by policymakers and politicians, that monetary authorities cannot control their own money supply and at the same time maintain a fixed or even a semi-fixed exchange rate. The stock of money will reflect the demand for it in open economies, or even semi-open economies, that are not reserve currency countries. One may find causality running from supply of money to GNP in a reserve currency country but not typically in those that are not creators of international reserves.\textsuperscript{4} And causality can become confusing to track in a relatively underdeveloped financial system in which the government exercises influence over lending practices of a banking system that bears no resemblance to how U.S. and European banking systems function. How this affects the standard set of policy implications must be examined to gain a better understanding for what is possible and not possible within the Chinese system.

With China’s semi-fixed exchange rate system in place, then, the theory suggests that the authorities cannot control its domestic money supply. The demand for money (if policy is deemed not to supply that demand with domestically created money) will be met with resources from abroad. (One could argue that is the flipside of the current account balance surplus but that is for a different set of discussions). By all accounts, however, we find that the domestic monetary authorities do operate in a manner to control money growth, “to support economic stimulus” or “slowdown an overheating economy” (to use the words of the press). How this is manifested in reserve buildup and domestic policy in the context of the theory is to be examined. Before one can make policy inferences for China, an empirical analysis of the MBOP as applied to Chinese data must be undertaken. To do so, we utilize the model laid out by Putnam and Wilford (1978). We choose this methodology among the possible ones since it allows us to consider the fact that domestic data for interest rates, prices, as well as other series may be insufficient or poor measures of what is actually transpiring in an economy. The approach, no matter the estimation procedure, is in keeping with the simple specification of the model for the basic MBOP. It follows that of Johnson (1973).

The basic model:

$$M^d = kPYe^u/d$$
$$M^s = aH$$
$$M^d = M^d$$

where,

- $M^d$ = the demand for money,
- $M^s$ = the money supply,
- $k$ = a constant,
- $P$ = the price level,
- $i$ = the interest rate,
- $a$ = the money multiplier,
- $H$ = the stock of high-powered money.

In the basic framework, prices are assumed determined by world markets and are, thus, a proxy for world prices. Domestic Chinese price data no doubt will differ. Similar assumptions are typically made for interest rates. We investigate the possible choices of measures during the analysis since there are many concerns over the data.\textsuperscript{3,5}

Next the balance sheet of the central bank can be described as

$$M^a = a(R + D),$$
where $R$ = stock of international reserves held by the bank, and $D$ = domestic credit.\textsuperscript{6}

Combining the basic equations noted above and moving to percentage change terms (d log terms) we can write the reserve flow equation as

$$\left(\frac{R}{H}\right)_{gR} = gY + gP – di – ga – \left(\frac{D}{H}\right)_{gD},$$
where $gX$ refers to the rate of growth in $X$.

\textsuperscript{2} The literature has expanded greatly over the years. Whereas the basic concepts have remained the same based upon the early work as outlined by Connolly, many ‘tweaks’ both theoretical and empirical have been contributed. We are not ignoring these adjustments, which have value. It is the intention of this paper to focus on very basic arguments so that the overall issues in managing a semi-fixed exchange rate are highlighted and analyzed in the Chinese context. Thus for the interested reader we have assembled a large bibliography for further reading.

\textsuperscript{3} See Sims (1972), Goodhart et al. (1976) and Putnam and Wilford (1978) for a discussion of the causality of money and the implications for an open economy versus a reserve currency country. With flexible exchange rates these points have been taken for granted for the G-7 economies, but intervention and sterilization even in these economies suggests that the losses are often ignored by the authorities.

\textsuperscript{4} Indeed one of the items that must be considered in this analysis is that the data are fraught with problems, adjustments are slow, and any analysis will thus be affected. For this reason we employ a semi-fixed exchange rate as the model.

\textsuperscript{5} Further, the exchange rate changes over the period and one of the assumptions of this model is that it does not. Movements are not continuous and we have introduced the exchange rate variable to compensate for its effect, both on the economics of demand for reserves and on the accounting effects that may occur.

\textsuperscript{6} Domestic credit is defined as the bank’s holdings of domestic assets minus domestic liabilities other than high-powered money. One could think of this as the monetization of domestic government debt. It can also be the monetization of debt; that is, by liquidating the monetary authorities’ holding of government debt or replacing it with foreign reserves on the balance sheet, the authorities may decrease that debt, potentially doing so as a tool to sterilize the purchases of foreign government debt.
This is the standard fixed exchange rate model for determining reserve flows or the overall balance of payments. The question now remains as to whether Chinese reserve flows are consistent with what one would expect from this model’s predictions for a non-reserve currency country.

Rewriting the basic equation in a simple OLS estimation form with an intercept term yields:

\[(R/H)gR = \beta_0 + \beta_1 gY + \beta_2 gP - \beta_3 di - \beta_4 ga - \beta_5 (D/H)gD + \mu\]

For our purposes, we will make a few more simplifying assumptions to deal with data issues as we move to the empirical analysis.

The MBOP prefers to treat separately permanent real income and prices, but in our statistical analysis we sometime combine the two back into a long-term nominal income growth concept. We do so on this occasion because of data collection issues in China. Real GDP data collection begins with estimates of nominal GDP which is then deflated by a price series. Given the large price discrepancies across regions in China, we feel that the process of estimating an appropriate price deflator is significantly more difficult than it would be in the case of U.S. and Europe. (In one investigation we do separate the variables and find that the results, while acceptable, may be corrupted by the general weakness in the price series; the results are discussed as well.) We explore the use of both reported income data and reported industrial production and inflation data to create a nominal income series. Using nominal income, and smoothing it over time, adheres to the essence of the monetary approach and modestly minimizes some of the Chinese data issues that all researchers confront.

The high-powered money multiplier is another issue. In a developing economy in which the evolution of the banking system is rapid and dynamic, the measurement of money and the calculation of the multiplier can be viewed as a potential distraction. So, in some cases, we focus only on the quantity of foreign reserves and on government debt held by the central bank as measures of assets active in monetary policy operations, and we let the empirical study adjust estimated coefficients to accommodate our omitting the high-powered money multiplier from the estimation equation. Essentially, we are viewing the high-powered money multiplier as endogenous to the banking system and not necessarily as appropriate for inclusion as an independent variable. (Again we separately test its movements in an alternative set of empirical work for consistency, although ex-ante we do not believe that the results will shed a different light on monetary issues and why the reserve build up is as it is.)

Finally, in certain estimation equations we add the percentage change in the exchange rate. As Connolly (1986) noted, in a managed exchange rate regime, the central bank makes decisions to allow the currency to appreciate if it wants to slow the accumulation of foreign reserves and not change any other domestic policies.

Thus, the simplified estimation equation used in some of the empirical analysis is of the form:

\[(R/H)gR = \beta_0 + \beta_1 gY + \beta_2 gP - \beta_3 di - \beta_4 ga - \beta_5 (D/H)gD - \beta_6 gX + \mu,\]

where gX is the year over year percentage change in the yuan per U.S. dollar.

**Expectations from the literature**

From the early estimations of this type of MBOP model and from observations of the growth of China through the years, one would expect that the betas for income and prices would be near one and significant. In the case where we estimate nominal income (PY), again we would expect the estimated beta coefficient to be near one. Moreover, in a country with a very high growth rate over a long period of time, we would expect this factor to be one of the most important in determining foreign reserve accumulation. Figure 1 illustrates the sustained high rates of nominal income growth (using smoothed industrial production data from China as a proxy for real income).

The interpretation of the interest variable is complex. In general, the use of a global interest rate in a fixed exchange rate world means that changes in the interest rate are to be interpreted as changes in inflation expectations. From an empirical perspective, though, when using the U.S. Federal Funds Rate as the global interest rate, we are mixing market expectations with U.S. policy decisions. We include the interest rate in the empirical analysis, but we do not view its estimated beta as critical to the focus of this study.

In the literature, the significance of the interest rate variable varies from
study to study. Different choices for measuring it affected results as well. Since the studies covered a great deal of different types of economies, from developed to extremely less developed, from developing to stagnant, and across various global economic conditions, one would expect the interest rate’s impact to vary. And, since most countries’ data series report set rates, which from time to time did not reflect market conditions (for policy reasons), the arbitrage that normally makes the variable relevant (in an open economy) was not exhibited in the published rate series (the data could be misleading).

On the supply side, when estimated, the high-powered money multiplier betas were typically around .5 although the range of estimates varied dramatically, depending on how the monetary authorities managed policy. The significance of the coefficients varied, based upon the tests used for monetary policy, the nature of the financial systems, and the freedoms with which the banking system operated. As noted, this factor may not be independent for China, given the nature of how the Chinese banking system has evolved.

Domestic credit, effectively that portion of central bank monetary assets that is not represented by foreign reserves would be expected to be important. The Chinese, however, have not continuously used purchases of domestic government debt by the central bank as an active policy tool. Figure 2 illustrates the evolution of this factor. We use the measure of domestic government debt held by central bank as provided by the IMF IFS data.7

Purchases by the central bank of Chinese government debt were used as a policy tool until 1995, when they were phased out. Following the 9-11 terrorist attack in 2001 on the U.S., markets were disrupted and the central bank bought Chinese government debt for a short period of time before it was halted again. Finally, in the financial panic of 2008, the central bank again used domestic government debt purchases as a temporary policy measure to insulate China from the banking system mismanagement occurring in the U.S. and Europe. Given that the domestic credit variable was used mostly in crisis times after 1995, we would not put too much emphasis on the estimated coefficient.

In other studies using MBOP, typically the estimated betas for domestic credit are significant and close to one in most all cases. Again varying circumstances yielded some differences.8

With flexible exchange rates the norm for most countries, the model could be reworked to explain foreign exchange movements instead of reserve flows. As such, the models would reflect inflationary expectations, policy forecasts, various indicators for the independent variables, and so forth. In our estimations, as noted earlier, we include an exchange rate variable and would expect this variable to offset foreign reserve accumulation in the context of an MBOP model; some of the effect will be real and some of it simply reflecting accounting and may be misleading, thus making the beta coefficient difficult to interpret.

China, with its capital restrictions and semi-fixed system is neither fully fixed nor floating, at least against the reserve currency, the dollar. It was, however, operating in a flexible exchange rate world and thus any MBOP estimations would have to reflect this fact and slippage in estimation would be expected. Similarly, discussions of price arbitrage and thus the meaningfulness of reported inflation rates with respect to effects on reserve flows would also be suspect relative to the original specie-flow theories espoused by Hume. Having said all this, the general model should still apply. With these caveats the MBOP sets the general guidelines for considering the sizable buildup in reserves by China.

The literature on Chinese exchange rate and reserves management policy

For the purposes of this paper articles by Lardy (2005) and Mussa (2007) are apropos. In particular, Mussa discusses at length the reserve adjustment mechanism in the context of the MBOP as we are. Referring to the basis of the adjustment process he states: “This has effectively frustrated the other mechanism for adjustment of China’s real exchange rate and its balance of payments, namely the classic price-specie-flow mechanism

7 This measure can vary from other estimates of domestic credit in minor ways, although in the case of China the absolute measures will look different. We will examine an alternative measure of domestic credit as well which deviates from that reported by the IFS database to see if this alternative measure impacts our conclusions.

8 In fairness at this point, it is important to note that several authors have argued that the results are not as robust as once thought for several reasons, suggesting that the estimates overstated the predictability of monetary policy or rather the lack of effect of domestic monetary policy for many non-reserve currency countries. Further, the movement to semi-flexible exchange rates could have made the data less meaningful.
described by David Hume. The exceptional nature, scale and duration of these policies, together with the extraordinary upsurge in China’s balance of payments surplus over the past five years leaves no reasonable doubt that China is in violation of its specific obligation under Article IV Section 1(iii) to “…avoid manipulating exchange rates or the international monetary system in order to prevent effective balance of payments adjustment or to gain unfair advantage over other members.”

In concluding these findings, he focuses on the use of sterilization of reserves to manipulate policy and support the buildup in reserves. “Applied to China today, the point is that the sterilization of the monetary effects of very large foreign exchange inflows chokes off the normal mechanism of adjustment of the real exchange rate through domestic price inflation. How large might this effect be if it was not being frustrated?” “As a useful thought experiment, is to consider what would probably have happened in China if the monetary effect of the foreign exchange inflow in 2006 had not been sterilized, but had been allowed to have its full impact on China’s monetary base, assuming for simplicity that the net domestic assets of the monetary authority had been held constant at their 2002 level (rather than being expanded to account for at least half of the growth of the base). By end 2006, the hypothetical result would have been a monetary base of 10, 677 billion yuan rather than the actual monetary base of 7,776 billion yuan, an increase of 37 percent assuming, as is reasonable, a roughly proportional reaction of China’s price level to the massive increase in the money supply implied by this large hypothetical increase in base money, China’s real exchange rate would have appreciated by 37 percent relative to its actual level. This would be equivalent to a nominal appreciation of the yuan, relative to the U.S. dollar to 5.7 yuan to the dollar, rather than the actual rate at end 2006 of 7.81 yuan to the dollar. The real appreciation of the yuan brought about, in the absence of sterilization, through David Hume’s price-specie-flow mechanism would have eliminated all, or virtually all, of what most estimates suggest is the present undervaluation of the yuan.”

Musser is focused on overvaluation and undervaluation and utilizes the MBOP analytics to discuss the degree of undervaluation of the yuan. Dunaway and Li (2005) find similar results utilizing alternative methodologies. Lardy (2005) worried about the undervaluation and implications for overinvestment in fixed assets. And today the popular press focuses on this issue of valuation.9

We agree that the MBOP is the correct methodology for considering the overall balance of payments issues for China, to better understand its monetary policy and objectives, and to understand the reserve build up. Our objective is not to try to calculate over- or undervaluation of the currency and the degree of deviation from equilibrium. Ours is simply to ask if the MBOP can explain the build up in reserves that we all know about and some commentators are concerned about. Ancillary to that analysis are others. Do the data that we have for China make sense? Many domestic rates are managed, just as are loans from time to time, in a manner that is not the case for other economies, even those where the MBOP was applied in the Bretton-Woods period. Can the MBOP shed light on these issues? Domestic interest rates, as they would be useful in helping define money demand, may or may not be meaningful signaling devices for money demand. The income data for China have been challenged.10 And, where the model may be lacking in explanatory power can it, in and of itself, enlighten us about the economy and how policy was conducted. A better understanding of the MBOP model and its usefulness (and lack thereof) in explaining reserve movements in China may shed light on some of these discussions.

**Empirical results**

We will approach the empirical section in various ways. We first examine a simplified approach to the modeling of the balance of payments and utilize monthly data, which has been smoothed. With this large dataset, we can dig deeper into the sensitivities of the explanatory variables. Secondly, we analyze the MBOP implications using a different dataset, which is quarterly. It utilizes a different measure for domestic credit (based upon an alternative calculation methodology of the data) as well as the real GNP series reported by the authorities. This alternative dataset allows for a check on the consistency of the results from our initial (larger data set) analysis.

**Primary empirical results: a simplified approach**

Our estimation equation is as follows:

\[ (R/H)_g = \beta_0 + \beta_1g(Y^*P) - \beta_2d(i) - \beta_3gX + \mu \]

We used data from the mid-1980s, but with data issues, smoothing, year-over-year percent change calculations, our first feasible estimation point was January 1990. Our end point was December 2009. We are using monthly data in the initial work presented in this section, even though some data are not reportedly monthly. The foreign reserve data and the domestic government debt held by the central bank are taken from the IMF IFS database, as is. In this part of the study we focus on these two components of the central bank monetary policy, and high-powered money, H, is the sum of the two.

The income variable is a nominal income proxy using industrial production and price data. Since the industrial production and price data are reported by the Chinese government in a year over year percent change

---

9 Though few and far between, some authors have argued that the yuan may not be undervalued against the major currencies. The Economist (2005) countered received wisdom in a thoughtful piece of analysis.

10 See Razak (2001) for a discussion of the fallacy of the income data provided by the authorities. We do not agree with much of this analysis and believe there is value in the income data as a prime explanatory variable for the demand for money.
form, we smooth the data with 12-month moving averages, and interpolate, where needed, to produce a monthly series.

The exchange rate data are from the IMF IFS database as is the U.S. Federal funds rate.

We estimate the whole period, as well as the first eight years and the last eight years. These two sub-periods are meant to shed light on policy shifts before the Asian contagion of 1997, and policy in the 2000s as part of the era of growth and modernization.

The results conform reasonably well to our expectations.

Most importantly, the income variable has an estimated beta coefficient close to one and is significantly different from zero in all periods. This strongly supports our contention that China’s economic growth is a primary driver of the accumulation of foreign reserves. When the whole balance of payments is considered, one need not look to arguments about an undervalued currency to explain reserve accumulation.

The domestic credit variable has the expected negative sign and is highly significantly different from zero only in the 1990s period running up to the Asian contagion. As already noted, after the mid-1990s, purchases by the central bank of domestic government debt was only used as a policy variable in crisis situations such as post 9-11 in 2001 and the global financial panic of 2008. In a short-lived crisis, all bets are off on beta coefficient estimates, so we do not place much importance on the full period estimation (expected negative sign, not quite significant) or the 2000s (positive sign, not significant).

The exchange rate is measured in yuan per U.S. dollar, so a smaller numerical value indicated yuan appreciation. Yuan appreciation can be viewed in MBOP terms as a substitute for foreign reserve accumulation, so a negative sign is expected. This occurs for the whole period and the 1990s, but not for the 2000s. Given the U.S.-China politics over the adjustments in the Chinese exchange rate in the last decade, as well as the relatively small adjustments, the role of this variable may have been more correlated with, and endogenous to, foreign reserve accumulation than an independent factor.

As can be seen in Figure 3, the exchange rate was quite active in the 1990s before 1996, then fixed from 1996 through 2006, when very small

| Panel A – Whole period: 1990-2009 monthly MBOP estimation results |
|---------------------------------|-----------------|-----------------|-----------------|
| Dependent variable: Foreign reserve growth (R/HgR) |
| Start period: January-1990 |
| End period: December-2009 |
| R-square statistic: 22.37% |
| Independent variable | Estimated beta coefficient | Standard error | T-statistic |
| Intercept term | 0.14 | 0.03 | 4.77 |
| Nominal income (gPY) | 0.94 | 0.16 | 6.02 |
| Domestic credit growth (D/HgD) | -0.26 | 0.19 | -1.38 |
| Exchange rate g(FX) | -0.55 | 0.15 | -3.70 |
| U.S. rate changes g(I) | 3.98 | 0.84 | 4.72 |

| Panel B – Before Asian contagion: 1990-1997 |
|---------------------------------|-----------------|-----------------|-----------------|
| Dependent variable: Foreign reserve growth (R/HgR) |
| Start period: January-1990 |
| End period: December-1997 |
| R-square statistic: 58.19% |
| Independent variable | Estimated beta coefficient | Standard error | T-statistic |
| Intercept term | 0.36 | 0.06 | 6.13 |
| Nominal income (gPY) | 0.71 | 0.23 | 3.08 |
| Domestic credit growth (D/HgD) | -4.85 | 0.59 | -8.17 |
| Exchange rate g(FX) | -1.46 | 0.24 | -6.16 |
| U.S. rate changes g(I) | -4.20 | 2.15 | -1.95 |

| Panel C – Recent growth period: 2002-2009 |
|---------------------------------|-----------------|-----------------|-----------------|
| Dependent variable: Foreign reserve growth (R/HgR) |
| Start period: January-2002 |
| End period: December-2009 |
| R-square statistic: 77.53% |
| Independent variable | Estimated beta coefficient | Standard error | T-statistic |
| Intercept term | 0.12 | 0.04 | 3.15 |
| Nominal income (gPY) | 1.14 | 0.23 | 4.94 |
| Domestic credit growth (D/HgD) | 0.12 | 0.10 | 1.19 |
| Exchange rate g(FX) | 0.59 | 0.22 | 2.64 |
| U.S. rate changes g(I) | 2.52 | 0.47 | 5.32 |

Table 1 – Empirical results
upward adjustments were allowed. When the exchange rate was active, the empirical results support the MBOP.

We also used recursive techniques with an 8-year window rolling through time, to observe the stability of the estimated coefficients in a much different way than can be gleaned from whole period standard error and T-statistics.

What follows are the rolling estimated beta coefficient charts. The dates on the charts refer the end date for each 8-year period. So, December 1998, refers to January 1990 - December 1998 period.

The estimated beta coefficient is consistently positive and shows a long-term trend increasing over time, only to come back to earth when the periods start to include the 2008 global financial panic. The growth of the beta coefficient over time, we think, highlights the increasing importance of economic growth as the driver for Chinese foreign reserve accumulation as China’s economy went from an emerging market to a dynamic developing country with an economy second only to the U.S. in size in the last decade.\textsuperscript{11}

The estimated beta coefficient for the domestic credit variable is seen in the recursive chart as behaving exactly as expected when it was an active policy variable, then losing its explanatory power (beta close to zero) as it came to be used only in emergency situations like 9-11 or the 2008 global financial crisis.

Relationship between domestic credit and reserves

Considering the movement through time in the estimated beta for the domestic credit defined narrowly as central bank purchases of domestic government debt, we expanded our investigations to include a broader view of domestic credit. In particular, the balance sheet of the central bank contains a number of items of a domestic nature, sometimes referred to in MBOP studies as ‘other domestic assets’ and ‘other domestic liabilities’ that may be considered as being part of the non-monetary policy functions of the central bank. These ‘other assets and liabilities’ were excluded in the results presented so far, but it is appropriate to examine their role.

Especially since about 2003, China’s foreign reserves began to build-up in earnest. We are arguing that most of this foreign reserve accumulation was the natural consequence of rapid economic growth in the context of a managed exchange rate regime. But depending on one’s perspective on the proper definition of domestic credit, one can ask some additional questions. Indeed, with a broader definition of domestic credit, during the period of the huge influx of foreign reserves, domestic credit was trending downward. Depending on the methodology used, domestic credit growth may have become negative, as ‘other assets’ held by the central bank declined relative to ‘other liabilities,’ which were accumulating as the central bank’s issuance of bonds grew dramatically in the latter period from 2002. Since the bank’s role is different than that of, say, the Bank of England, such issuance is more consistent with its ‘treasury’ type function and may explain why the data could be distorted. In any case, the accumulation in foreign reserves is impressive and the change in the way the balance sheet looked in 2000 compared to 2010 is fairly obvious as well.

\textsuperscript{11} Since income seems to be the driving force for reserve build-up, given the semi-fixed exchange rate, the issue noted about its quality is interesting in and of itself. The typical alternative variable, industrial production may be an overestimation of output for the economy as a whole, given the huge transition from a rural economy to an industrial, city dwelling one. Others argue for electricity output as a measure of economic growth (see Xu (2010) for a graph of these data and Maddison (1998) for discussions on this topic). Here we utilize this combined measure based on industrial production and with the smoothing we find that it is consistent with expectations.
Foreign reserve accumulation, as a result of rapid expansion in the economy and resulting money demand not being satisfied through domestic sources, would normally drive the money supply and in time affect the price of goods and eventually wages. To a point, sterilization was used to offset the effect of foreign reserves on the money stock, creating a drop in domestic credit leaving demand unsatisfied, thereby setting in motion another phase of reserve buildup.\(^{12}\) In the long-run, though, sterilization simply does not work, so long as the managed exchange rate policy is maintained. In the medium term, the cycle can be supported, which leads to more accumulation of reserves.

Using quarterly data and the broader definition of domestic credit, we tried to demonstrate a causal link running from foreign reserves to domestic credit, but were unable to find such a relationship. After taking first differences of both series and running Granger tests of causality we rejected this hypothesis; instead only lagged values of changes in domestic credit were significant in explaining changes in domestic credit, especially the four- and eight-period quarterly lagged values.

Seasonality is very important in Chinese data and can often disturb statistical relationships. In Figure 7 we see that there is indeed a very strong seasonal pattern to changes in broadly defined domestic credit; major increases in domestic credit during the fourth quarter of each year are followed by significant declines the following quarter.

This pattern is due to the lunar New Year, which occurs every year in late January or February, during which Chinese take a one-month vacation. Plants close, schools are on vacation, and an estimated 300 million people travel by bus, train, and airplane returning to their cities, towns, and villages to spend time together with their families. Virtually all industrial activity ceases. Thus, the fourth quarter is marked by heightened activity in preparation of the mass exodus home; presents are bought and plants work overtime to produce inventories to ship before the recess.

This expansive monetary demand during the fourth quarter quickly collapses with the reduction of economic activity.

Because our central bank balance sheet data are not seasonally adjusted, it is impossible to find a correlation between the highly seasonal credit series and the relatively non-seasonal foreign reserves data. But in order to visualize the true relationship, we smoothed both series, using

---

\(^{12}\) Domestic credit calculations from the IMF database can take two forms. In this instance we looked at OA – OL from reported data from the central bank’s balance sheet. This measure is slightly different from that calculated in our monthly estimates above, which utilize the IMF’s approach. Both methodologies are consistent, however, in the picture they paint for policy. The qualitative differences in implications are unimportant although the series are quantitatively different in minor deviations. For the estimations using this alternative methodology, we used quarterly data. Our goal was to check on causality using a more raw dataset that may be interpreted differently if we achieved results inconsistent with the domestic credit calculations above.
centered moving averages. There has been a very clear inverse relationship, especially since 2003, between changes in the two variables, as can be seen in Figure 8, in which we plot the two smoothed series. The correlation between smoothed changes in reserves and domestic credit between 2002 Q4 and 2010 Q1 is -0.66311, which is significantly different from 0 at the 1% level.

These results are consistent with those reported in the previous section using monthly smoothed data. The estimated beta movements over time for domestic credit support a notion that the domestic credit variable is, from time to time, a driver of policy, money supply, and thus foreign reserves while at other times it is passive. It also supports the notion that China changed policy during the global financial crisis in an effort to insulate the impact of the West’s problems on domestic monetary policy.

Using alternative GNP and price data (quarterly)

Further we considered the implications for a MBOP calculation using a different series for GNP data reported quarterly. In this case we broke out GNP, reported prices in China, the exchange rate, domestic interest rates, as well as the money multiplier and domestic credit. The data were quarterly and the period was 1994 through 2010, reflecting the time period when this reported data for GNP was begun.

Summarizing the results for the whole period provide some interesting insights. Significance and theoretically expected signs are found for the monetary variables and real income. It appears that real income is the driver, with the price series not providing any information. Overall the adjusted R2 is acceptable at about .7. The overall F-statistic suggests reasonable explanatory power. The real income variable is the primary driver of the results with a much higher than expected beta. The poor state of the domestic price data could be biasing the size of the real income coefficient upward. Also, as our monthly recursive work suggested, when the Chinese economy shifted to high gear, that structural development issues resulted in the upward trend for the estimated beta for real income. For our purposes, it suffices to note that the overall results of this alternative set of results are consistent with our earlier ones using monthly data.

Some observations from the empirical work

In the quarterly data investigations, the money supply variables do not appear to be as reliable indicators of policy as one would like, but they do show importance. The fact that the money multiplier shows up as important probably tells us very little, however. This variable was considered useful as a measure of monetary policy in several of the studies during the Bretton Woods period. It was significant for countries that tended to use mechanisms, such as reserve requirement changes, for policy in better developed financial markets, such as the U.K. In less developed economies, the results were mixed.

In China, given the nature of the banking system, which is under greater state control over lending, many mechanisms for delivery of policy outside the standard protocols of a semi-developed non-government controlled banking system are available to the authorities. As such, the policy conduits are naturally more difficult to delineate. Thus, the money multiplier variable is less likely to reflect any consistent conduit for policy for China. The fact that it shows up in this period with the correct sign and is significant in our alternative quarterly estimates may suggest that when the authorities do use a reserve requirement tool it does have a direct effect, even if it is not necessarily a consistently powerful tool.

Second, and much more interesting, are the mixed interpretations of the broadly defined domestic credit variable’s size, stability, and importance in all of our research. When the domestic credit variable conveys little effect on the money supply and central bank policy, then one may conclude that the data are incorrect (a possibility, but in this case not likely the problem), the country is somehow a reserve currency country (again not the case), or monetary policy targeting makes causality assumptions in the MBOP inconsistent. That is, statistical causality tests from money supply to income, etc., à la the original Sims (1972) tests, is clearly correct for the reserve currency country, the U.S., but not for those who hold dollars as foreign reserves. As pointed out by Putnam and Wilford (1978), the causality runs from demand to supply for the non-reserve currency countries. Measurement of this causality effect may be hampered, however, if policy does not consistently utilize domestic credit as a policy tool but instead allows it simply to adjust to the flow of reserves. In this case, reserves are demand driven by, say, growth and for sterilization reasons domestic credit adjusts to the buildup in reserves. Considering the results from our different approaches to applying the MBOP, it appears that the domestic credit variable’s usage changes depending on the period and thus its impact differs from period to period.

In a sense, if the target for policy is controlled growth in economic output commensurate with population dynamics, then monetary policy may be implemented in two ways: (1) through changing domestic credit or (2) by simply absorbing reserves (or allowing outflow of reserves) with domestic credit acting as a residual to the demand for money generated by economic growth. In the first case, one would expect causality to run from domestic credit to foreign reserves; that is domestic credit policy would move to satisfy or constrain domestic money demand and foreign reserves would adjust accordingly. In this case, domestic credit measures

---

13 Goodhart (1976) noted this observation for the U.K., as a counter to Sims. Putnam and Wilford (1978) explained why this should have been the case for the U.K. and for all non-reserve currency countries under a fixed exchange rate. Within the MBOP literature, the phenomenon was discussed as a critical element by other authors such as Wilford (1986), Gupta (1984), and Blejer (1979); some studies have shown the direction of causality is not as clear as one would expect from reserves to domestic credit. Usually the issues arose due to structures of the domestic economy and the nature of monetary and capital controls.
reflect the policy of the authorities. In the second case, other mechanisms for controlling policy would occur such as loan targeting (state control may allow this), moral suasion (albeit with a stronger hand), or other policy dictates (even wage and price controls). If these are the instruments of policy, foreign reserves would then adjust, followed by domestic credit adjustments (as the bank absorbs the reserves). Simultaneously attempting to sterilize reserve impact on the money supply would show up on the balance sheet as a reduction in domestically created high powered money. The flow of causality would run, however, from economic growth to money demand to foreign reserve flows and, subsequently, to the bank reaction and contraction of domestic credit during it sterilization attempts.

For China there appear to be two distinctly different periods of policy reflected in the domestic credit variable. During the pre-Asian crisis of the end of 1997, a more classic (in the sense of Bretton Woods) use of monetary policy to support growth appears to have been taken. Domestic credit creation was clearly strong and helped fuel growth. Monetary policy was proactive in supporting growth through stimulating high powered money and its effect on the money supply to support the growth efforts of the government.

The Asian contagion crisis in the fall of 1997 dramatically changed the demand for foreign reserves by almost all Asian central banks. It became accepted that a country needed a much larger stockpile of foreign reserves to institute an effective and credible insulation from a policy mistake or economic miscalculation of a neighboring country. Policies in virtually every Asian country adjusted to the reality of the Asian crisis and the dependence on domestically created assets in most Asian central bank balance sheets was lessened dramatically to accumulate foreign reserves [See Putnam and Zecher (2008) for a discussion of the policy adjustment of the authorities for the Asian Tigers following the crisis of 1997-1998].

In the case of China, even though they did not face the same issues that the Asian Tigers did in the late 1990s, central bank policy appears to have adjusted to strong economic growth driven by a more liberal trading and investment policy, which meant for the monetary authorities a greater focus on these mechanisms for fueling the economy and less on proactive domestic monetary policy. Domestic credit tended now to simply adjust (through partial sterilization mechanisms although they may or may not be adjustments made through traditional mechanisms) to the inflows of foreign reserves.

Simultaneously in the post-2001 period, the U.S., as the reserve currency country, was producing reserves at a very rapid rate by any measure. Taken together, China became the recipient of massive investment stimulated by U.S. policy while being supported by its fixed exchange rate focus and its desire to keep the expected rate of return on investment high (on the margin) as it kept labor prices from adjusting (either through high domestic inflation or a stronger currency). In this model of economic growth then, given the rise in global reserves in general, one could surmise that causality would run from reserves to domestic credit.

Why this change occurred can be associated with the Asian contagion as suggested above, but there may be another compelling reason. That is, as China began to solve its underutilization of manpower by facilitating the movement of people from the countryside to the city, it fueled economic growth. What many analysts fail to recognize, although Chinese policymakers appear well aware of the issue, is that China faces another labor issue: China’s population is aging quickly and the best way to mitigate its affect on society is to build up global IOUs. Doing so during a period of exceptional growth deals with two issues simultaneously – the need to make existing labor relatively more productive more quickly to enhance employment and provide the cash flow and income which future demographic difficulties will demand.14 The labor migration from rural to urban centers has therefore masked the challenge of the aging population in China, but toward the end the 2010-2020 decade, the aging challenge will start to dominate the rural-to-urban migration in terms of its impact (i.e., drag) on economic growth. Chinese authorities should not necessarily be criticized for seeking to augment savings (i.e., foreign reserves) during the decades of rapid economic growth.

**Policy and risk issues raised by the analysis**

Focusing policy on trade issues misses the point of what is actually happening in China during the last decade and the one to come. Aging of the population must be considered as well as the implications of aging for economic policy; and, in particular the effect on foreign reserves accumulation. Growth could have been supported by domestic provision of the financing (and not utilizing foreign reserve growth) to satisfy the implicit money demand created by strong growth. The movement of labor in a controlled manner from rural areas to the cities changed the marginal productivity of labor supporting the growth and, ceteris paribus, the buildup in reserves. And, finally, the reserve currency country was all too ready to supply excess reserves (well beyond what would be needed to support low inflation, slower growing domestic economy). The confluence of these four factors during the recent decade all contributed to the strong demand for foreign reserves in China. Obviously the reserve flows could have been neutralized if a floating exchange rate had been allowed or if domestic credit had been utilized to support money demand. Taking

---

14 Three papers that discuss Asian demographic and reserve management issues should be considered. Putnam and Zecher (2008) look at reserve management policy changes following the Asian crisis and its implications for monetary policy. Risk characteristics of demographics differences under democracies leading to different policy perspectives exist for the seventies and eighties are discussed in Putnam and Wilford (2002) and China’s demographic characteristics are highlighted in Kaoth et al. (2009).
these factors individually one may have prescribed a different policy set but in conjunction conclusions about the best course of action are not as clear.

Consider the aging population question that is rapidly descending on Japan, Europe, and soon, China. Just as Japan and Germany accumulated assets and excess claims on the productivity of younger economies over the past 30 years, good long term economic planning suggests that China should do the same. The question is what younger populations should be the recipient of those investments. One can argue that the status of the U.S. as the reserve currency country makes it the likely choice for ‘coupon clipping’ by an older population, even if it itself is getting older, although at a slower rate (especially given immigration). And one might argue that the dollar global bond rates we are observing for countries such as Chile and Brazil suggest that they are also receiving a portion of this demand. The supply of secure fixed income instruments most readily available and most easily traded and discounted are, however, still those of the reserve currency country; thus the U.S. remains the real beneficiary.

With the Chinese population aging quickly, much of it still underutilized in rural areas, and in a country where traditionally personal property rights cannot be taken for granted, the population itself necessarily has a high savings rate. Where one cannot diversify savings directly, there is often a greater demand on the banking system for fixed income investments and eventually through high-powered money, ceteris paribus, a greater demand for foreign reserves (at the central bank level). With a banking system still in development and capital controls a long way from being phased out, this need for savings passes through the monetary authorities and manifests itself in the accumulation of foreign reserves. As wage earners increase in productivity based on the movement from the countryside to the city, the income rise creates a demand for money. Capital markets, with banking system evolution over time to something more akin to the Western system, will satisfy these needs, but for the past decade the burden has fallen on the monetary authorities, exacerbating the overall demand effect related to foreign reserves.

One has to appreciate the rural to urban migration’s impact on growth, reflected in a surge in the marginal productivity of labor. Higher labor productivity leads to higher return on capital, unless distinct policy choices are made to compensate for the higher labor productivity. That is, if the marginal laborer can garner in wages the marginal increase in productivity then there will be no excess return to capital and thus a much lower increase in the capital stock to be applied against the next laborer venturing from the subsistence farm economy. Government policy has been to keep, on the margin, a high expected rate of return on capital to promote capital investment and create more demand for labor. The Chinese authorities would have chosen a flexible exchange rate regime only in the case in which they perceived no need to accumulate foreign reserves. In the flexible exchange rate regime, labor prices would be adjusted through the exchange rate as capital flows (or rather is free not to flow into China for new investment) to the highest risk adjusted return. But, a flexible exchange rate regime would have meant zero accumulation of foreign reserves. That is, there would be no insulation from policy mistakes of other countries (including especially the reserve currency country) and no international saving for the future of an aging population.

With a fixed exchange rate regime, the process of adjustment in labor prices would occur through wages (perhaps following increases in domestic inflation). Wage demands following demand-pull inflation would be the normal route for adjustment if the money supply growth reflected the buildup in high-powered money as reserves grew. To prevent this from occurring, the monetary authorities have used various tools to “cool off the economy” fearing inflationary expectations driving wages. This intervention, of course, supported increased demand for foreign reserves as part of the process of the economy seeking the long run equilibrium. But as David Hume noted, the mechanism will reach equilibrium over time and the authorities, while building up those foreign reserves, may find that they only delay the inflationary (or exchange rate) adjustment. The disequilibrium showing up in reserve buildup (perhaps wanted for the reasons stated above) and a continued high return on the marginal capital invested will ultimately disappear as all misallocations of resources do and move back to equilibrium, possibly even in a violent shift such as has occurred in the West’s banking and housing industries over the past three years. This could occur through a rapid change in the exchange rate, a rapid change in wages and domestic inflation, or more slowly through some combination (most likely the desired goal of the authorities).

For the investor expecting a sustained high real rate of return on the marginal dollar invested in China, these adjustments will lower those expectations and should show up in prices of freely traded equities. Of course the timing of these shifts toward equilibrium of marginal prices of labor and capital, as well as the mechanism through which the adjustment occurs, are uncertain.

This brings one to the other side of the reserve story; the creation of excess reserves by U.S. policy helps feed the disequilibrium in Chinese markets, just as Chinese policy affects demand for foreign reserves. Ceteris paribus, both factors are working the same way to create a greater buildup in official holdings of foreign reserves. Some have called this symbiotic relationship a win-win! Others take the opposite view. They argue Chinese demand created the (low interest rate and thus housing) bubble in the dollar markets. Lower interest rates then fueled the emerging markets bubble economies. Finally it is argued that this affect is exacerbated by Quantitative Easing I and II. The situation may have been a
win-win for politicians (American politicians could more easily place their debt and thus expand government spending and the Chinese politicians could create jobs in the hope of absorbing more underemployed labor). Whether it has been a win-win for either economy will not be clear until the disequilibrium that is obviously there has adjusted.

In the West, if the pessimists above are correct about the effect of Chinese demand for treasuries on the housing bubble through lower interest rates, then the negative side of the adjustment is already occurring. It will likely follow that China will experience inflationary pressures or an adjustment in the price of labor, a shift in expected return from capital (stock market implications are obvious) downward, and/or a wage adjustment through the exchange rate mechanism. The MBOP story outlined above will not tell one when or what channel an adjustment will occur. It simply suggests that the buildup in reserves has been logical and the fundamental factors that may have caused it to be excessive will shift, showing up in potentially unpleasant ways.

References

• Gupta, S., 1984, “Causal relationship between the domestic credit and reserve components of a country’s monetary base,” Kredit und Kapital, 17:2, 261 – 271
• Hume, D., 1752, Of the balance of trade, Essays, moral, political and literary, essay V
• International Financial Statistics: various issues
• Johnson, H., 1973, Further essays in monetary economics
• Kamp, D. S., 1975, “A monetary view of the balance of payments,” Federal Reserve Bank of St. Louis Review, April, 14 – 22
• Maddison, A., 1996, Chinese economic performance in the long run, OECD, Paris
• Mundell, R., 1968, International economics
• Mundell, R., 1971, Monetary theory
• Musa, M., 2007, “IMF surveillance over China’s exchange rate policy,” Paper presented to the Conference on China’s Exchange Rate Policy at the Peterson Institute
• Peoples Bank of China, various reports
• Stone, R., 2010, “Asia’s looming social challenge: coping with the elder boom,” Science, 330, 1599
PART 2

Asset Allocation: Mass Production or Mass Customization?

Brian J. Jacobsen — Chief Portfolio Strategist, Investments Group, Wells Fargo Funds Management, LLC

Abstract
Assuming a client’s goals, resources, and constraints have been clearly identified, when constructing an asset allocation instead of using a generic efficient frontier a client should have his or her own efficient frontier. What is efficient for one person may not be efficient for another. In this paper, I outline a framework for creating these customized allocations.

1 The views expressed are as of December 15, 2009 and are those of Brian Jacobsen and not those of Wells Fargo Funds Management, LLC. The views are subject to change at any time in response to changing circumstances in the market and are not intended to predict or guarantee the future performance of any individual security, market sector, or the markets generally. Wells Fargo Funds Management, LLC, is a registered investment advisor and a wholly owned subsidiary of Wells Fargo & Company. Not FDIC insured, no bank guarantee, may lose value.
An advisor’s greatest value can be in helping a client identify goals, recognize constraints, and to then formulate, implement, and monitor a plan that accomplishes the goals with the highest probability of success. When possible, instead of identifying a single solution for a single goal, many times it is better to identify a single solution to many goals at once. This is the underlying philosophy of ‘customized asset allocation.’ Each client will have unique needs, exposures, resources, and preferences.

Financial planning is client-centered. It focuses on the goals and constraints of each individual client. One part of financial planning, investment management, is organized around asset allocation and security selection [Jahnke (2003)]. In this paper, I focus on the asset allocation decision.

What asset allocation is and is not

Asset allocation refers to the proportion of a portfolio that should be placed in the relevant asset classes [Blair (2002)]. An asset allocation strategy is the set of decision rules that enables the asset manager to determine an asset allocation at a particular point in time. Even though some have defined asset allocation as building a diversified portfolio utilizing different asset classes or investment categories [Cardona (1998)], in fact, diversification and asset allocation are related, but not identical. An asset allocation could involve making a concentrated investment. Whether to make a diversified investment is a follow-up consideration to the asset allocation decision.

Typically, investors approach asset allocation in two steps. First, they determine the optimal allocation to each broad asset class. Second, active and passive managers are selected to implement the allocation within each asset class [Clarke et al. (2002)]. Using this two-step approach can inadvertently add systematic risk to a portfolio. An active manager’s investment strategy may increase or decrease systematic risk at different times. For example, if you measure the active risk of a manager by his or her tracking error, this tracking error will likely depend on whether markets are up or down. The Wells Fargo Advantage Common Stock Fund has a daily tracking error of 0.36% in down markets (i.e., when the Russell 2500 had a negative daily return), but a daily tracking error of 0.27% in up markets. If active risk, which is the risk added by an active manager, was separable from systematic risk, represented by broad exposure to a benchmark, then the down market and up market tracking error would be nearly identical.

Because of transaction costs, agency problems, and limited knowledge on the part of consumers, asset allocation will and should become an activity performed by financial intermediaries, rather than by their retail customers [Bodie (2003)]. There are numerous benefits of using mutual funds to implement an asset allocation strategy, including the following: professional management, diversification (across securities and within an investment style), economies of scale, and flexibility (i.e., the exchange of fund shares easily) [Cardona (1998)]. Investors can obtain manager diversification along with security diversification by including multiple mutual funds to represent a particular asset class.

What is an asset class?

It is often useful to group similar investments together into asset classes. This begs the question, though, of what makes different investments similar? Categories can be useful, but are not perfect. A danger inherent in asset allocation is assuming that asset categories or asset classifications create homogeneous groups of securities. Assuming securities are more correlated or subject to the same risks within a category than between categories can be a mistake. Another problem is what Ennis (2009) refers to as “category proliferation and ambiguity.” It is always possible to further differentiate between categories, and some securities seem to be so unique that they defy categorization, going into an ‘alternative’ category.

All securities are imperfect substitutes for all others, but some are better substitutes than others. One idea that caught on in the investment community is the idea of ‘style boxes.’ This typically categorizes equity securities into whether they are large, mid, small, and then value, core, or growth stocks. Different index providers provide their own categorization schemes. However, the whole notion that style boxes are useful for asset allocation is flawed in that the original justification as presented in Fama and French (1992) would suggest that everyone should hold small cap value stocks and that no one should invest in large cap growth stocks. Yet, these categories, which are useful for identifying a few broad common risk factors, have become thought of as silos into which one should invest.

Alternatively, instead of thinking of asset classes, we prefer to think of exposures to systematic and active risks. The allocation decision is then...
how much systematic and active risk the investor wants in the portfolio. The risk of a security can be described by its exposure to systematic factors, the volatility of those systematic factors, and the residual (or active) risk. An investor can get passive exposure to systematic risks with little effort or cost. Active risk comes from ongoing decisions made in managing and revising a portfolio.

Systematic risk exposure can be defined in a number of ways. Generally, it is anything that indiscriminately impacts a large part of the market. For example, sector exposure can be thought of as a risk exposure. Another systematic risk exposure can be the type of security. For example, all debt securities can be subject to factors that equities are not directly exposed to. Taxation and bankruptcy law changes are the two factors that seem most significant in terms of the differential risk factors between equities and debt. There are also factors related to interest rates that more immediately impact debt compared to equities.

As a starting point, asset allocation then depends on determining how much exposure you want to the systematic risks. Anson (2004) refers to these exposures as the ‘beta drivers’ of a portfolio. In the long run, this decision will tend to dominate the contribution to total portfolio return. After determining the systematic risk exposure, investors then need to decide whether to add active risk to the portfolio through active management. Active managers would be what Anson (2004) refers to as the ‘alpha drivers’ of a portfolio.

Investors can be thought of as having a ‘risk budget.’ This is a summary of the risks the investor is willing and able to bear. This risk budget then needs to be allocated across systematic factors. Bearing systematic risk is rewarded with an expected market-wide risk premium. Active returns, on the other hand, depend on the skill of particular managers. Investors may have some concerns about managers’ abilities to generate positive active returns consistently, and thus may be less tolerant of bearing active risk.

Determining what assets are allowed in a portfolio
Not every investment that is available will be appropriate for every client. Most investors will likely be constrained to investing in assets that are available through their advisor. Even all of those assets that are available might not be appropriate. A screen of some sort needs to be constructed to filter out what is or is not appropriate. A good initial screen is based on the risk tolerance of an investor.

a. For conservative investors, predictability may trump diversification
Securities have different characteristics in terms of the predictability of cash flows. As Bodie (2003) points out, a person’s welfare depends not only on her end-of-period wealth but also on the consumption of goods and leisure over her entire lifetime. As a result, multi-period hedging rather than diversification can be a preferred way to manage market risk over time. This could be interpreted as meaning that a ‘conservative’ investor would prefer a more concentrated portfolio over a diversified portfolio, if it means the cash flows generated are more predictable.

The typical way to deal with risk aversion is to increase cash and bond holdings relative to equities. We prefer a framework wherein conservative investors will want relatively more self-liquidating assets than an aggressive investor. If you view any security as a bundle of claims to both predictable and unpredictable cash-flows, self-liquidating securities are those which have a predetermined structure over which they are paid-off. For example, a preferred stock has a predetermined cash flow stream (albeit, subject to some uncertainty) whereas a common stock does not. A share of preferred stock is not as self-liquidating as a fixed rate coupon bond, though. An investor can hold onto a fixed-rate bond until maturity and receive the contractual cash flows (ignoring the possibility of bankruptcy).

Inflation risk obviously plays a role here too since fixed-rate securities are fixed in nominal terms, not real terms (inflation adjusted). Thus, a variable rate or adjustable rate security may be more self-liquidating than a fixed coupon bond. The consummate self-liquidating security is a Treasury Inflation Protected Security (TIPS). It is only subject to the default risk of the government. At maturity, it pays off the inflation adjusted principal or the face value, whichever is greater.

Because of the changing nature of clients and the economy, you cannot systematically favor a single asset. Overly simplistic or extreme strategies are a disservice to the client. Instead, a systematic set of rules is needed to transform expectations into a strategy [Arnott and von Germeten (1983)]. In Bodie et al. (1992), they highlight the importance of an individual’s labor income (also referred to as human capital) in determining an asset allocation strategy. Because human capital is usually less risky than returns on equities, they analogize human capital to a fixed income security. At any given age, the greater the flexibility to alter one’s labor...
supply, the greater the amount that should be invested in risky assets. Individuals may be able to offset financial asset losses through adjusting the amount they work. Bodie (2003) says the value, riskiness, and flexibility of a person’s labor earnings are of first-order importance in optimal portfolio selection at each stage of the life cycle. From our perspective, one shortcoming of analogizing human capital to a fixed-income instrument is that it downplays the ‘optionality’ of labor: a worker can choose to change jobs or extend/contract his working life. This means that it may be more appropriate to compare human capital to a convertible bond or a bundle of a bond and an option. This can create a justification for a higher allocation to active management strategies as an individual gets older. As the individual ages and the human capital ‘loses its optionality,’ the financial wealth should have more optionality. A holistic approach to financial planning requires recognizing that human capital and financial capital are inextricable in building an appropriate asset allocation strategy for a client.

b. Liquidity is valuable and its price changes with market conditions

Allocations should be modified according to an investor’s need, or lack thereof, for liquidity. Investors with a short investment horizon have a need for liquidity and, therefore, should have a relatively higher proportion of the portfolio in liquid securities. Liquidity does not just refer to how quickly, but also to how predictable the price at which, a security can be sold. Thus, for a fixed income allocation, a short-term investor should have more in cash equivalent assets (i.e., money market funds). Long-term investors have less of a need for liquidity. Thus, a long-term investor can be relatively short liquid securities. This means things like private equity, hard assets, and equity securities should be a relatively larger part of the long-term investor’s portfolio.

Liquidity is also important since it helps determine how returns should be measured for the inputs to a portfolio optimization system (see below about how to optimize a portfolio). An investor with only 1 month of cash needs to be concerned about one month returns on the various asset classes since they are living ‘one month at a time.’ An investor with six months of cash needs to consider risks and returns over a six-month holding period since the worst case scenario is that the investor will run out of cash in six months and needs to recapitalize after that point.

The most liquid security is usually the ‘risk-free’ security. But, there is no one universal ‘risk-free security.’ What is risk-free depends on the client’s funding needs. This can be thought of as employing a cash-flow matching scheme where cash becomes available just as it is needed. Portfolios structured this way are sometimes called ‘laddered portfolios’ (Bodie and Crane (1997)).

Customized optimization for a client

One of the most well-documented, and perhaps most controversial, methods of determining an asset allocation is through Modern Portfolio Theory (MPT). The concept behind MPT is simple: if securities behave like random variables, then based on the assumed probability distributions of the securities, an investor must examine the combinations of securities to find the most attractive combination. The typical method of measuring portfolio ‘attractiveness’ is through selecting amongst the possible risk-return combinations of the portfolios. A typical decision rule is to minimize the risk of a portfolio for any given target return. Alternatively, an investor answers the question, “what is the combination that maximizes the return for a targeted level of risk?” These are referred to as the “efficient portfolios,” in the language of MPT.

Central to MPT is the premise that investment decisions are made to achieve an optimal risk/return trade off from the available opportunities. The critiques of MPT typically focus on the inappropriateness of particular measures of risk or return. Because asset allocation – just like investing – is forward-looking, these measures should be based on expectations of the future, but most implementations are based on historical averages, standard deviations, and correlations.

The first step in MPT is to quantify ex-ante measures of risk and return for the appropriate set of assets. The next task is to isolate those combinations of assets that are the most ‘efficient,’ in the sense of providing the lowest level of risk for a desired level of expected return, and then to select the one combination that is consistent with the risk tolerance of the investor. A key element is to define risk. Typically, for mathematical tractability, standard deviations or variances are chosen as the measures of risk. However, an asymmetric measure of risk that focuses on returns below a specified target or benchmark return level can also be used (Harrow (1991)). These asymmetric measures of risk are also referred to as ‘downside-risk measures.’ In practice, using a downside-risk framework

10 If a worker’s human capital is very risky (i.e., it is not easy to change jobs or work more) or wages become systematically less risky over the life cycle, the optimal equity fraction could actually increase with time.
11 A familiar proposition is that investing in common stocks is less risky the longer an investor plans to hold them. If this proposition were true, then the cost of insuring against earning less than the risk-free rate of interest should decline as the investment horizon lengthens. Bodie (1995) shows that the opposite is true (whether stocks follow a random walk or even if stock returns are mean reverting in the long run). The case for young people investing more heavily than older people in stocks cannot, therefore, rest solely on the long-run properties of stock returns.
12 Harry Markowitz is considered the founder of MPT. The Markowitz (1952) model assumes individuals make decisions ‘myopically’ in a static, single-period framework with risks and returns defined for a single holding period. This model has been extended in various ways. Robert Merton extended the model, assuming that individuals make decisions dynamically over time, behaving as though they are trying to maximize their expected utility from consumption of goods and leisure over their lifetimes, and they are free to change their choices at any time (Bodie and Crane (1997)). The equilibrium counterpart to Markowitz’s model is the Capital Asset Pricing Model. The counterpart to Merton’s is the inter-temporal capital asset pricing model (Bodie and Crane (1997)).
produces significantly higher bond allocations relative to stocks. This composition increases downside protection while offering the same or greater level of expected return [Harlow (1991)].

The optimization of the portfolio allocation can be achieved in either one stage or two. In the single-stage optimization, allocations are determined by satisfying both active and systematic risk preferences simultaneously. In the two step process, investors first choose a long-run benchmark portfolio based on the risk-return relationships for bearing systematic risk. Once this allocation is established, the investor then decides how to structure the portfolio using a combination of passive and actively managed strategies. The allocation to active strategies can be structured by maximizing the trade-off between expected portfolio return and active risk while holding the level of systematic risk constant as embodied in the long-run benchmark allocation. This is the typical process used when advisors determine a client needs specified exposures to broad asset classes and then a recommended list of mutual funds is referred to in order to populate the portfolio. As we noted above, this is a suboptimal method of optimizing a portfolio: active managers bring systematic risk to a portfolio and that systematic risk exposure can change over time. Thus, a two-step process is inferior to a one-step process. The advisor must evaluate active risk and systematic risk jointly, and not separately. All sources of systematic and active risk need to be monitored and evaluated to make sure they are consistent with the client’s risk preferences.

For example, domestic equity exposure could be modeled by using three asset classes represented by the Russell top 200, the Russell mid-cap index, and the Russell small-cap index. Using the historical risk-return numbers, as of September 30, 2007, with monthly returns would have given the following descriptive statistics as inputs to an optimization tool:

<table>
<thead>
<tr>
<th>Index name</th>
<th>Average monthly return (annualized)</th>
<th>Standard deviation of monthly returns (annualized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russell top 200</td>
<td>13.23%</td>
<td>14.79%</td>
</tr>
<tr>
<td>Russell mid-cap</td>
<td>15.57%</td>
<td>16.02%</td>
</tr>
<tr>
<td>Russell small-cap completeness</td>
<td>10.12%</td>
<td>19.93%</td>
</tr>
</tbody>
</table>

Source: Zephyr Allocation Advisor

Table 1 – Average and standard deviation (annualized) of monthly returns for indices, from April 1999 to September 2007

A portfolio with a target risk of 14.76% (the minimum risk portfolio, assuming non-negative investments in each index), would have been 12.4% in the Russell top 200 index and 87.6% in the Russell mid-cap index. The average return for this portfolio was 13.52%.

If, instead, the ‘Wells Fargo advantage small cap opportunities fund’ was used as a replacement for the Russell small-cap completeness index, then the same target risk (14.76%) would have been achieved with a 37.4% allocation to the Russell top 200, 30.8% allocation to the Russell mid-cap, and 31.8% allocation to the small-cap opportunities fund. The portfolio’s average return would have also increased from 13.52% to 14.40%. This partially reflects the fact that active risk and systematic risk are not necessarily separable.

Determining the “neutral” allocation

Instead of generating forecasts of risks and returns and optimizing a portfolio based on that outlook, some prefer to appeal to the competitiveness of markets to develop asset allocations. If markets are competitive, in the sense of incorporating relevant information in current prices, then the market allocation should be an optimal allocation for a ‘typical’ investor.

One way to measure the market allocation is based on the calculated market capitalization of the various assets. This is commonly calculated as the closing price of the security times the number of securities outstanding at any given time. One relatively recent innovation is to base the market capitalization on the ‘free-float’ of the shares outstanding. This includes in the number of shares outstanding only those that are available to domestic investors.

There is a philosophical question as to whether this is actually appropriate for calculating the market capitalization, though: if only 2 shares of stock sold for $100, would it be appropriate to say that all shares of that company’s stock are worth $100 a piece? If markets are ‘continuous,’ in the sense that prices do not make large jumps, then this method of calculating market capitalization is not too objectionable. However, if prices can jump and the size of the order matters, then this method of calculating market capitalization, which goes into determining the relative weights of securities in a portfolio, seems suspect. We prefer a ‘liquidity weighted’ metric, where what matters is the relative dollar trading volume of securities. There is actually no statistically discernible relationship between the dollar trading volume of a security and the calculated market capitalization of a company (Figure 1).

Different weighting schemes reflect different views about how security prices will likely move. For example, a price-weighted portfolio is equivalent to a momentum investing strategy. An equal weighted strategy or a fundamental-weighted portfolio is equivalent to a contrarian investment strategy. The trading volume weighting scheme is not materially different from an equally weighted weighting scheme. Based on this evidence,
an equally weighted portfolio may be the best starting point for an allocation called the ‘neutral allocation’ or ‘the market allocation.’

Institutional investors as well as individuals have to face the problem of managing their assets in a way that ensures that all their liabilities (including future liabilities) can be fully met and their financial goals can be achieved. This type of investment management is widely known as asset and liability management (ALM) [see Hocht et al. (2008) for a summary]. An optimal strategy for a client here means achieving financial and personal goals under a given set of financial constraints. In contrast to ordinary financial planning methods, the ALM approach simultaneously considers the joint uncertain evolution of assets and liabilities. This approach can be superior if, for example, assets and liabilities share common risk factors and liabilities hedge some assets.

For example, according to the September 17, 2009 flow of funds release of the Federal Reserve (Table B.100 balance sheet of households and nonprofit organizations), households had U.S.$18 trillion in real estate holdings and U.S.$42 trillion in financial assets. Thus, a typical household will have 30% of its portfolio in real estate. For anyone who owns a home, he or she may have a very concentrated exposure to real estate. Adding exposure to real estate through a real estate investment trust (REIT) may throw off the asset allocation. If it were possible to short a home, then a diversified exposure to real estate could be achieved by shorting the personal residence and investing in a well diversified REIT.

Single step customization
Asset allocation and asset location have to be considered jointly

When it comes to asset allocation, asset location also matters: different ways in which you title an asset or the type of account carries profound tax and other legal implications. Assets and asset strategies should be allocated to different vehicles based on their tax efficiency, but the primary purpose of developing an asset strategy is to achieve the client’s goals, not just to have a tax efficient portfolio. There may be very good reasons a client will prefer to subordinate tax efficiency to other goals.

There are some basic rules of asset location that arise due to differential tax treatment of different types of accounts that Waggle and Englis (2000) outline. Typically, long-term equity holdings should be in nonretirement accounts while taxable bonds should be in retirement accounts to improve tax efficiency. Generally, only those in the highest marginal tax bracket should hold municipal bonds. An advisor needs to locate assets in the most advantageous places, in the most advantageous mix to meet various client goals with the highest probability [Fowler and Vassal (2006)].

Conclusion
Putting this all together suggests that each individual client’s portfolio needs to be custom-optimized. So, in our framework where you have short-term, medium-term, and long-term investors, you will have ‘efficient frontiers’ for each time horizon because what is optimal in the short run is not necessarily optimal in the long run. You will also have different efficient frontiers, due to the different allowable investable assets. As a result, instead of one efficient frontier, you will have nine efficient frontiers.

The appropriate portfolio for a client needs to be modified according to the investor’s investment horizon and willingness to accept risk. As Table 3 illustrates, a basic framework for asset allocation is to think of nine different types of investors, categorized according to the client’s investment horizon (short-, medium-, and long-term) and risk tolerance (conservative, moderate, and aggressive). This is a highly stylized framework as each dimension should be thought of as a continuum and not simply a matrix of boxes. We center the framework on the neutral allocation (equally weighted, as described above), defining it as the medium-term/moderate allocation. The investment horizon and risk tolerance is always measured relative to the neutral allocation.

Individual investors have their own goals and constraints that make it such that cookie-cutter solutions are rarely optimal. An advisor can help a client articulate those goals, identify constraints, and develop a coherent strategy. This involves developing an efficient frontier for each client. The client’s risk preferences will dictate which types of securities go into the investable universe. The client’s need for liquidity will determine over which horizon risks and returns should be measured. As such, there is no one single efficient frontier. This is why asset allocation cannot be viewed as a mass production technology. At best, it is mass-customization.

Table 3 – Asset allocation framework

<table>
<thead>
<tr>
<th>Allocation framework</th>
<th>Risk tolerance</th>
<th>Conservative (i.e., fewer U.S. treasuries)</th>
<th>Moderate</th>
<th>Aggressive (i.e., more U.S. treasuries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investment horizon</td>
<td>Long term</td>
<td>Short liquidity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medium term</td>
<td>Bias toward self-liquidating securities</td>
<td>Neutral</td>
<td>Bias away from self-liquidating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(i.e., more fixed income)</td>
<td></td>
<td>securities (i.e., more equity)</td>
</tr>
<tr>
<td></td>
<td>Short term</td>
<td>Long liquidity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References

- Blair, B., 2002, “Conditional asset allocation using prediction intervals to produce allocation decisions,” Journal of Asset Management, 2, 325-335
- Kiefer, N., 2008, “Annual default rates are probably less than long-run average annual default rates,” Journal of Fixed Income, Fall, 85-87
- Louton, D., and H. Saragouzi, 2008, “How many mutual funds are needed to form a well-diversified asset allocated portfolio?” Journal of Investing, Fall, 47-63

14 De Brouwer (2009) writes about how well the findings of Harsh Shethrin and Meir Statman (2000) summarized in their Behavioral Portfolio Theory are in agreement with the more general psychological theory about the hierarchy of needs as formulated by Abraham H. Maslow (1943). Maslow’s Hierarchy of Needs is the theory in psychology that Abraham Maslow proposed in his 1943 paper “A Theory of Human Motivation”. The basic idea is that human needs are not all addressed simultaneously, but layer by layer. One only feels the need to fulfill the needs of a certain layer if the layer below is already fulfilled, if not then the layer below will get all of one’s concern. Maslow recognizes five levels. The first four lowest levels are grouped together as ‘deficiency needs’ and are associated with physiological needs. When these levels are met, the individual will not feel anything special, but when they are not fulfilled, one will become anxious. The top level is called ‘growth needs,’ and is associated with psychological needs. Deficiency needs must be satisfied first, and once these needs are met, one seeks to satisfy growth needs and hence seeks ‘self actualization.’
Abstract
This paper discusses a general framework for risk measurement for an asset liability management (ALM) process for a bank. Market value- and accrual earnings-based risk measurements are typically used in any ALM process. Market value-based measurements, however, are difficult to decipher in the presence of highly volatile market factors. The adverse impact from some of these factors can be controlled, but some factors are considered non-controllable. As a result, to be effective, the bank must evaluate risk measurements so that the hedging actions executed to mitigate risks are addressing relevant risks and contribute positively to the long-term profitability or viability of the bank. An effort is usually made so that the bank does not attempt to hedge the transitory risk signals arising from volatile non-controllable factors. This paper provides a practical approach and an illustrative example to demonstrate how to isolate impacts from controllable and non-controllable factors to facilitate effective ALM hedging actions for a held-to-maturity (HTM) portfolio.
ALM is one of the key processes at any bank. A bank typically leverages its capital by investing in assets and then funding these assets using equity, deposits, and debt with the intention of earning a spread between the asset yield and the liability yield. The ALM process involves managing credit, market, liquidity, and operations risks so that the bank can earn a return that is commensurate with the risk tolerance established by its board of directors. The risk management aspect of the ALM process, particularly the market risk management, becomes overwhelming due to multiple market factors affecting the risk measurement and management process.

The primary source of market risk under an ALM framework is interest rate risk (IRR). The Basel Committee on Banking Supervision considers the interest rate risk arising from both trading and banking activities, namely “interest rate risk in trading book” and “interest rate risk in banking book.” The ALM process in a bank typically seeks to manage the “interest rate risk in banking book.”

The Office of the Comptroller of the Currency (OCC) discusses two common approaches for assessing interest rate risk in the banking book – the ‘earnings’ approach and the ‘economic’ approach. The earnings approach, also known as ‘earnings simulation,’ considers the sensitivity of the bank’s accrual earnings (net interest income) to changes in interest rates. The economic approach, also known as the ‘market value or mark-to-market (MTM)’ approach, determines the sensitivity of the market value (net present value) of assets, liability, and equity to changes in interest rates. In the measurement of interest rate risk, both approaches are complementary and are used in conjunction with risk management. For each of the above approaches, there is a variety of metrics that are used in risk management. These are discussed in the next section.

**ALM and risk measurement metrics**

The traditional approach to asset liability market risk management involves measurement of the interest risk sensitivity of the asset liability portfolio. These metrics include:

- Effective duration of asset, liability and duration of equity (DOE).³
- Partial durations or key rate durations of asset, liability, and equity.
- Vega and convexity of asset, liability, and equity.
- Market value of equity, market value to book value of equity (MV/BVE).
- Value at risk (VaR).
- Earnings at risk (EaR).

The first four measurement types provide information on the level of mismatch of risk sensitivities between assets and liabilities. They are based on the market value or MTM perspective and provide sensitivities to changes in market value of asset or liability or equity due to changes in interest rates and other market factors, such as volatility. These measurements capture risks arising primarily from maturity and repricing gaps in the banking book.

The market value of equity or net portfolio value equals the market value (using either mark-to-market or mark-to-model) of all assets less the market value of all liabilities including all off-balance sheet items. Changes in this measure capture the impact of interest rate changes on all future cash flows of the bank. The change in market value of equity is a leading indication of changes in future accrual earnings of the bank. A similar measure is the market value/book value of equity (MV/BVE) which equals the market value of equity (as defined above) divided by the book value of equity.

Duration of equity (DOE) is a risk metric that equals the percentage change in the market value of equity when interest rates move by a parallel shift of one percent. Positive DOE implies a loss in market value if interest rates shift up in parallel fashion.

Value-at-risk (VaR) is also an MTM type of measurement and establishes a number that represents possible loss to market value of equity based on a full valuation run of the balance sheet for a distribution of scenarios for a certain holding period at a given confidence level.

The last measurement, earnings at risk (EaR), is based on an ‘earnings simulation’ perspective and estimates loss of future accrual income over a certain horizon at a given confidence level based on a distribution of changes in market factors. It is typically achieved through net interest income simulation over a mid-term horizon. EaR is calculated using static balance sheet (run-off of volumes with no replacement) assumptions as well as dynamic balance sheet (incorporating new business volumes) assumptions.

The MTM measurements are naturally based on a static balance sheet in which the calculations assume that existing assets and liabilities run off. EaR, on the other hand, can be run with both static and dynamic balance sheets and can accommodate new business forecasts. EaR measurements are usually based on a mid-term horizon and thus do not fully capture the long-term risks in the portfolio. MTM measurements consider all cash flows of all instruments on the balance sheet and therefore provide a comprehensive picture of all risks present.

---

1. Basel Committee on Banking Supervision, 2004, “Principles for the management and supervision of interest rate risk,” BIS.
3. Appendix A.
The MTM-based and accrual earnings-based measurements together provide the ALM process with the necessary information to understand risks and corresponding returns. Analyses of these measurements facilitate ALM actions according to the bank’s established risk/return objectives and guidelines. Depending on specific situations or style or philosophy of the ALM process, the bank may put more emphasis on one or both types of approaches. A detailed analysis and review of ALM risk measurements typically results in specific hedging actions resulting in appropriate changes to the risk profile of the balance sheet. Once hedging actions are executed, MTM-based risk measurements are immediately impacted. However, since EaR measurements involve calculation of future accrual earnings, the full effectiveness of hedging actions can only be realized through time and is subject to changes in market conditions.

In addition to these measurements, a bank also performs ‘stress tests’ and ‘scenario analysis’ to understand the extent of possible losses or gains under various scenarios. Depending on the risk appetite of the board of directors, certain boundaries are established in the form of risk limits to ensure that the bank earns returns by taking risks within acceptable risk limits.

**MTM valuation methods for ALM instruments**

The MTM valuation methodology includes a combination of market-observed prices and model-derived values using observable market inputs. Consider a wholesale bank with an HTM portfolio of assets. HTM assets imply that the bank is planning to hold these assets to maturity and these kinds of assets are generally held in the banking book. Typical assets for a commercial bank include mortgage-backed securities (MBS), whole loan mortgages, and short-term investments. Through the ALM process, the bank would typically hedge the interest rate, volatility, prepayment, and other risks in its portfolio of assets through a portfolio of liabilities that may include short-term debt instruments (cash and cash like), medium- and/or long-term bullet, and callable debt. Also, depending on the accounting treatment and suitability of derivative hedges, the bank may mitigate risks using off-balance sheet instruments like swaptions and caps/floors.

The prices for MBS and whole loans are determined using market-observed prices. Using appropriate prepayment and interest rate models along with volatility assumptions, Monte Carlo simulation is typically used to calculate option adjusted spread (OAS) to the LIBOR swap curve. The short-term investments are valued by a ‘discounted cash flow’ (DCF) approach using the LIBOR swap curve with appropriate spreads and models for optionality. The liabilities are valued using a DCF approach using the bank’s funding curve, and a lattice-based model is used for instruments with optionality. Derivatives are valued using a DCF approach off the LIBOR swap curve with volatility assumptions that are consistent with market-observed option prices.

**Market risk factors affecting the mark-to-market valuations**

Table 1 depicts the specific market factors that affect the valuation of assets, liabilities, and off-balance sheet items in the bank’s portfolio.

The LIBOR swap curve represents changes (parallel and non-parallel) to LIBOR/swap curve; basis movement represents the relative spread changes (parallel and non-parallel) between LIBOR/swap and the funding curve; volatility represents market value change due to changes in volatility of mortgage assets (short volatility) and callable debt (long volatility); prepayment risk primarily arises due to rate changes and the prepayment option in mortgages and is captured via prepayment model; and OAS represents market value changes due to changes in mortgage OAS. Widening of OAS reduces market value of existing mortgages and vice versa.

The market value of equity in the asset liability portfolio is affected by all the market factors described above. Market risk is defined as the potential for loss due to adverse changes in market rates and prices, including changes in interest rates. Typically, the MTM component of market risk in the bank is manifested by a decline in market value of equity due to a decline in asset values or an increase in liability values.

As explained earlier, two of the MTM-based risk metrics for market risk management are DOE and MV/BVE. One way to monitor the impact of market risk factors is to assess and quantify changes in these two metrics due to market risk factors. The main market risk factors are general interest rate movements in the LIBOR swap curve (A), basis movement between the funding curve and LIBOR curves (B), changes in option-adjusted spread (OAS) for mortgages (C), changes in implied volatility (D), and others, such as changes in portfolio, capital stock, trading strategy, time, prepayments, and other miscellaneous factors (E).

<table>
<thead>
<tr>
<th>Asset</th>
<th>Mortgages and MBS</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liability</td>
<td>Short-term investments</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bullet debt</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Callable debt</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off-balance sheet</td>
<td>Derivatives</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equity</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1 – Market risk factors
Attribution methodology for MTM-based market risk metrics

A common framework for fixed income attribution is based on a multifactor approach [Fong et al. (1983)]. Changes are decomposed based on a number of key factors including the coupon and roll return, interest rate environment (including rate and yield curve movements), credit spreads, volatility, optionality, trading and timing, etc.

The attribution methodology presented here decomposes changes in MTM risk metrics based on the factors A through E discussed in the prior section.

\[ \Delta \text{MTM risk metric} = f(\text{LIBOR, basis, OAS, volatility, other}) \] (1), where, \( \Delta \)MTM risk metric is the change in the MTM-based risk metric. Libor is the factor representing interest rate movements (level, slope, curvature, etc) in the LIBOR swap curve (henceforth referred to as Factor A or FA). Basis is the factor representing relative spread changes between the funding curve and the LIBOR swap curve for each maturity (henceforth referred to as Factor B or FB). OAS is the factor representing option adjusted spread changes for mortgages (henceforth referred to as Factor C or FC). Volatility is the factor representing changes in implied volatility (henceforth referred to as Factor D or FD), other is the factor representing other factors (including error terms and cross correlation effects) in the LIBOR swap curve (henceforth referred to as Factor E or FE).

For example, the change in market value of the portfolio could be analytically computed as:

\[ \Delta \text{Market Value} = \sum_{i=1}^{M} \text{KeyRateDur}_i \times \Delta \text{LIBOR}_i \times \text{convexity} + \text{other high order terms} + \sum_{i=1}^{M} \text{PartialBasisDur}_i \times \Delta \text{basis}_i \times \text{convexity} + \text{other high order terms} + \text{SpreadDur} \times \Delta \text{OAS} + \sum_{i=1}^{M} \text{Vega}_i \times \Delta \text{volatility}_i + \text{changes due to other factors} \] (2), where \( i = 1 \) to \( M \) represent the key rate yield curve term-points. KeyRateDur\(_i\) is the key rate/partial duration for the Libor swap curve for each key rate term point, \( \Delta \text{LIBOR}_i \) is the change in the LIBOR swap curve at each key rate term point, PartialBasisDur\(_i\) is the key rate/partial duration for the funding curve for each key rate term point, \( \Delta \text{basis}_i \) is the relative change between the funding curve and the LIBOR swap curve at each key rate term point. SpreadDur is the percentage change in value due to option-adjusted spread movements, \( \Delta \text{OAS} \) is the change in option-adjusted spread, Vega\(_i\) is the Vega for each key rate term point, and \( \Delta \text{volatility}_i \) is the change in implied volatility for each key rate term point.

Alternatively, instead of analytical computations one could directly compute the effects due to each factor (FA through FE) through full valuation in the ALM risk system. For example, the effect of Factor A or FA, which can be analytically computed (as above) by \( \sum_{i=1}^{M} \text{KeyRateDur}_i \times \Delta \text{LIBOR}_i \times \text{convexity} + \text{other higher order terms} \) (3), could also be computed through full valuation in the ALM risk system by changing the LIBOR swap interest rate environment while keeping all other factors fixed and performing a portfolio revaluation.

Similarly, the effect of factor D or FD, which can be analytically computed (as above) by \( \sum_{i=1}^{M} \text{Vega}_i \times \Delta \text{volatility}_i \) (4), could also be computed through full valuation in the ALM risk system by changing the implied volatilities while keeping the interest rate environment and other factors fixed and performing a portfolio revaluation.

Thus, by varying the combination of portfolios and market environments using a series of setups in the ALM risk system, the impact of each factor can be isolated and computed. The contribution of each factor is calculated sequentially; the final factor effect FE is calculated as the difference between the total month-to-month change and the sum of all the other factor effects.

Results from full valuation are compared with results based on analytical computations (using factor sensitivities and factor movements as discussed in the equation above). Although, the methodology computes and isolates impacts due to each of these market factors, there is usually cross correlation between these factors.

Tables 2 and 3 introduce the terminology used. Table 4 details the ALM system batch process output that is used to compute each of the factor effects. Table 5 details how each factor effect is calculated from the runs above.
Separation of market risk factors into controllable and non-controllable

We introduced a factor-based attribution analysis methodology to decompose the month-to-month changes in the risk metrics MV/BVE and DOE using the factors (F_A through F_E). The primary focus of the attribution is from a risk management perspective and it aims to quantify the effect due to hedging actions that can be executed to maintain long run profitability of the bank by controlling the impacts due to the various factors.

All factors affect MV/BVE and DOE. However, all of these factors are not directly controlled by the ALM process. Recall that the ALM process is primarily concerned with managing the asset and liability (along with associated derivatives) positions to earn a return (net interest income, margin, spread, etc.) commensurate with the risk appetite of the bank. The ALM process typically manages the held-to-maturity (HTM) portion of the balance sheet, also known as the banking book. To the extent that some of these factors have an impact which is extraneous to the ALM process, they would be considered as non-controllable factors. These non-controllable factors could have a significant impact on key risk metrics and this may have an effect on hedging decisions and risk-adjusted profitability.

For example, if the assets are held to maturity, typically the ALM process will hedge the interest rate risks for the general level of interest rates, i.e., interest rate movements in the LIBOR/swap curve (factor A) and manage the volatility risk (factor D) by a combination of hedging instruments through derivatives (swaptions, caps) or cash instruments (callable debt). At the same time, impacts due to portfolio changes or hedging strategy are the other factors that the ALM process would manage. All these factors have a direct impact on the return metrics for ALM (accrual net interest income, margins, spreads, etc.) and thus need to be controlled by the ALM process.

On the other hand, for HTM portfolios, factors such as OAS, basis movement between LIBOR/swap and the funding curve can be considered extraneous factors or factors that are outside the scope of control of the ALM process. These factors are transitory in nature and have a tendency to revert to long-term mean values over a period of time. For example, for HTM asset-liability portfolio, the positions are held to maturity. The market values of these positions prior to maturity are subject to changes in OAS, basis movement between funding curve and LIBOR curve, changes in volatility, etc. However, at the time of maturity, the market prices converge to par and hence the price impacts from such extraneous factors disappear. Under certain assumptions, the impact of these factors on market value and risk measurements, while relevant for a trading book/MTM-based portfolio, is not realized for the banking book/HTM-based portfolio if there is an intent and ability to hold all balance sheet instruments or positions to maturity (no forced sales/liquidations). For example, in the absence of actual credit losses, changes in OAS will not have an impact on future accrual income and spreads of an HTM portfolio. Similarly, changes in credit spreads reflecting downgrade risk, while relevant for a traded portfolio, will not affect accrual earnings if there is no actual default for the HTM portfolio. Thus, these factors are non-controllable and are not directly hedged. Again, changes in spreads reflecting asset liquidity premiums or discounts in the market would not be a relevant risk for an HTM portfolio if there is an ability to hold assets to maturity. The main controllable factor for an HTM portfolio is the interest rate risk that is managed by a funding strategy based on the risk appetite of the bank.

Therefore, the above factors can be divided into two categories, controllable and non-controllable.

Controllable factors
- General interest rate movements in the LIBOR/swap curve (A).
- Changes in implied volatility (D).
- Other: changes in portfolio, capital, trading strategy, time, prepayments, miscellaneous factors (E).

Non-controllable factors
- Basis movement between the funding curve and LIBOR curves (B).
- Changes in option-adjusted spread (OAS) for mortgages (C).

In order to implement ALM strategies, it is important to isolate impacts due to controllable factors and non-controllable factors. For example,
the recent mortgage crisis has significantly widened OAS of mortgages due to illiquidity and dislocations in the mortgage market. The OAS's in the marketplace often resulted from forced liquidations/distressed asset sales and are not reflective of prices (and hence intrinsic values) that would result in an orderly market. The impact of widening OAS is a decrease in MV/BVE and an increase in DOE. However, if these market factors are non-controllable, the ALM process will need quantification of its impact so that this factor is not inadvertently hedged by the bank through the ALM process.

The methodology presented here provides a tool to the ALM process to isolate the impact on the risk management metrics due to controllable and non-controllable market factors. The ALM process should carefully evaluate the magnitude of the extraneous factors affecting the ALM risk measurements and appropriately hedge only those market movements or factors that affect long-term accrual income of the bank.

**Simplified illustration of the attribution methodology for a sample bank**

Consider a bank with a mortgage portfolio (MP) and a non-mortgage portfolio (NMP). For the sake of simplicity, assume that the bank is equally invested in both portfolios, i.e., asset sizes of 50% each for MP and NMP.

The MP contains residential mortgage whole loans (unsecuritized mortgage pools) and residential mortgage-backed securities (RMBS). Assume that the assets are funded primarily with a mix of bullet bonds, callable bonds, and discount notes or cash-like instruments. This combination of liabilities hedges the interest-rate, prepayment, and volatility risks of the portfolio of mortgage assets. In choosing the proportion of various liabilities of each type, the ALM process attempts to maximize long-run net interest income and margin subject to minimizing interest-rate and liquidity risks. The primary source of market risk at the bank arises from the mortgage portfolio.

Assume that the NMP contains short-term money market investments and swapped investments. Further assume that the money market investments are match funded with liabilities of equal duration. All other investments are swapped to synthetic floating rate assets and are funded with matched debt that is also swapped to synthetic floating rate liabilities. All optionality on the asset side is equally offset on the liability side. Thus, the NMP is not a significant source of market risk at the bank.

Attribution analysis is performed on a monthly basis on this sample bank from Sep. 30, 2007 to Sep. 30, 2010. In this example, it was assumed that the MV/BVE on Sep 30, 2007 was 100%. This three year period was marked by unprecedented changes in the markets due to the credit/financial/liquidity crisis resulting in dramatic declines in interest rates, steepening of the curve, widening of credit spreads and option-adjusted spreads, and increases in implied volatility.

Figure 1 shows the LIBOR/swap curve at the beginning and end of this period. Interest rates declined in a dramatic fashion across the curve with more movement at the front end of the curve. Three-month LIBOR decreased by 494 bps, two-year swaps decreased by 405 bps, and ten-year swaps by 264 bps. The LIBOR/swap curve steepened and separately the FNMA current coupon mortgage rate also fell by 259 bps.

Figure 2 shows the implied black volatilities for various swaption structures. As can be seen, volatilities on shorter structures declined quite significantly while volatilities increased on longer structures. As interest rates fell, credit spreads widened significantly. As a result, while the bank’s funding curve dropped, the drop in the funding curve was less than the corresponding drop in the swap curve. As Figure 3 shows, between 09/30/07 and 09/30/10, the basis movement (or the relative change...
between the swap curve and the funding curve) was negative since credit spreads widened. At the two-year point, the basis change was -14 bps, at the five-year point the change was -20 bps, while at the ten-year point the basis change was -47 bps.

At the same time, due to widespread dislocations in the mortgage markets, credit deterioration, and heightened illiquidity, option adjusted spreads (OAS) on mortgages fluctuated widely going from record lows to record highs. As can be seen in Figure 4, the FNMA current coupon OAS increased to 97 bps at the height of the crisis. This reflected primarily liquidity and credit concerns. Non-agency OAS widened even more significantly, especially on more esoteric products. Once again, this was reflective of demand and supply in the mortgage marketplace along with the underlying credit risk.

Due to these unusually large movements in the various market factors, the market risk metrics, namely MV/BVE and DOE, of the sample bank exhibited wide fluctuations. Figure 5 shows the monthly trend of MV/BVE ratio for the sample bank plotted on the left axis. MV/BVE declined from 100% to 42% through this period and then increased back up to 130%.

On the right axis are the monthly changes caused by each of the factors A through E (factors are defined in earlier section). As can be seen from the graph, the OAS factor (B) and the basis movement factor (C) caused the largest changes in the market value primarily in the period Sep 2008 to Jun 2009. This was not surprising given that this period was marked by huge changes in OAS and widening of credit spreads due to the credit and mortgage crisis. Widening of OAS on mortgages (factor B) in the final quarter of 2008 caused huge declines in the bank’s MV/BVE. This was offset to some extent by the negative basis movement between the funding curve and the LIBOR/swap curve (factor C). Since the funding curve dropped by less than the LIBOR/swap curve, the assets increased more in value than the liabilities; hence, equity increased in value due to factor C.

The change in MV/BVE due to controllable factors (factors A, D, and E) is much smaller relative to the OAS and basis (factors B and C). This shows that the bank was relatively well hedged to changes that the ALM process controls, i.e., interest rates movement, volatility changes, and portfolio changes.

The change due to non-controllable factors (factors B and C) overshadowed the change due to the controllable factors. Over the analyzed period, the controllable factors (sum of A, D, E) increased MV/BVE by 85% and the non-controllable factors decreased MV/BVE by 55%. Over the simulation period, total MV/BVE ratio increased by 30%. However, the part of the MV/BVE changes that is directly hedged and managed by the ALM process actually increased by 85%. Thus the ALM process has actually added more value than it would appear based purely on
The changes in MV/BVE. A similar attribution analysis is performed on the changes in DOE for the sample bank. Figure 6 shows the monthly changes in DOE throughout the simulation period (on the left axis) and the monthly changes attributable to each of the factors A through E (on the right axis).

The period of Sep 08 through to Dec 08 was marked by large increases in DOE due to the OAS widening (factor B) as mortgage assets extended in duration. The OAS widening was driven by dysfunctional markets, illiquidity, and credit issues. Negative basis movement (factor C) also caused large changes in DOE. From an ALM perspective, the increase in DOE was driven by factors that were external to the ALM process (non-controllable factors). For this sample bank all assets are held to maturity.

If the bank has the intent and ability to hold assets to maturity, then the extension of DOE due to non-controllable factors like OAS widening and basis movement is something that the ALM process would not hedge.

The interest rates factor (factor A) also caused large changes in DOE. Again, this was a period characterized by a great deal of volatility and changes in rates and that manifested itself in the changes attributable to Factor A. The controllable factors decreased DOE by 11 and the non-controllable factors increased DOE by 3. The total change in DOE over this entire horizon was -7.7. It can be clearly seen that the non-controllable factors had a fairly large role to play in the changes in DOE.

**Attribution analysis and the ALM process**

Figure 5 shows a trend in attribution analysis due to controllable factors and non-controllable factors. In the illustrative example above, the OAS and basis movement factors are non-controllable and have contributed significantly as MV/BVE declined from 100% to 42% and then rose to 130%. The cumulative impact of non-controllable factors and controllable factors was -55% and +85%, respectively. The net impact was a gain in MV/BVE of 30%. However, it is important to note that the significant monthly changes to MV/BVE were caused by non-controllable factors. These factors were very volatile from Sep 2008 to Jun 2009 due to the credit crisis and illiquidity in the RMBS markets. Similar observations can be made for the trend in DOE charts in Figure 6. The interest rate factor is the most controllable factor and its contribution to changes in DOE was as significant as the contribution of non-controllable factors (namely OAS factor and basis movement factor) to changes in DOE.

Thus, at the time of hedging, it is prudent for a bank to consider these impacts. Impacts due to non-controllable factors without such attribution analysis may lead the bank to incrementally put on a hedge that might be unwarranted. The factor-based attribution analysis provides the ALM manager with a tool to correctly assess impacts due to controllable factors, to communicate ALM risk management strategies to the senior management, the ALCO committee, auditors, regulators, and boards of directors. With this practical analysis, the ALM process will be able to isolate risks due to controllable factors and to hedge the balance sheet for a long-term profitability of the bank and will systematically be able to avoid inefficient and unwanted hedging decisions.

**Conclusion**

The ALM process is very critical to any bank. A practical risk attribution analysis presented in this paper is a tool that facilitates a better understanding of the contribution to changes in MV/BVE from controllable and non-controllable market factors. A simple methodology using a good risk management system provides analytical and quantitative rigor that is essential in understanding the impacts of these market factors. Over the long term, such attribution analysis will enable a bank to hedge proactively by isolating impacts due to controllable factors and accumulate good risk measurement information to assess and develop risk appetite and hedging guidelines. The methodology presented here considers MTM-based risk measurements. Contribution to changes due to controllable factors for MV/BVE and accrual-based income is expected to be correlated and can be tested using income simulation or economic value analysis.

The approach presented here is simple, practical, and intuitively appealing. It provides not only good information for decision making but also clarity and vocabulary for communication within the bank and with other interested parties such as regulators, external auditors, and boards of directors.
References

• Basel Committee on Banking Supervision, 2004, “Principles for the management and supervision of interest rate risk,” BIS
• Comptroller’s handbook, 1998, “Interest rate risk,” OCC
• Federal Deposit Insurance Corporation (FDIC), Risk management manual of examination policies, Part II, Section 7.1 – sensitivity to market risk
• Lawton, P., and T. Jankowski, 2009, Investment performance measurement: evaluating and presenting results (CFA Institute investment perspectives), Wiley

Appendix A -Definitions

DOE (duration of equity) – is a measure of interest rate risk. It estimates the percentage change, expressed in years, in the bank’s market value of equity caused by a parallel shift in the interest rate curve. A DOE of 5 years means that the market value of equity would drop by about 5 percent if the rates shifted up by one percent or 100 basis points. When rates increase, a positive (negative) DOE would lead to a decline (increase) in market value of equity.

To understand DOE pictorially, consider the market value profile in the graph below. The market values in various interest rate scenarios are plotted on the vertical axis with the horizontal axis representing rate shifts or interest rate scenarios. The equity market value profile depicted in the right graph is simply a curve that represents the difference in market value of assets and liabilities at various rate shifts.

The DOE at any interest rate shift is proportional to the slope of the tangent drawn to the equity market value profile. In the right graph, a tangent line is shown for the base case, –200bp, and +200bp shifts. If the tangent line is horizontal, the slope of the line would be zero and the DOE would be zero. A zero DOE means that the bank would not lose market value if the rates moved up or down. In order to make money, however, the bank typically takes risk by targeting a certain level of DOE. The bank controls such risk-taking activity by establishing risk limits.
PART 2

Hedge Funds
Performance Ratios
Adjusted to Market
Liquidity Risk

Pierre Clauss — Associate professor of finance at CREST (Ensai) and CREM (UEB)

Abstract
Market liquidity is complex to measure empirically. This explains why there is no consensus about performance ratios adjusted to its risk. We summarize market liquidity by two major characteristics: a costly one because of the loss of the illiquidity premium, and a profitable one when investors can withdraw when they want. In this paper, three new performance indicators are proposed to integrate, to a certain extent, market liquidity risk, especially for hedge funds investment: liquidity-loss ratio will capture the cost characteristic whereas liquidity-Sharpe ratio and liquidity-profit ratio will represent the profitable alternative. These new ratios try to be simple and precise enough to help investors choose between hedge funds strategies according to their liquidity profile: do they want to capture illiquidity risk premium, or do they want to be free to withdraw?

1 The opinions expressed in this paper are those of the author and are not meant to represent the opinions or official positions of his affiliated organization and research unit. I am grateful to Charles Lacroix, Edmond Lezmi, and Jean Rousselot for helpful and precise comments. The usual disclaimer nonetheless applies and all errors remain mine.
Liquidity: a concept studied again
Since the subprime crisis, liquidity has appeared as the new financial grail. A fundamental concept of financial markets (without liquidity, there would be no markets [Keynes (1936)]), liquidity has been somewhat forgotten in recent years, with the focus shifting to profitability alone.

Historically, capital liquidity became extraordinarily abundant in the 1970s with the increased depth of capital markets and the development of derivatives facilitating the transfer of risk. Forming part of the traditional financial landscape ever since this era, investors paid little attention to liquidity, assuming that it would never be an issue. Unfortunately, this was not the case. It must now be addressed, not just by investors, many of whom now swear by it, but also by regulators: the future Basel agreements will doubtless find a special place for it.

Looking at academic research, we also see this new interest in liquidity issues and the subsequent innovations needed. To illustrate this, some articles in the spring 2010 edition of the Journal of Portfolio Management were rather enlightening:
- Engle, R. F., “How to forecast a crisis” – predicting illiquidity has become an important challenge.
- Lo, A. W., “Survival of the Richest” – it is essential to take into account investors’ copycat behavior in new financial theory.
- Shiller, R. J., “Crisis and innovation” – financial innovation is more than desirable to improve the markets after the recent crisis.
- Golub, B. W., and C. C. Crum, “Risk management lessons worth remembering from thecredit crisis of 2007-2009” – the article focuses on “the supreme importance of liquidity.”

Liquidity risk and hedge funds
At first glance, liquidity is a relatively simple concept: synonymous with fluidity, it shows how easily investors can enter or exit financial markets. It becomes more complicated on closer analysis and particularly if we try to measure it [cf. Clauß (2010), for portfolio selection with liquidity considerations]. Furthermore, if one is looking to invest in hedge funds, liquidity is obviously not going to be the main selling point. Indeed, it is hard to argue that hedge funds are liquid when one takes into account: lock-up periods, where investors cannot withdraw their money when they want and redemptions must take place on specific dates or after minimum periods; leverage, which ties the hedge fund to a prime broker whose liquidity is limited; investment in illiquid but high yielding products; and the domiciliation of hedge funds in tax havens with little transparency.

More precisely, there are two types of liquidity, funding liquidity and market liquidity. The first relates to the possibility for an investor to raise funds: can hedge funds experience mass waves of redemptions? The second relates to investors’ freedom to withdraw from an investment: at first sight, hedge funds’ lock-up periods and illiquidity investment seem to restrict this market liquidity.

It is the second type of liquidity that we will study in greater detail in this paper with the objective of helping investors choose between hedge funds strategies and also single hedge funds.

Measuring hedge funds performance adjusted for market liquidity risk
First of all, performance must be a risk-adjusted measurement. Furthermore, a measurement of risk must be simple and understandable to allow communication between mathematics and investment specialists. It must be conventional. This explains why, at a time when statistical models and probabilities are complex and more accurate, the likes of beta, volatility, Sharpe ratio, and even Gaussian Value-at-Risk models are still as popular, despite their well-documented and generally well-known shortcomings [Shojai and Feiger (2010); Shojai et al. (2010)].

By risk, in this paper, we mean market risk; that is, the risk of investors experiencing changes in the value of their portfolios. For performance, we will calculate the risk-adjusted return on a portfolio: the rationale is that of the Sharpe ratio, defined by Sharpe in 1966 as a “reward for variance” and improved in 1994, or information ratio, two ratios that we are about to transform somewhat to adjust for market liquidity.

In the case of hedge funds, due to a lack of daily data it does not take into account ‘true’ variance but provides a ‘smoothed’ performance pattern. For extreme cases, Sharpe ratios, as in the case of Madoff [Clauss et al. (2009)], could be suspicious. But, it seems to be simple enough and very interesting to use them nevertheless, after a light but essential transformation, to puzzle out the market liquidity risk.

Then, market liquidity is complex to measure empirically. This explains why there is no consensus about performance ratios adjusted to its risk. We will provide a summary of market liquidity with two major characteristics in following sections: a costly one because of the loss of illiquidity premium and a profitable one when investors can withdraw when they want.

Two indicators are put forward specifically for this paper: liquidity-loss or I-loss ratio, developed in the next section, and liquidity-Sharpe or I-Sharpe ratio, developed in the following section, corresponding to each of the two characteristics of market liquidity. We will also develop a third ratio: liquidity-profit or I-profit ratio to help make a comparative synthesis of the two aspects of market liquidity to conclude the paper.
Liquidity cost: l-loss ratio

Theoretical framework

L-loss ratio definition

Liquidity cost relates to the fact that illiquidity can be rewarding. Indeed, investing in an illiquid product is riskier and therefore better compensated by the market. Junk bonds, equity tranches of CDOs made up of subprime securities, and even Greek bonds are examples of financial securities with a high risk premium (measured here by the spread) as they are riskier. Liu (2006) developed an asset valuation model that takes into account the illiquidity risk priced in by the market. Hedge funds are investments that use strategies on relatively illiquid products. They generate absolute performance through, amongst others, the lucrative illiquidity premium.

For a simple measure of outperformance resulting from this illiquidity premium, the idea here is to compare hedge fund strategies with strategies that try to replicate them (or clone them) with the help of highly liquid financial instruments. This is the rationale of the l-loss ratio.

Definition 1: The l-loss ratio is an information ratio calculated between the hedge fund strategy and its clone:

\[
\text{L-loss} = \frac{\text{E}(R_{\text{clone}} - R_{hf})}{\sigma(R_{\text{clone}} - R_{hf})}
\]

with \(R_{\text{clone}}\) being the return of the clone and \(R_{hf}\) the return of the hedge fund strategy. This ratio is determined with annual mean and annual volatility.

This ratio is simple and effective measure of outperformance based on the illiquidity premium. Limitation can appear with the imperfections of the clones.

Principles of hedge funds’ liquid replication

Replication is a fairly recent topic. With hedge fund performance attracting a number of investors, some funds try to replicate this performance using more liquid products, freeing themselves from the constraints of lock-up periods. There are two dominant quantitative families for cloning the performance distributions of hedge funds:

- Factorial replication uses the classic linear Gaussian model to build a portfolio that replicates hedge funds returns [Roncalli and Teilletche (2008)].
- Statistical replication does not clone hedge fund returns month by month, but rather their general distribution over a longer period [Kat and Palaro (2005, 2006), and Hocquard et al. (2007)].

Replication often involves the use of highly liquid instruments such as futures.

More in line with our l-loss ratio objective, we have used factorial replication, following the methodology of Roncalli and Teilletche (2008). The liquid instruments are the following indices (expressed in total returns):

- S&P 500.
- A long/short between the Eurostoxx 50 (hedged in U.S.$) and S&P 500.
- A long/short between the Topix (hedged in U.S.$) and S&P 500.
- The EUR/U.S.$ exchange rate.

To be more reactive and more precise, we will construct replication from a robust method, as in Roncalli and Teilletche (2008): the Kalman filter.

The Kalman filter

The Kalman filter was originally used in industrial and signal processing: radar, photography, radio, computers, and more recently finance. This filtering allows us to eliminate interference inherent to the measurement of observations. For the purposes of our study, the Kalman filter will allow us to obtain a highly accurate calculation of changes in factor exposures. This technique is similar to focusing with a pair of binoculars: the beta determined by the Kalman filter is like a moving image that is no longer blurred.

Let us study the application of Kalman filter modeling in the case of linear regression.

Let \(y_t\) be a temporal series (returns of a hedge fund index in our case study). Let us then assume that this observed variable is correlated with \(p\) explanatory variables (previously described) via regression coefficients, i.e., betas. These are grouped together in what we shall call the state vector. This is denoted \(\beta_t\) with dimension \(p\) and is linked to \(y_t\) via the measurement equation:

\[ y_t = x_t^\prime \beta_t + \epsilon_t. \]

\(\epsilon_t\) is a white noise (meaning that it is a static and not self-correlated process) assumed to be centered Gaussian, with variance-covariance matrix \(R\), assumed to be constant. In the methodology of the Kalman filter, state vector \(\beta_t\) is unobservable. Let us then assume that it is generated by a particular process (that we assume to be first-order Markovian) described by the following transition equation or state equation:

\[ \beta_t = \beta_{t-1} + \eta_t. \]

\(\eta_t\) is a white noise, independent of \(\epsilon_t\), and centered Gaussian with variance-covariance matrix \(Q\), assumed to be constant.

We have thus established the space-state representation on the basis of which we shall determine the Kalman filter, a recursive procedure that will allow us to build an optimal estimator of the unobservable state vector \(\beta_t\) using the information in \(t\).
We must establish a starting point for the filtering. Let us, therefore, assume $\hat{\beta}_{t-1}$ is the known optimal estimator of $\beta_{t-1}$ given the information available in $t-1$. Let $P_{t-1}$ be the variance-covariance matrix of the estimation error between the real value of $\beta_{t-1}$ and its estimate $\hat{\beta}_{t-1}$.

Given $\hat{\beta}_{t-1}$ and $P_{t-1}$, we draw up the following predictive equations:

$$\hat{\beta}_{t|t-1} = \hat{\beta}_{t-1}$$
$$P_{t|t-1} = P_{t-1} + Q$$

When new observation $y_t$ is available, estimator $\hat{\beta}_{t|t-1}$ can be updated. We then obtain the following updated or corrected equations:

$$\hat{\beta}_t = \hat{\beta}_{t|t-1} + K_{t}(y_t - x'_t \hat{\beta}_{t|t-1})$$
$$P_t = (I - K_{t} x'_t)P_{t|t-1}$$

with $K_t = P_{t|t-1}x'_t(x'_t P_{t|t-1}x_t + R)^{-1}$.

Matrix $K_t$ is called the Kalman gain matrix. We will not prove its calculation here based on the minimization of variance-covariance matrix $P_t$ [Harvey (1989)]. When the new estimator in $t$ is determined with the corrected equations, we repeat the process calculating the new estimator of predictive equations, which will enable us to establish gain $K_{t+1}$ and thus determine the estimator of corrected equations.

The following difficulty is the estimation, using the maximum likelihood method, of parameters $Q$ and $R$ [see Harvey (1989) for more details].

**Empirical results**

Each month, we regress the excess-returns of HFRI indices on the excess-returns of the liquid products mentioned above. For greater accuracy, we use the Kalman filter to estimate positions at month $t$ in liquid products that would allow us to replicate the hedge fund strategy in month $t+1$. This inevitably means there is a one-month lag, which is offset by the responsiveness of our position estimators. We can finally establish the $l$-loss ratio as the information ratio between the clone and the hedge fund strategy.

In Figures 1 to 3 and Table 1, we tested three different periods and noted that in the two financial bubbles studied, liquidity did have a cost as $l$-loss ratios are all negative. However, during a downturn such as the one experienced during the subprime crisis, liquidity would have been significantly profitable, except for the following two strategies: macro and emerging markets.

<table>
<thead>
<tr>
<th>Period</th>
<th>HFRI (Global index)</th>
<th>Equity hedge</th>
<th>Event-driven</th>
<th>Macro</th>
<th>Relative value</th>
<th>Fund of funds</th>
<th>Emerging markets</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 1998 – March 2003</td>
<td>-1.40</td>
<td>-1.34</td>
<td>-0.93</td>
<td>-0.40</td>
<td>-1.15</td>
<td>-0.18</td>
<td>-0.45</td>
</tr>
<tr>
<td>March 2003 – February 2009</td>
<td>-0.60</td>
<td>-0.18</td>
<td>-0.80</td>
<td>-0.66</td>
<td>-0.73</td>
<td>-0.21</td>
<td>-1.05</td>
</tr>
<tr>
<td>October 2007 – February 2009</td>
<td>0.29</td>
<td>0.71</td>
<td>0.84</td>
<td>-0.77</td>
<td>0.16</td>
<td>0.63</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Table 1 – L-loss ratios during 3 different periods
Liquidity profit: I-Sharpe ratio

Theoretical framework

L-Sharpe ratio definition

After the liquidity-cost, we will try to estimate the liquidity-profit. As we have already seen, investing in a hedge fund is relatively complicated, mainly because of lock-up constraints. We will try to estimate the gain that would result from investors being free to withdraw from and invest in a hedge fund whenever they want. In reality, it is not possible but we will act as if the investor could.

We have to establish a signal for when to enter and exit a hedge fund: we base this on price momentum (see below). We will determine this signal in the following way: if the hedge fund strategy has performed positively for six months, investors decide to enter; conversely, if it has had a negative return for six months, investors decide to exit and invest in risk-free assets. This signal based on a stop-loss criterion can reflect realistically the behavior of a rational investor who wants to withdraw. Although imperfect, this type of signal based on 6-month momentum does allow us to take into account persistence and minimize the risk of saloon door problems (entering just before a loss; exiting just before a gain). Lastly, turnover is reduced to an average of 10% (we do not take transaction costs into account in the ratio). With the help of this new active strategy, we determine a new Sharpe ratio that we call the liquidity-Sharpe ratio or simply the I-Sharpe ratio.

Definition 2: The I-Sharpe ratio is determined as a Sharpe ratio of an active strategy following an adequate signal from price momentum:

\[ \text{L-} \text{Sharpe} = \frac{\mathbb{E}[R_{\text{active}} - r_t]}{\sigma(R_{\text{active}} - r_t)} \]

with \( R_{\text{active}} \) being the return of the active strategy and \( r_t \) the risk-free rate. This ratio is determined with annual mean and annual volatility.

This ratio is a simple and effective measure of the gain that would result from investors being free to withdraw from and invest in a hedge fund whenever they want. Limitations can appear with the imperfection of the signal from the active strategy and the non-consideration of transaction costs.

Hedge funds momentum

In finance, momentum refers to price trends (in Latin it means movement). For example, six-month momentum is determined by a portfolio’s return over the previous six months.

In 1993, Jegadeesh and Titman showed the profitability of a strategy of buying U.S. equities for which returns over the previous six months had been positive and selling stocks for which returns over the previous six months had been negative. This profitability could not be explained by traditional systematic risk factors: market portfolio (CAPM), size factor, and valuation factor [Fama and French (1993)]. Carhart (1997), therefore, developed a model taking this new risk factor – momentum – into account. Today, momentum is well documented and accepted by the financial community. Hong and Stein (1999) put forward one appealing explanation. They explained it by the fact that economic agents have not completely integrated the available information and may take more than six months to adjust. Agents can underreact to the spread of information, creating this price momentum. For example, good news (better-than-expected profits) about a stock that is off the radar will take time to be factored in. Accordingly, the price will rise as and when agents on the financial markets discover this news and want to buy the stock.

Does this momentum, originally visible on the equity markets, apply to hedge fund strategies? Or, in other words, is there a persistence of hedge fund performance? Most literature on this controversial subject argues for a persistence of hedge fund performance for between three and six months [Agarwal and Naik (2000), Capocci et al. (2005), and Eling (2009)]. Of course, results vary according to how persistent the performance of the strategies is.
Empirical results
We will measure l-Sharpe ratio for different hedge fund strategies as well as for the equity market over the two bubble periods studied for the l-loss ratio and compare it with the classical Sharpe ratio determined on the basis of passive strategy (buy-and-hold or constant mix) on HFRI indices, i.e., without entry or exit signals.

In Figures 4 and 5 and Table 2, we note that equities indices (CAC 40, S&P 500, and NASDAQ) take advantage of liquidity freedom since l-Sharpe ratios are always better than Sharpe ratios.

For the hedge funds universe, it is less obvious. We note several points:
- During the technology bubble, liquidity benefits would be relatively mild for all hedge funds strategies and even slightly negative for Macro and Event-driven strategies. We can then conclude that investing in hedge funds does not require freedom to come and go during this period.
- This was not the case with the subprime bubble: the liquidity would be highly profitable for all strategies except Macro, which could have done without it again. The Relative value strategy is affected the most. Certainly the intensive use of leverage, but also the ban on short selling, weighed heavily on the performance of most strategies. It is worth remembering that the recent crisis is primarily a liquidity crisis.

Concluding remarks
To conclude, we will develop a third ratio: liquidity-profit or l-profit ratio to help make a comparative synthesis of the two aspects of market liquidity: the cost one and the profitable one. Like the l-loss ratio, this is an information ratio but this time between a passive strategy, i.e., without entry or exit signals, and an active and liquid strategy (used to determine the l-Sharpe ratio).

Definition 3: The l-profit ratio is an information ratio between the active strategy following the momentum signal and the concurrent passive strategy:

\[ \text{L-profit} = \text{E}[R_{\text{active}} - R_{\text{passive}}] + \sigma R_{\text{active}} - R_{\text{passive}} \]

with \( R_{\text{active}} \) being the return of the active strategy and \( R_{\text{passive}} \) being the return of the passive one. This ratio is determined with annual mean and annual volatility.

In Figures 6 and 7 and Table 3, we have studied again the two different bubbles with two different sets of liquidity conditions. During the technology bubble, the illiquidity premium could be more attractive (negative l-loss ratio) than the gain in investment freedom (l-profit ratio negative or lower in absolute value terms) particularly for Macro, Relative value and Event-driven strategies; Funds of funds are neutral between the two ratios. On the contrary, during the recent bubble, the spread between the two ratios is restabilized: the liquidity-profit has been more important especially for Equity hedge and Funds of funds. The Macro strategy alone comes out of it well.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HFRI (global index)</td>
<td>0.71 0.91 0.39 1.27</td>
</tr>
<tr>
<td>Equity hedge</td>
<td>0.73 0.83 0.11 1.25</td>
</tr>
<tr>
<td>Event-driven</td>
<td>0.83 0.59 0.43 1.61</td>
</tr>
<tr>
<td>Macro</td>
<td>0.56 0.30 1.09 0.82</td>
</tr>
<tr>
<td>Relative value</td>
<td>2.47 2.57 0.08 1.69</td>
</tr>
<tr>
<td>Fund of funds</td>
<td>0.42 0.63 0.03 0.73</td>
</tr>
<tr>
<td>Emerging markets</td>
<td>0.42 0.56 0.61 1.59</td>
</tr>
<tr>
<td>S&amp;P 500</td>
<td>-0.37 -0.21 -0.38 0.70</td>
</tr>
<tr>
<td>NASDAQ</td>
<td>-0.17 0.41 -0.15 0.18</td>
</tr>
<tr>
<td>CAC 40</td>
<td>-0.44 0.18 -0.15 0.84</td>
</tr>
</tbody>
</table>

Table 2 – L-Sharpe ratios and Sharpe ratios during 2 different periods

Figure 6 – L-loss ratios versus l-profit ratios during technology bubble (August 1998 – March 2003)

Figure 7 – L-loss ratios versus l-profit ratios during subprime bubble (March 2003 – February 2009)
The three new ratios developed in this paper (\(l\)-loss, \(l\)-Sharpe and \(l\)-profit ratios) can be very interesting in helping with allocation decisions. Then, empirical results seem to suggest that Macro strategy funds in the universe of hedge funds tend not to be really interested in dealing with liquidity freedom but to be interested in taking advantage of illiquid premium in general, even during subprime crisis.

Further works are to be made in specifying these ratios and their statistical properties. We think about replication methods and relations between hedge funds strategies’ persistence and \(l\)-Sharpe ratio values.

### References

- Keynes, J. M., 1936, *General theory of employment, interest and money*, Macmillan
Regulating Credit Ratings Agencies: Where To Now?

Amadou N. R. Sy — Deputy Division Chief, IMF

Abstract
This paper reviews current proposals to regulate credit rating agencies. The proposals can be classified in two broad areas: (1) micro- and (2) macro-prudential measures. While the previous regulations relied on micro-prudential measures, experience from the U.S. subprime crisis and the European sovereign debt crisis shows that such an approach did not address the negative externalities from credit ratings, prompting the need to expand the regulatory perimeter. This paper highlights two types of macro-prudential regulation: (i) those that attempt to take credit ratings out of regulation such as Section 939A of the 2010 Dodd-Frank Act and (ii) those that aim to reduce the importance of credit ratings in regulation such as the recent changes to the ECB collateral rules. This paper argues, however, that it remains important that policymakers conduct formal assessments of the impact of the use of credit ratings on financial markets. This can be done through stress tests of institutions that would be affected by rating downgrades. Additional capital requirements and/or liquidity buffers could be used if necessary to mitigate the systemic risk of credit ratings.

1 The views expressed herein are those of the author and should not be attributed to the IMF, its Executive Board, or its management. The author thanks Ana Carvajal, Nadege Jassaud, and Kate Shields for useful comments. All errors or omissions remain mine.
Typically, micro-prudential regulation is the main approach for regulating credit rating agencies (CRAs). This type of regulation is concerned with issues such as abuse of monopoly power, consumer protection, and asymmetric information. The oligopolistic nature of the CRA industry, concerns about conflicts of interest in their ‘issuer-pays’ business model, and massive losses following the abrupt downgrades of structured products during the U.S. subprime crisis provide convincing arguments for the micro-prudential regulation of CRAs. From a political economy view, this type of prudential regulation is appealing to politicians as it gives a clear signal that quick and strong action is being taken to address voters’ concerns.

However, recent and current developments in global financial markets have highlighted the need to also address externalities endogenous to financial markets. Credit rating downgrades can have important spillover effects and contribute to systemic risk. Indeed, recent events in the euro area illustrate how rating downgrades can (i) increase borrowing costs and reduce access to capital markets, (ii) disrupt the functioning of money markets as ratings-based collateral rules are triggered, and (iii) have spillover effects across countries and asset classes. The U.S. and European crises show that downgrades can also lead to (iv) a drying up of liquidity, (v) collateral calls through ratings-based triggers such as those in credit default swaps (CDS); and (vi) mark-to-market losses through ratings-based institutional investment mandates such as money markets funds.

CRAs regulation should, therefore, seek to avoid externalities. Financial stability national authorities and international bodies have identified existing gaps in the micro-prudential regulation of CRAs and are seeking to expand and refine the regulatory perimeter to address systemic risks. Current efforts are focused on (i) increasing CRAs transparency and (ii) reducing reliance on CRAs. While increased transparency can be seen as a part of micro-prudential regulation, efforts to reduce the reliance on CRAs take a broader, more systemic perspective. For instance, one of the goals of the 2010 Dodd-Frank Act is to remove any reference to, or requirements of reliance on, credit ratings in regulation and substitute in their place other standards of credit-worthiness. Similarly, the ECB collateral framework attempts to reduce the regulatory reliance on credit ratings.

However, current regulatory efforts may prove to be insufficient to manage all the types of risks stemming from CRAs. Indeed, credit ratings are literally everywhere and in particular in private contracts, such as credit default swaps (CDS) and the investment mandates of institutional investors. It is, therefore, important for risk managers and senior management and boards of institutional investors as well as regulators to fully understand the type of risk credit ratings attempt to capture, the methodology and data credit rating agencies use, and the potential systemic effects of credit ratings. Policymakers should take a systemic approach to the regulation of CRAs, which would require them to also take into account the effects of credit ratings in private sector contracts.

Policymakers should assess the possible impact of credit ratings on financial markets as a whole, across different asset classes and instruments. This can be done, for instance, through stress tests of institutions that would be affected by rating downgrades and as a result precautionary measures such as requiring increased capital or liquidity buffers could be used to avoid the possible negative effects associated with credit downgrades. This approach would be consistent with the objective of macro-stress tests in Pillar 2.

The micro-prudential regulation of CRAs

In August 2007, the U.S. Securities and Exchange Commission (SEC) initiated in-depth examinations of the three major rating agencies (Moody’s, Standard and Poor’s, and Fitch) and over a ten-month investigation uncovered significant deficiencies in the rating agencies’ policies, procedures, and practices.

The examinations found that:

- The CRAs struggled significantly with the increase in the number and complexity of subprime residential mortgage-backed securities (RMBS) and collateralized debt obligations (CDO) deals since 2002.
- None of the CRAs examined had specific, comprehensive, written procedures for rating RMBS and CDOs.
- Significant aspects of the rating process were not disclosed or even documented by the firms.
- Conflicts of interest were not always managed appropriately.
- Processes for monitoring ratings were less robust than the processes used for initial ratings.

The theoretical literature stresses the importance of restoring the ‘investor-pays’ business model. Most analytical studies on the role of CRAs in the current crisis stress that the current model under which issuers of securities pay CRAs to rate their securities gives rise to (1) conflicts of interest, (2) perverse effects of ‘shopping’ for rating, and (3) issues related to the quality of disclosed information.2 As a result, the academic consensus is that the ratings industry must be regulated to address agencies’ fundamental conflicts of interest. In addition, shopping for ratings should be banned to reduce the conflict of interest of issuers. This could be achieved through a return to the ‘investor-pays’ system which would replace the current ‘issuer-pays’ business model.

The academic literature offers a number of recommendations on how to better address these micro-prudential considerations. For instance, Calomiris (2009) argues that an ‘investor-pays’ system would lead to inflated...

---

2 See Freixas and Shapiro (2009) for a concise review of this literature which includes Bolton et al. (2008), Mathis et al. (2008), Pagano and Volpin (2008), and Sivrida and Veldkamp (2008). See also Bennmelech and Dlugosz (2009) who find evidence that ratings shopping may have played a role in the current crisis.
ratings. This is because buy-side investors reward rating agencies for underestimating risk as high ratings loosen regulatory restrictions on the types of instruments they can invest in.\(^3\) The literature also suggests the establishment of centralized clearing platforms for ratings. Richardson and White (2009) argue that there is a free rider problem in the ‘investor pays’ model which competition may not solve. They, therefore, recommend the creation of a centralized clearing platform for rating agencies within the SEC. In this scheme, the platform would assess a flat fee for the rating of a security, depending on its attributes. It would also choose a rating agency from a sample of approved CRAs, which would then rate the security. Mathis et al. (2008) also suggest creating a platform that would take payments from issuers and assign securities to one or more CRAs, which would then rate them. They argue that this scheme would eliminate conflicts of interest and ‘shopping’ for rating.

A recent review of regulatory developments in the U.S., Europe, Japan, and Australia [IMF (2010)] shows that most proposals to regulate credit rating agencies are of a micro-prudential nature. This type of regulation is concerned with issues of abuse of monopoly power, consumer protection, and micro-manifestations of asymmetric information as illustrated above. For instance, the IMF review notes that the 2009 E.U. regulation introduced mandatory registration for all CRAs operating in the E.U..\(^4\) Registered CRAs will have to comply with a comprehensive set of rules to make sure that ratings are not affected by conflicts of interest; that CRAs remain vigilant, ensuring the quality of the rating methodology; and that they act in a transparent manner. The regulation also includes a surveillance regime for CRAs. In particular, CRAs:

- May not provide advisory services.
- Will not be allowed to rate financial instruments if they do not have sufficient quality information on which to base their ratings.
- Must disclose the models, methodologies, and key assumptions on which they base their ratings.
- Must differentiate the ratings of more complex products by adding a specific symbol.
- Should have at least two independent directors on their boards whose remuneration cannot depend on the business performance of the rating agency.

According to the regulation, the Committee of European Securities Regulators will be in charge of the registration and day-to-day supervision of the CRAs. However, in June 2010 the EC proposed the introduction of centralized E.U. oversight of CRAs, entrusting the proposed new European Securities and Market Authority (ESMA) with exclusive supervisory powers over CRAs registered in the E.U., making CRAs the first type of institution subject to centralized E.U. supervision. Under the proposal, the ESMA will have powers to request information, launch investigations, and perform on-site inspections. Furthermore, issuers of structured finance products will have to provide all other interested CRAs with access to the information they give to the CRA rating their product, enabling the other CRAs to issue unsolicited ratings.

Similarly, the IMF review notes that the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act increases internal controls for CRAs, requires greater transparency of rating procedures and methodologies, and provides the SEC with greater enforcement and examination tools regarding NRSROs. In particular, the bill:

- Requires each NRSRO to have a board of directors of which at least half (but not fewer than two) are independent members, some of whom must be users of NRSRO ratings.
- Introduces the possibility of exposing NRSROs to liability as experts.
- Suggests that the SEC should exercise its rulemaking authority to prevent conflict of interest arising from employees of NRSROs providing services to issuers of securities that are unrelated to the issuance of credit ratings.
- Requires each NRSRO to establish, maintain, enforce, and document an internal control structure to govern implementation of and adherence to policies, procedures, and methodologies for determining ratings.
- Asks the SEC to adopt rules that require each NRSRO to establish, maintain, and enforce policies and procedures that clearly define and disclose the meaning of any ratings symbol and apply this symbol consistently for all instruments for which the symbol is used.
- Establishes an SEC ‘office of credit ratings’ that will put in place fines and other penalties for violations by NRSROs, administer SEC rules with respect to NRSRO practices in determining ratings, and conduct an annual examination of each NRSRO.

The bill also asks for a number of studies. In particular, the SEC is required to undertake a study of the credit rating process for structured finance products and the conflicts of interest associated with the issuer-pay and subscriber-pay models, the range of metrics to determine the accuracy of ratings, and alternative means of compensation to create incentives for accurate ratings. The SEC must also study the feasibility of establishing an independent organization to assign NRSROs to determine credit ratings for structured finance products, and create and oversee a Credit Rating Agency Board that would assign a ‘qualified’ CRA to rate each new issue of asset-backed securities, unless it determines that an alternative system would be more appropriate. The SEC is also asked

---

\(^3\) Calomiris (2009) recommends that NRSROs provide specific estimates of the probability of default and the loss-given-default for any rated instruments. Regulators would then penalize NRSROs that systematically underestimate risk with a six-month ‘sit out’ during which their ratings would not be used for regulatory purposes. Such reduced demand for their ratings would affect their fee income, thereby giving them an incentive to correctly estimate risk.

\(^4\) Specific treatment can be extended on a case-by-case basis to CRAs operating exclusively from non-E.U. jurisdictions provided that their countries of origin have established regulatory and supervisory frameworks as stringent as the one now put in place in the E.U.
to provide a study of the independence of NRSROs and how this affects ratings issued, while the Government Accountability Office must conduct a study of alternative means for compensating CRAs in order to create incentives to provide more accurate ratings. However, one section of the Dodd-Frank bill broaden the regulatory perimeter for CRAs. Indeed, Section 939 A of the Dodd-Frank bill requires the removal of certain statutory references to credit ratings and requires that all federal agencies review and modify regulations to remove references to or reliance upon credit ratings and substitute an alternative standard of creditworthiness.

Broadening the regulatory perimeter of CRAs to build a macro-prudential framework

Use of ratings in legislation, regulations and supervisory policies (LRSPs)

A recent international stocktaking exercise conducted by the Joint Forum (2009) reveals that credit ratings are generally used in member jurisdictions for five key purposes, especially in their LRSPs covering the banking and securities sectors: (i) determining capital requirements; (ii) identifying or classifying assets, usually in the context of eligible investments or permissible asset concentrations; (iii) providing a credible evaluation of the credit risk associated with assets purchased as part of a securitization offering or a covered bond offering; (iv) determining disclosure requirements; and (v) determining prospectus eligibility. A key finding of the Joint Forum (2009) exercise is that no member authority had conducted a formal assessment of the impact of the use of credit ratings in LRSPs on investor behavior.

In the U.S., the SEC started using ratings by NRSROs in 1975 to determine capital charges for broker-dealers. The term ‘NRSRO’ which stands for nationally recognized statistical rating organizations, was originally adopted by the U.S. SEC that year solely for determining capital charges on different grades of debt securities under the Net Capital Rule. The rule allowed broker-dealers to apply lower ‘haircuts’ to debt securities that were rated investment grade by a NRSRO. Partnoy (2009) notes that private reliance on ratings has typically followed its public use. This phenomenon predates the 1975 rule and traces its origins to the aftermath of 1929 Crash.

The regulatory use of ratings expanded quickly to other segments of the financial markets. SEC (2003) illustrates how issuers of commercial paper find it difficult to sell paper that does not qualify for investment by money market funds under Rule 2a-7 of the Investment Company Act (1940) limits money market funds to investing in only high quality short-term instruments, and NRSRO ratings are used as benchmarks for establishing minimum quality investment standards. It also notes that most money market funds voluntarily limit themselves to investing in securities rated higher than necessary to be eligible under Rule 2a-7.

Policymakers’ reliance on credit ratings is even illustrated in the resolution of the current crisis. Indeed, the U.S. government continues to rely on AAA ratings as illustrated by their use in the Term Asset-Backed Securities Loan Facility (TALF), established in November 2008. Indeed, the U.S. authorities will allow use of the TALF only for the purchase of AAA-issues [Ng and Rappaport (2009)].

Under the 2005 Basel Committee on Banking Supervision (BCBS) new capital adequacy framework (Basel II), banks can use ratings assigned by a recognized CRAs in determining credit risk weights for many of their institutional credit exposures. The objective of Basel II, Pillar 1 is to align a bank’s minimum capital requirements more closely to its risk of economic loss. To do so, a bank capital is made more sensitive to such a risk by requiring higher (lower) levels of capital for those borrowers with higher (lower) credit risk, and vice versa. Under the ‘standardized approach,’ any bank ‘may’ use external measures of credit risk to assess the credit quality of its borrowers for regulatory capital purposes. In 2009, the Basel Committee revised its risk-based capital framework so as to strengthen it. For instance, it introduced operational criteria to require banks to undertake independent analyses of the creditworthiness of their securitization exposures.

Over time, marketplace and regulatory reliance on credit ratings has increased to the point where ratings are widely used for distinguishing among grades of creditworthiness. The regulatory use of credit ratings may have increased the demand for highly rated products, in particular those issued by off-balance sheet entities. It may have also reduced incentives for investors to conduct appropriate due-diligence on the quality of their investments and manage risks adequately. As a result, some argue that policymakers should consider withdrawing financial regulation that imposes the use of ratings while others stress that they should recognize the

---

5 The Joint Forum (2009) received a total of 17 surveys from member authorities, representing 26 separate agencies from 12 different countries and five responses, describing international frameworks.

6 The NCR requires broker-dealers, when computing net capital, to deduct from their net worth certain percentages of the market value of their proprietary securities positions. SEC (2003) notes that the Commission determined that it was appropriate to apply a lower haircut to securities held by a broker-dealer that were rated investment grade by a credit rating agency of national repute, because those securities were typically more liquid and less volatile in price than securities that were not so highly rated. A primary purpose of these ‘haircuts’ is to provide a margin of safety against losses that might be incurred by broker-dealers as a result of market fluctuations in the prices of, or lack of liquidity in, their proprietary positions. The requirement that the credit rating agency be ‘nationally recognized’ was designed to ensure that its ratings were credible and reasonably relied upon by the marketplace.

7 See SEC (2003) for more regulatory use of ratings, including in a wide range of financial legislation at the federal, state, and foreign laws and regulations such as the definition of ‘mortgage related security,’ institutions that wish to participate in student financial assistance programs, or appropriate investment for insurance companies.

8 The more advanced (i.e., foundation and advanced internal ratings based) approaches may be used only when the bank satisfies the supervisor that it meets the requisite higher standards.

9 The Joint Forum (2009) notes that respondents to its survey were split as to whether their use of credit ratings and/or reference to CRAs has had the effect of implying an endorsement of such ratings and/or agencies, although a slight majority answered in the affirmative.
limits of regulation. Richardson and White (2009) suggest that one policy option would be to allow regulated financial institutions to take advice from any source that they consider to be most reliable. Financial institutions would, however, justify their choice of advisor to their regulator. They conjecture that this would open the advisory information market to new ideas and new entry. In contrast, the Turner Review (2009) notes that factors other than regulation may have a bigger influence on the use of ratings and on the extent to which they are procyclical. These include investor wariness, especially with instruments such as complex CDO structure, CDO2, higher capital requirements for trading books, countercyclical macro-prudential policies relating to capital, accounting, and liquidity.

FSB (2010a) reviews progress in the implementation of the G20 recommendations for strengthening financial stability, including those related to CRAs. The report shows an important shift from the sole reliance on micro-prudential regulation to one where the regulator perimeter is being expanded and refined. The international community has now realized the need for (i) increased transparency of CRAs, and (ii) reducing reliance on CRAs.

Regarding transparency – a key objective of micro-prudential regulation – FSB (2010a) notes that the revised IOSCO “Code of conduct fundamentals for credit rating agencies” (IOSCO code of conduct) has been substantially implemented by the major rating agencies. It also reviews national efforts in the U.S., E.U., Japan, and Canada to strengthen oversight of CRAs.

Beyond micro-prudential regulation, FSB (2010b) presents 14 “principles for reducing reliance on CRA ratings,” which are of a macro-prudential nature. The rationale for these principles is to increase the resilience of the financial system by reducing herding and cliff-effects that arise from the rating thresholds that are present in laws, regulations, and market practices. The principles are quite exhaustive and cover the reliance on CRA ratings in standards, laws, and regulation, their use by banks, market participants, and institutional investors, as well as in central bank operations and bank supervision. At the national level, the Dodd-Frank Act in the U.S. is a useful example of attempts to go beyond micro-prudential regulation.

Removing ratings from regulation: Section 939A of the 2010 Dodd Frank Act

The U.S. regulatory agencies include various references to and requirements based on the use of credit ratings issued by NRSROs. In particular, Section 939A of the 2010 Dodd-Frank Act (July 21, 2010) requires them to review and modify their regulations to remove any reference to, or requirements of reliance on, credit ratings. An important implication is that U.S. regulatory agencies will have to rely on standards of creditworthiness other than credit ratings.

A recent advanced notice of proposed rulemaking (ANPR, U.S. Treasury, 2010) describes the areas in the U.S. regulatory agencies’ risk-based capital standards for federal banks and Basel changes that could affect those standards that make reference to credit ratings and requests comment on potential alternatives to the use of credit ratings. Although the use of credit ratings in regulation is much broader than in the setting up of risk-based capital standards, this is an important area of regulation as capital offers banks a cushion against unexpected risks. In addition, risk-based capital offers a useful area to benchmark developments in the U.S. with global regulatory efforts, such as the Basel standardized approach for credit risk which relies extensively on credit ratings to assign risk weights for various exposures. As noted in the ANPR, implementation in the U.S. of the changes to the Basel Accord would be significantly affected by the need for the agencies to comply with Section 939A of the Dodd-Frank Act (Table 1 below compares the use of credit ratings in the U.S. and Basel regulation on risk-based capital).

The ANPR flags possible alternatives to credit ratings in the risk-based capital standards which fall in two broad categories, as risk weights can be either based on (i) an exposure to a particular category – such as to a sovereign, public sector entity, bank, corporate, securitization, and also credit risk mitigation – or (ii) a specific exposure.

Under the first approach, the determination of risk weights based on exposure category would drop references to credit ratings. Instead, non-securitization exposures generally would receive a 100 percent risk-weight unless otherwise specified. This approach would, however, allow for some flexibility to provide a wider range of risk-weights and increase the risk sensitivity of the risk-based capital requirements. For instance, some sovereign and bank exposures could be assigned a zero or 20 percent risk weight, respectively. The U.S. regulatory agencies could also consider the type of obligor, for example sovereign bank, public sector entities, as well as other criteria such as the characteristics of the exposure.

In contrast, the second approach could assign risk weights to individual exposures using specific qualitative and quantitative credit risk measurement standards established by the U.S. regulatory agencies for various exposure categories. Such standards would be based on broad credit-worthiness metrics. For instance, exposures could be assigned a risk weight based on certain market-based measures, such as credit spreads, obligor-specific financial data, such as debt-to-equity ratios, or other sound underwriting criteria. Alternatively, banks could assign exposures

10 See for instance Partnoy (2006). See also SEC (2009) and the April 2009 SEC “Roundtable to examine oversight of credit rating agencies” for current views regarding the oversight of CRAs.
11 U.S. regulators in this paper include the Office of the Comptroller of the Currency (OCC), the Board of Governors of the Federal Reserve System (FRB), the Federal Deposit Insurance Corporation (FDIC), and the Office of Thrift Supervision (OTS).
12 This is comparable to the original Basel 1 approach.
13 Market-based measures were debated and discarded by the Basel Committee when devising Basel 2.
to one of a limited number of risk weights categories based on an assessment of the exposure’s probability of default or expected loss.\textsuperscript{14}

The ANPR suggests different ways that could be used to assign specific risk weights. For instance, banks could be permitted to contract with third-party service providers to obtain quantitative data, such as probabilities of default, as part of their process for making credit-worthiness determinations and assigning risk weights. An alternative to third-party service providers would be the approach used by the National Association of Insurance Commissioners (NAIC), under which a third-party financial assessor would inform the U.S. regulatory agencies’ understanding of risks and their ultimate determination of the risk-based capital requirements for individual securities. The ANPR recognizes, however, that the use of third-party service providers involves trade-offs between risk sensitivity and consistency, while that of a third-party financial assessor may lead to an excessive reliance on a single third-party assessment of risk.

### Reducing the reliance on ratings: the ECB collateral framework

A consequence of Eurozone sovereign downgrades, which is typically overlooked relates to collateral rules in the eurozone money markets, which are based on credit ratings. Although such collateral rules do not exist in Asia and the U.S., they provide an interesting illustration of the use of credit ratings and their systemic importance. Gyntelberg and Hördahl (2010) note that one concern in the aftermath of the initial downgrade of Greek sovereign debt was that Hellenic banks – which according to rating agencies and analysts, depended more on ECB funding than institutions in other countries did – would not be able to post Greek government bonds as collateral in the ECB’s refinancing operations. The possible loss of this funding source for Greek banks pushed up CDS premia and yield spreads on Greek government debt even further, as it increased the perceived financial risks of holding government bonds. Data from Blundell-Wignall and Slovik (2010) illustrates the exposure of European banks to Greek sovereign credit risk (Table 2).

### Under the Eurosystem Credit Assessment Framework (ECAF), there are specific rules governing the quality of the Government bonds that banks can use as collateral in exchange for funding. To be eligible for collateral, securities have to be assigned a credit rating above a preset minimum of BBB-. As a result, banks cannot obtain funding for collateral with a rating lower than the minimum. In contrast, the higher the rating, the lower the haircut banks will pay.

Following S&P’s downgrade of Greece to BB+ on April 27, 2010, one notch below BBB-, and the announcement by Moody’s that it may downgrade further Greece’s A3 rating, the ECB indefinitely suspended its minimum credit rating threshold in early May. This decision helped banks in Greece and other eurozone countries retain the use of Greek government bonds as collateral to obtain funding. It also helped avoid systemic consequences in the eurozone money markets, as such markets are typically the cornerstone of any financial markets and a key source of interbank

\textsuperscript{14} This is a Basel 2 methodology.
liquidity. Previously, on March 25, 2010, the European Central Bank announced that it would delay its plans to raise by the end of this year the minimum credit rating required for assets provided as collateral by eurozone banks (from BBB- to A-). The ECB’s decision not to tighten collateral rules reduced the risk of Greek as well as other Eurozone banks becoming ineligible to use their holdings of Greek government securities as collateral if the country’s sovereign credit rating was downgraded below the minimum required by Eurozone rules.

The use of credit ratings can, however, be made more flexible when the level of haircut required for different rating grades do vary. Indeed, the ECB changed its collateral framework in April 2010 by announcing graduated valuation haircuts for lower-rated assets (to be implemented as of 1 January 2011). While keeping the minimum credit threshold at investment-grade level (i.e. BBB-/Baa3) beyond the end of 2010, except for asset-backed securities (ABSs), the ECB will, as of 1 January 2011, use a schedule of graduated valuation haircuts to the assets rated in the BBB+ to BBB- range (or equivalent). This graduated haircut schedule will replace the uniform haircut add-on of 5 percent that is currently applied to these assets.

In particular, the ECB announced that the detailed haircut schedule will be based on the following parameters (Table 3):

- The new haircuts will be duly graduated according to differences across maturities, liquidity categories, and the credit quality of the assets concerned. The lowest haircuts will apply to the most liquid assets with the shortest maturities, while the highest haircuts will apply to the least liquid assets with the longest maturities.
- The new haircuts will be at least as high as the haircut currently applied, which is a flat 5 percent add-on for the assets concerned over the haircut that would apply to similar assets with a higher credit quality.
- No changes will be made to the current haircut schedule foreseen for central government debt instruments and possible debt instruments issued by central banks that are rated in the above-mentioned range.
- The new haircuts will not imply an undue decrease in the collateral available to counterparties.

In addition to varying the level of haircut, the ECB also restricted a number of assets from being eligible as collateral (as of January 1, 2011). These include marketable debt instruments denominated in currencies other than the euro, such as the U.S. dollar, the pound sterling, and the Japanese yen, and issued in the Euro Area; debt instruments issued by credit institutions, which are traded on the accepted non-regulated markets; and subordinated debt instruments when they are protected by an acceptable guarantee.

Table 3 – Haircut schedule for assets eligible for use as collateral in Eurosystem market operations (in percent)

<table>
<thead>
<tr>
<th>Credit quality</th>
<th>Liquidity categories</th>
<th>Residual</th>
<th>Category I</th>
<th>Category II</th>
<th>Category III</th>
<th>Category IV</th>
<th>Category V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>maturity (years)</td>
<td>fixed coupon</td>
<td>zero coupon</td>
<td>fixed coupon</td>
<td>zero coupon</td>
<td>fixed coupon</td>
<td>zero coupon</td>
</tr>
<tr>
<td>Steps 1 and 2</td>
<td>(AAA to A-)</td>
<td>0-1</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>1.5</td>
<td>1.5</td>
<td>2.5</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-5</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-7</td>
<td>3.0</td>
<td>3.5</td>
<td>4.5</td>
<td>5.0</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-10</td>
<td>4.0</td>
<td>4.5</td>
<td>5.5</td>
<td>6.5</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;10</td>
<td>5.5</td>
<td>8.5</td>
<td>7.5</td>
<td>12.0</td>
<td>11.0</td>
</tr>
<tr>
<td>Source: ECB</td>
<td>Steps 3 (BBB+ to BBB-)</td>
<td>0-1</td>
<td>5.5</td>
<td>5.5</td>
<td>6.0</td>
<td>6.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-3</td>
<td>6.5</td>
<td>6.5</td>
<td>10.5</td>
<td>11.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-5</td>
<td>7.5</td>
<td>8.0</td>
<td>15.5</td>
<td>17.0</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-7</td>
<td>8.0</td>
<td>8.5</td>
<td>18.0</td>
<td>20.5</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-10</td>
<td>9.0</td>
<td>9.5</td>
<td>19.5</td>
<td>22.5</td>
<td>29.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;10</td>
<td>10.5</td>
<td>13.5</td>
<td>20.0</td>
<td>29.0</td>
<td>29.5</td>
</tr>
</tbody>
</table>
Credit ratings in private contracts

Previous crises have shown the central role that credit ratings play in the investment decisions of financial market participants. SEC (2003) provides a good review of the role of rating agencies in securities markets. Issuers of securities seek credit ratings to improve the marketability or pricing of their securities, or to satisfy investors, lenders, or counterparties who want to enhance management responsibility.

Buy-side firms, such as mutual funds, pension funds, and insurance companies, are substantial users of credit ratings, even though they claim to typically conduct their own credit analysis for risk management purposes or trading operations. Buy-side firms use credit ratings to comply with internal by-law restrictions or investment policies that require certain minimum credit ratings. Finally buy-side firms use credit ratings to ensure compliance with various regulatory requirements.

Sell-side firms also use credit ratings in addition to their own credit analysis for risk management and trading purposes. Many broker-dealers maintain rating advisory groups which generally assist underwriting clients in selecting appropriate credit rating agencies for their offerings and help guide those clients through the rating process. In addition, sell-side firms often act as dealers in markets that place significant importance on credit ratings. For instance, in the OTC derivatives market, broker-dealers tend to use credit ratings (when available) to determine acceptable counterparties, as well as collateral levels for outstanding credit exposures. Finally, large broker-dealers themselves frequently obtain credit ratings as issuers of long- and short-term debt.

Credit ratings even play a key role in private contracts, which in turn enhance their importance to the marketplace. SEC (2003) notes the widespread use of ‘ratings triggers’ in financial contracts. These are contractual provisions that terminate credit availability or accelerate credit obligations in the event of specified rating actions, with the result that a rating downgrade could lead to an escalating liquidity crisis for issuers subject to ratings triggers.

The question at stake is, therefore, how can investors identify, measure, and manage the risk of credit rating downgrades. A useful contribution from Turnbull (2009) suggests that investors and risk managers ask themselves the following four questions with a focus on problem areas: (i) what criteria do credit rating agencies use to assign ratings? (ii) What methodology do credit rating agencies use? (iii) What data do credit rating agencies use? (iv) What use is a rating? Similarly, IMF (2010) recommends that policymakers discourage the mechanistic use of ratings in private contracts, including investment manager internal limits and investment policies. However, they should recognize that smaller and less sophisticated investors and institutions that do not have the economies of scale to do their own credit assessments will inevitably continue to use ratings extensively. Hence, any steps to reduce overreliance on ratings should differentiate both according to the size and sophistication of the institution and the instruments concerned.

A proposal for the systemic regulation of ratings

In spite of the recent experience with credit ratings, a key finding of a recent stocktaking exercise by the Joint Forum (2009) finds that policymakers do not conduct formal assessments of the impact of the use of credit ratings on financial markets. In addition to micro-prudential regulation and efforts to broaden the regulatory perimeter, policymakers should consider the use of credit ratings in the entirety of financial markets, including in private contracts such as CDS contracts. Policymakers should better assess the nature and extent of the use of credit ratings in financial markets as well as their potential impact on financial stability. Such an approach should require both micro- and macro-level analysis, include all market participants, and take a global approach. The determinants of the supply and demand for ‘rated assets,’ especially in ‘good times’ and the implications of unanticipated abrupt downgrades in ‘bad times,’ should be assessed carefully. For instance, the proposed U.S. Financial Stability Oversight Council of prudential regulators should include credit rating agencies in its mandate.

Such an approach requires an assessment of the systemic effects of rating downgrades. Key components include: the type of institutions and markets that would be affected by downgrades, whether directly or indirectly and how systemic and interconnected they are; the consequences for financial markets and the economy in terms of market losses, liquidity shortages, loss of access to credit, and reduced liquidity; the factors that can increase downgrade risk, idiosyncratic and systemic; the measurement of systemic downgrade risk; and the management of downgrade risk at the systemic level through increased capital requirements or liquidity buffers, or other means.

Sy (2009) proposes a three-step approach to address the systemic risk of credit ratings (see Appendix I). As a first step, policymakers should have a good grasp of the risks inherent to credit ratings. In particular, policymakers should use ‘rating maps’ to identify the different channels through which rating downgrades can lead to systemic risk. Given the potential procyclicality of ratings, questions will also need to be asked in ‘good times,’ during boom cycles. Credit ratings can encourage the growth of the rated market where rated securities are transacted. This growth may also be accompanied by a higher volume of highly rated

---

16 Sell-side firms include broker-dealers that make recommendations and sell securities to their clients.
securities. This ‘rating inflation’ was a key development prior to the current crisis, and policymakers will need to get a full grasp of the determinants of ‘rating inflation’ incentives and the methodology used to justify substantially larger volumes of highly rated securities.

Second, questions will need to be asked about market participants’ exposure to abrupt rating changes. Policymakers will need to measure risks inherent to ratings once they are identified. A useful method for measuring systemic exposure to downgrade risk during boom cycles would be for regulators and regulated institutions to stress test their balance sheet and off-balance sheet positions. Risk managers have long been aware of the risks of credit downgrades, especially for fixed income portfolios. The current crisis has brought to the forefront the need for policymakers to manage the risks of such downgrades but, this time, at the systemic level. One first step would be to conduct scenario analysis in which the consequences of ratings downgrades for systemically important institutions and different types of rated securities are analyzed. Such an approach will depend on increased transparency in the rated markets. For instance, it will be key to have a clear sense of ‘rating triggers’ and other contractual arrangements, where ratings downgrades can lead to systemically important market portfolio rebalancing or a dry-up of liquidity. Finally, systemic institutions that are vulnerable to abrupt ratings downgrades may have to hold more capital or liquidity buffers.

Conclusion

The debate on CRAs emphasizes the need to reduce their oligopolistic power and the conflicts of interest inherent to the ‘issuer-pays’ model, and enhance transparency in their operations. Not surprisingly, the types of regulatory proposals are of a micro-prudential nature as this approach is the most suited to address such issues.

One key lesson of the U.S. subprime crisis and the unfolding European sovereign debt crisis is, however, that credit rating downgrades have negative externalities and can threaten financial stability. Micro-prudential regulations alone are not well suited to reduce the potential risks to financial stability that credit ratings can create. As a result, a number of principles to broaden the regulatory perimeter of CRAs have been proposed by the international community. In addition, a number of initiatives have emerged in the U.S. and the E.U. These include Section 939A of the Dodd-Frank Act, which aims to take ratings out of regulation, and the ECB collateral rules, which reduce the role of credit ratings in legislation, regulation, or supervisory policies.

In spite of all these efforts, it is important that policymakers take a systemic approach to credit ratings and conduct formal assessments of the impact of their use on financial markets, especially in private contracts such as CDS or institutional investors’ investment policies. Additional capital and/or liquidity buffers may then be used to mitigate the risks inherent to credit ratings. This approach would be consistent with the objective of macro-stress tests in Pillar 2.

References

- CFA Institute, 2009, “Member poll on credit rating agencies,” CFA Centre for Financial Market Integrity, April
- Partnoy, F., 2006, “How and why credit rating agencies are not like other gatekeepers,” Research Paper No 07-46, Legal Studies Research Paper Series, University of San Diego School of Law, May
- Richardson, M., and L. White, 2009, “The rating agencies: is regulation the answer?,” In Acharya, V. V., and M. Richardson (eds.) Restoring financial stability: how to repair a failed system, An independent view from New York University Stern School of Business, John Wiley & Sons
Appendix 1

- Rate securities
- Rated issuers
- Rated insurers
- Rating-based investment mandates
- Contractual rating triggers

2

- Lower capital requirements
- Higher market gains

3

- Higher capital requirements
- Larger market losses

4

- Higher funding costs
- Loss of market access

Abrupt and unanticipated rating downgrades

1

- Financial crisis
- Systemic shocks (e.g., subprime loans defaults, macro-economic shocks)
- New material information on issuer
- Revision of rating methodologies
- Recognition of conflicts of interest
- New regulation

High demand for rated

- Few collateral calls
- Available funding

Banks, insurers
- Investors such as banks’ off-balance sheet entities (OBSEs) such as conduits and SIVs

Issuers such as corporates, banks, and OBSEs
- Insurers including monolines

Figure 1 - “Ratings map”: the systemic risk of credit rating downgrades (bust cycle)

- Financial innovation
- Model risk in methodologies
- Low interest rate environment
- Conflicts of interest
- New regulation

2

- Low funding costs
- Increased issuance
- Increased insurance

3

- Higher funding costs
- Loss of market access

4

- Larger collateral calls
- Larger funding needs

Fire sales of securities

- Issuers such as corporates, banks, and OBSEs
- Insurers including monolines

- Investors including mutual funds and other buy-side funds

- Insurers of CDS such as AIG
- Banks sponsoring OBSEs

Figure 2 - “Ratings map”: the systemic risk of credit rating downgrades (boom cycle)
Insurer Anti-Fraud Programs: Contracts and Detection Versus Norms and Prevention

Sharon Tennyson — Associate Professor, Department of Policy Analysis and Management, Cornell University

Abstract
Opportunistic claims fraud is undertaken by claimants who did not contemplate the fraud prior to experiencing a loss event, and is often characterized by claim exaggeration rather than outright falsification of a loss. Because opportunistic fraud usually falls short of standards that define fraud as criminal, a hidden cost to insurers from policing opportunistic fraud is the possible damage to customer relationships and trust. Damage to customer relationships may in turn increase consumers’ tolerance of insurance fraud. This article reviews what we know from research on consumers’ attitudes toward insurance fraud, and on the social and psychological determinants of opportunistic cheating, and discusses its implications for insurer anti-fraud programs. The article concludes by suggesting that the success of insurer anti-fraud programs may be increased by distinguishing opportunistic fraud from planned (criminal) fraud, and applying insights from the research literature to the design of programs to deter opportunistic fraud.
Because fraudulent insurance claims may go undetected or may be dealt with indirectly, estimating the true prevalence of fraud is difficult. A number of different methods have been used to obtain such estimates, including expert analysis of closed claims, surveys of insurers, and crime statistics [Viaene and Dedene (2004)]. Other studies use statistical methods to estimate the relationship between a documented incentive to engage in fraud and the existence of excess loss costs or claims frequency, or to extrapolate the true magnitude of fraud based on characteristics of detected fraud cases. These studies can provide only indirect evidence of fraud, and their results are limited to the insurance market or line of business studied, but they have the advantages of rigor and of capturing both detected and undetected fraud. Although quantitative estimates of fraud differ across these different approaches to measurement, the conclusion from each is that the costs of fraudulent claims are substantial [Tennyson (2008)].

One of the difficulties in devising a management strategy toward claims fraud is the variety of ways in which fraud may present itself. Claims fraud may arise out of casual opportunity or from deliberate planning [Weisberg and Derrig (1991)]. If no loss event occurred but a claim is filed then the fraud is planned or outright. If a legitimate insured event occurred but circumstances are falsified in order to get excessive payments, then the fraud is opportunistic. Outright fraud may be perpetrated by an individual claimant on a one-time basis or may be carried out repeatedly by professional fraud-rings. Opportunistic fraud is undertaken by claimants who did not contemplate the fraud prior to the loss event, and is often characterized by claims exaggeration rather than outright falsification of a loss. Although planned fraud tends to receive more attention in the fraud-fighting agenda, much of insurance claims fraud is thought to be opportunistic in nature.

The distinction between planned and opportunistic fraud is important because optimal management responses may differ. Consider, for example, the question of law enforcement. Fraud is viewed as criminal when it displays characteristics sufficient to be prosecuted, such as evidence of a clear and willful act of material misrepresentation that violates a law to achieve financial gain [Derrig et al. (2006)]. Planned fraud is likely to be criminal fraud, and referral to law enforcement is appropriate. Opportunistic fraud may be criminal, but is more likely to fall short of that definition. Suspected fraud that falls short of the criminal presents a difficult management task. Insurers must be careful of both the legal requirements for fair and prompt settlement [Tennyson and Warfel (2010); Sykes (1996)] and of customer relations [Ericson et al. (2000)].

This latter point highlights a hidden cost to insurers from policing opportunistic fraud – the possible damage to customer relationships and trust. Hyman (2001) notes that insurance claim professionals are trained to be suspicious and are habituated to viewing claims as fraudulent. Baker (1993-1994) has written eloquently about the contradictions in insurance company rhetoric during the sales process and the claims process, shifting from an emphasis on trust and caring during sales to contractual language and legality during the claiming process. This divergence can undermine trust relationships, leading to reduced demand for insurance and/or increased demand for government regulation of insurers. Due to the particular psychology of insurance claiming behavior, it may also increase consumer tolerance of insurance fraud. This implies that the way in which insurers respond to opportunistic fraud is important.

The psychology of opportunistic fraud

Academic studies of consumers’ attitudes toward cooperation with an institution or authority suggest that perceptions of the institution are a determining factor [Axelrod (1986)]. Cialdini (1989) posits that consumers’ perceptions of the fairness of an institution affect its legitimacy in their eyes, and that consumers are more willing to comply with the rules of legitimate institutions. Two notions of fairness have been found to be important: procedural fairness, which is based on the equity and consistency of the process by which outcomes are determined; and distributional fairness, which is based on the equity of the outcomes themselves. Consistent with this idea, Tennyson (1997) analyzes consumer survey data from the U.S. and finds that consumers with negative views of insurance markets and institutions are significantly more likely to view claims fraud as acceptable. The Coalition Against Insurance Fraud [CAIF (1997)] notes that the public’s perception that insurers are unduly profitable accounts for some consumers’ acceptance of fraud.

Researchers have also examined how consumers use techniques of neutralization to justify unethical behaviors in the marketplace [Duffield and Graboski (2001), Strutton et al. (1994)]. Neutralization techniques are the rationalizations used when consumers who generally adhere to mainstream values attempt to reduce feelings of guilt for violating social norms of behavior. Common techniques include assertions that no one was really harmed by the action (denial of injury), assertions that the victim deserved the behavior (denial of victim), and assertions that those who would find fault are hypocritical given their own behaviors (condemn the condemners) [Strutton et al. (1994)].

While the lack of a readily apparent victim (denial of victim) is often thought to be a rationale for consumers’ accepting attitudes toward insurance fraud, the neutralization techniques employed by consumers have not been extensively studied. One exception is Fukukawa et al. (2005), who present several different consumer fraud scenarios to a sample of U.K.

1 Dionne and Gagne (2001, 2002) provide studies in the first category; see Ai et al (2010) for an application of the second approach.

2 An extensive body of research supports the conclusion that consumer perceptions of fairness influence attitudes toward compliance or cooperation with an institution. See the discussion in Tennyson (1997).
consumers. A scenario relating to insurance fraud presents a consumer who exaggerates a claim, and respondents are asked to evaluate the behavior. The results show that respondents who identify positively with the consumer in the scenario are significantly more likely than other respondents to agree that the fraud could be justified based on the insurer’s unfairness in pricing, or that the insurer might deserve such treatment.

Additional evidence of the use of these neutralization techniques can be found in two separate surveys of U.S. consumers, presented below. The surveys were both administered by telephone to randomly selected samples of consumers, but the geographic sampling range differed. One survey was administered to a representative sample of consumers in a single state, and the other survey was administered to a representative national sample of consumers. The surveys were administered by different survey firms, and were undertaken several years apart. In addition, the questions about insurance fraud were only a small part of each survey and the main focus of the two survey instruments differed substantially.

In each survey, respondents were asked their views of the acceptability of two variants of opportunistic claims fraud: inflating an insurance claim in order to help cover the deductible on a policy, and misrepresenting the nature of an incident to obtain payment for a loss not covered by the policy. Respondents to each survey were also asked a question regarding their attitudes toward insurers’ behaviors. In the single state survey, respondents were asked whether they agreed with the statement “insurance companies argue that claims are fraudulent as an excuse to get out of paying claims.” In the national survey, respondents were asked whether they agreed with the statement “if insurance companies treated people with more respect, people wouldn’t lie to them as much.”

Table 1 shows the percentage of consumers in each survey who found the various fraudulent actions to be acceptable, and the percentage who agreed with the statement regarding insurers’ behaviors. The data show that the mean levels of fraud tolerance varied substantially across the two samples. Respondents to the single state survey are much more tolerant of opportunistic insurance fraud than are respondents to the national survey. Respondents to each survey were also asked a question regarding their attitudes toward insurers’ behaviors. In the single state survey, respondents were asked whether they agreed with the statement “insurance companies argue that claims are fraudulent as an excuse to get out of paying claims.” In the national survey, respondents were asked whether they agreed with the statement “if insurance companies treated people with more respect, people wouldn’t lie to them as much.”

The survey responses also reveal a significant link between views of insurers’ behavior and attitudes toward opportunistic fraud. Table 2 shows these relationships in the single state survey. Respondents who believe Inflating an insurance claim to help cover the deductible. 27.8% 19.8%*

Misrepresenting the nature of an incident to obtain insurance payment for a loss not covered by the policy. 17.0% 5.7%**

Table 2 – Single state survey sample – fraud attitudes by mistrust of insurers (percentage who agree that fraud is acceptable)

Inflating an insurance claim to help cover the deductible. 7.8% 3.0%**

Misrepresenting the nature of an incident to obtain insurance payment for a loss not covered by the policy. 5.1% 1.2%**

Table 3 – National survey sample – fraud attitudes by mistrust of insurers (percentage who agree that fraud is acceptable)

* and ** t-test statistics indicate means across categories are statistically significant at the 10% and 5% confidence levels, respectively.

Tennyson (2008) conjectures that the low levels of fraud tolerance expressed in the national survey are due to the fact that the questions were asked after the respondents were asked questions that led them to contemplate their ethical beliefs (i.e., whether certain behaviors are ethical and the reasons they are considered more ethical or less ethical). The low levels of fraud acceptance are consistent with evidence from psychology experiments which find that individuals are less likely to lie or cheat if they are first “reminded” of their ethical beliefs (Mazar and Ariely 2008).

Table 1 – Attitudes of consumers from two survey samples (percentage who agree with statement)
insurance companies use fraud as an excuse to avoid paying claims, are 40 percent more likely to find claims exaggeration to be acceptable, and nearly 300 percent more likely (three times as likely) to find misrepresenting a loss to be acceptable.

Table 3 reports these relationships for the national survey. In this sample those who believe that insurance companies do not treat people with respect are nearly three times as likely to find claim exaggeration to be acceptable, and are four times as likely to find misrepresenting a loss to be acceptable.

These results demonstrate that consumers are prone to rationalize opportunistic fraud by pointing to negative behaviors of insurers, providing a strong caution to insurers that opportunistic fraud must be managed in a way that does not undermine customer relationships.

Consumers’ attitudes toward fraud are important because they constitute a social norm regarding claiming behavior, and by doing so they create social costs to individuals of undertaking the behavior. Higher public tolerance for fraud or the perception that fraud is commonplace can lead to more accepting attitudes toward fraud and thus to lower social costs of engaging in it. Tennyson (1997) finds that a 10 percent increase in the percent of other nearby consumers who find fraud acceptable leads to a 5.9 percent increase in the chance that a consumer will find fraud acceptable. Researchers have also demonstrated that exposure to other people’s unethical behavior can increase an individual’s dishonesty. In one experiment, researchers planted a person in the room who engaged in obvious cheating on the task assigned to the experiment participants. When exposed to this behavior, other participants were themselves much more likely to engage in cheating, but only when the cheater was perceived to be a part of their peer group [Gino et al. (2009)].

These results suggest that both accepting fraud attitudes and actual fraud behaviors may spread across peer groups, translating into a higher prevalence of fraudulent behaviors. There has been little research linking insurance fraud attitudes to the actions of individuals, but empirical studies of aggregate claim costs support such a relationship. In an analysis of bodily injury liability claim rates in private passenger automobile insurance, Cummins and Tennyson (1996) show that the percentage of consumers in a state who find claim exaggeration to be acceptable is positively related to statewide automobile insurance claim frequency, after controlling for other characteristics of the driving and insuring environment. Similarly, Colquitt and Hoyt (1997) find that the number of fraudulent life insurance claims is positively related to the percentage of a state’s consumers who find claim fraud to be acceptable.

Another important dimension of opportunistic fraud is the dishonest behavior of the ‘honest’ consumer. Most individuals express views that fall within our accepted standards of ethics and morality, and most consumers view themselves as honest and ethical. And, most of the time they behave in a manner consistent with that self-image – just not all of the time. In experiments in which subjects are given the opportunity to cheat on a task, for example, by taking a test and being asked to grade and reward themselves based on the ‘honor system,’ most overstate their performance by at least a small amount [Mazar et al. (2005)]. Cheating is more likely if no detection method is apparent – that is, the more a person really feels unobserved in the act of cheating, the more likely he or she will cheat.

Experiment participants are less likely to take advantage of opportunities to cheat if they are reminded of their internal ethical standards before engaging in the assigned task. For example, having participants write down as many of the Ten Commandments as they could remember was found to reduce the prevalence of cheating [Mazar et al. (2005)].

Ethics and personality traits also moderate individual reactions to the incentives and opportunities to cheat that are provided in experiments. One study finds that individuals who are more likely to shade the truth to themselves (to their own advantage) are also more likely to cheat in experiments, and that those who cheat are more likely to engage in other risky behaviors [Nagin and Pogarsky (2003)].

Implications for managing opportunistic fraud
Over the past twenty years, insurance claims fraud has received increasing attention in the insurance industry. The resources that many insurers devote to combating fraud have increased substantially and there have been many advances in fraud detection methods [Derrig (2002), IRC (2001)]. Much of the focus has been directed toward criminalizing fraud. Insurers have successfully lobbied for stronger laws against insurance fraud, and demonstrate an increasing willingness to litigate fraud cases and to refer cases to law enforcement agencies [Derrig et al. (2006)].

Care must be taken in applying these approaches to address the problem of opportunistic fraud. A criminal focus applied to these cases may reinforce negative perceptions of insurance institutions and may create perceptions that they treat customers unfairly during the claiming process. Negative perceptions may in turn encourage more fraud by increasing fraud acceptance and by providing easy rationalizations for fraud. Moreover, experimental evidence suggests that small changes in context or interactions may affect the likelihood that people engage in opportunistic behaviors. These effects suggest that deterring or preventing opportunistic fraud may provide the greatest benefits at the lowest cost. Applying these insights from research on the social and psychological dimensions of cheating may increase the success of anti-fraud programs, and additional research to better understand these dimensions of opportunistic insurance fraud would be helpful.
References

- Tennyson, S., 2008, “Moral, social and economic dimensions of insurance claims fraud,” Social Research, 75:4, 1181-1204
Revisiting the Labor Hoarding Employment Demand Model: An Economic Order Quantity Approach

Harlan D. Platt — Professor of Finance, Northeastern University
Marjorie B. Platt — Professor of Accounting, Northeastern University

Abstract

The labor studies literature has for many years accepted the labor hoarding theory. That theory derives from seminal work by Oi (1962), Solow (1964), Miller (1971), and Fair (1985). Those studies argue that as a result of the absolute cost of hiring and training certain workers that even when the economy turns down, firms avoid layoffs as would be expected in a neoclassical framework. Consequently during such time periods companies develop a reserve supply of workers. If labor hoarding occurs the employment cycle should be less extreme than the production cycle. From December 2007 when there were 115.5 million employed workers, the American economy lost 8.5 million jobs by January 2010, a 7.35% reduction in employment. During the same period, real GDP fell by 1.25%. This paper presents a different view of the demand for labor that is based on Baumol’s (1952) cost minimization paradigm for determining the optimal level of inventories. In the framework, the stock of employees is built up beyond the current need level in good times to minimize hiring costs but during periods of slack demand by consumers the number of excess workers is reduced. This alternative model appears to fit the current changes in unemployment and GDP better than the labor hoarding theory.

1 We are grateful to the editor of this journal for useful comments on an earlier version of this paper. We also wish to thank Leah Boustan for helpful comments. The usual caveat applies.
How firms make decisions regarding the hiring and firing of workers has long intrigued economists and political scientists of all persuasions. From Karl Marx’s Das Kapital (1867) to neoclassical economists’ various theories have proposed explanations for how firms make employment decisions. Marx employed the labor theory of value and viewed workers as selling labor power to capitalists. He felt that workers became “an ever cheaper commodity” as their productivity increased and were subject to layoffs due to the availability of a reserve army of unemployed workers and the vagaries of the business cycle.

Neoclassical economics treats firms as profit maximizers who rationally hire the correct number and mix of workers. The firm determines each worker’s marginal productivity and from it the value of their marginal product (VMP), which also includes information on product price. Workers are hired until their VMP just equals their wage rate. Employing this decision calculus leads to an optimal employment level and maximum profits for the firm. Once the optimal workforce level is achieved, further hiring occurs in response to a reduction in the wage rate or when the VMP of workers increases due to a rise in the product price or an increase in worker productivity. That is, neoclassical economics assumes that the firm maintains its employment level until the fundamental hiring equation involving wage rates, productivity, and product prices is disrupted.

Layoffs occur in the neoclassical economic framework when the marginal worker’s VMP is not sufficient to cover their wage. This event follows from a rise in the wage rate, a decline in worker productivity, or the firm’s product price declines. Firms make these decisions rationally with a focus on the profitability of the incremental worker. Friedman and Wachter (1974) observe that this paradigm is critical to the relationship between the goods market and the labor market.

Precision of the neoclassical model
The fundamental neoclassical model supposes a constant flow of new hires and fires as a firm seeks to maintain the delicate balance between worker VMP and the wage rate. In fact, firms seldom adhere strictly to the formalized dictates of the neoclassical model and their hiring and firing decisions are less systematic than the theory implies [Kniesner and Goldsmith (1987)]. For example, few firms hire workers on Monday and then fire them on Tuesday even if Tuesday requires fewer workers. Other forces at work – ennui, cultural or regulatory mandates, and information limitations – that counteract the forces of strict neoclassical economics. Alternative theoretical models opposing the neoclassical viewpoint include the segmented labor market [Cain (1976)], the dual labor market [Piore (1969)] and labor hoarding [Solow (1964)]. The segmented labor market explains why different wage rates can occur for workers with similar characteristics while the dual labor market postulates the presence of good and bad jobs. Bulow and Summers (1986) argue that primary jobs, those having the best pay, result from an inability of employers to monitor workers while contingent jobs result from situations where the employer can easily monitor the worker.

The labor hoarding theory
The theory of labor hoarding seeks to explain the disconnection between changes in output and employment. The theory focuses on a variety of costs that include hiring, training, and firing which impact the firm when it modifies its labor force. The labor hoarding theory argues that these costs make workers a quasi-fixed factor of production. As a consequence, the firm partially disassociates its hiring and firing decisions from fluctuations in cyclical demand. While Oi (1962) was the first author to describe these employment costs, the name was coined by Solow (1964). Later work on labor hoarding includes Miller (1971) and Fair (1985). Ohanian (2001) utilized labor hoarding as one of five factors to explain the extraordinary 18% decline in labor productivity measured during the great depression. However, his five factors only explain about one-third of the actual decrease in productivity.

The labor hoarding theory argues that when firms have significant labor adjustment costs (i.e., human resources-related costs), they reduce the association between hiring and firing decisions and output. By contrast, in the neoclassical framework when the economy turns down and product demand and price decline (reducing the VMP), firms lay off workers because in a neoclassical framework workers are a true variable cost of production. The labor hoarding model relaxes the association between economic output and a firm’s labor demand. Instead, it argues that when the absolute cost of hiring and training is considerable, firms avoid layoffs during slow time periods. Consequently during such time periods companies develop a reserve supply of workers.

Labor hoarding has an intuitively appealing logic. Suppose it costs $100,000 to recruit and train a new key employee. In an economic downturn that worker might not be needed for several months. Letting that worker go would save the firm several month’s salary and benefits but unless the cost savings exceed the $100,000 recruitment cost it would be an uneconomic strategy. Instead, according to the theory, the firm would retain or hoard the worker, expecting to benefit from that employee’s labor when demand increases.

Empirical tests of the labor hoarding model have for the most part not rejected the theory’s implications. For example, Bernanke and Parkinson (1991) tried to understand procyclical labor productivity and compared labor hoarding against the real business cycle (procyclical technological

---

2 Thorstein Veblen coined the term neoclassical economics in Preconceptions of economic science (1900). Neoclassical economics assumes rational decision makers, with full information, seek to maximize their utility (individuals) or profits (businesses).

3 See Karl Marx, Economic and philosophic manuscripts, 1844.
shocks) and theories pointing to increasing returns. They found little evidence of a real business cycle effect but in various industries their tests support the labor hoarding and increasing returns hypothesis. In other words, labor hoarding appears to be a common practice though not in all industries; industries employing less skilled workers with higher labor supplies appear to find labor hoarding to be a suboptimal strategy. Burnside et al. (1993) studied the Solow residual and considered whether it was related to labor hoarding. The Solow residual measures productivity and is calculated holding capital and labor inputs constant. In fact, Burnside et al. (1993) found that a significant portion of changes in the Solow residual were attributable to labor hoarding. They estimated that standard real business cycle models overestimated the variance impact of innovations to technology by approximately 50%. More recently, Wen (2005) derived a theoretical model that supports the labor hoarding argument especially when information-updating technologies and inventory management techniques reduce inventory fluctuations. However, Wen notes that labor hoarding is less likely when the cost of holding inventories of goods and services is lower than the cost of hoarding labor.

Problems with labor hoarding
Despite empirical and theoretical support in the literature for the labor hoarding view of the labor market, recent observations from the financial crisis of 2007-2010 cast some doubt on the theory. If labor hoarding occurs then the employment cycle should be less extreme than the production cycle. Since December 2007, when there were 115.5 million employed workers, the American economy lost 8.5 million jobs by January 2010, a 7.35% reduction in employment. During the same period, real GDP fell by 1.25%. Granted that the labor hoarding theory only suggests that certain workers will be retained during periods of slack demand, the extraordinary gap between the change in employment and output argues for a reexamination of the theory. The business press has noted this extraordinary gap and has suggested that perhaps things have changed recently, upsetting historical norms.6

This paper presents a more general version of the theoretical labor hoarding model. Unlike Wen (2005) who considered how firms could inventory goods and therefore reduce their short term labor demand, this paper takes the view that workers themselves are an inventory (of ready labor) and that firms decide on the optimal inventory level of workers to hold. That is, the firm chooses a level of labor to hoard, not in response to a decline in economic activity but as an ongoing, day-to-day policy. The model derives the optimality conditions for the number of excess workers to be hired, subject to the cost of hiring and inventorying workers, relative to the demand for labor derived with neoclassical optimality conditions. That is, it identifies an optimal number of excess workers that firms should hire but not immediately put into gainful use as they anticipate a future need for workers.

The differences between this model and the labor hoarding model are twofold. First, in the labor hoarding model firms are assumed at any given time to hire the optimal quantity of workers (based on neoclassical terms) but to not fire them when a reduction in economic activity results in a lower optimal level of employment. The model presented in this paper argues instead that the optimal hiring level exceeds the level ordinarily derived with neoclassical conditions. The excessive hiring, according to the new model, is designed to reduce the average cost of hiring and training workers. The second difference between the two models is how they deal with declines in short-term labor demand. In the labor hoarding model some workers are retained during a downturn to avoid the future cost of rehiring them. The new model argues that the length of time expected to pass before excess workers are put into gainful employment increases in an economic downturn, thereby increasing the cost of holding excess workers. As a result, excess workers are not retained as is postulated by the labor hoarding model but instead are let go when the economy slows. In other words, there will be more layoffs than would be expected in the strict neoclassical or labor hoarding views. Labor demand would, in fact, be more volatile than output when output is falling.

Hiring more workers than are currently needed is not an unusual practice. In fact, it is unlikely that most firms actually hire on a daily or even weekly basis. Firms that hire seasonally (i.e., accounting firms or department stores) or even once a year (such as law firms, sports teams, or universities) typically hire enough new workers to avoid the need to reenter a limited labor market later in the year. Some newly hired workers may not be put to work immediately but may instead wait for demand to rise later in the year. Should sales not reach their expected level after some time has lapsed, these firms begin to lay off some of the excess workers in the hiring pool exactly as the new model suggests.

The EOQ type model of short-term labor demand
One of the most widely adapted economic models is Baumol’s (1952) cost minimization paradigm for determining the optimal level of inventories. While Baumol’s original work focused on the demand for money, ironically the trade-off model proposed by Baumol was originally described by Harris (1913) where he showed how a company could minimize its physical inventory costs. That model is now referred to as the Economic Order Quantity or EOQ model of inventories. The EOQ model assumes that firms trade off the average cost of purchasing inventories (a negative function of the size of the order) against the cost of holding excess inventories (a positive function of the size of the order). For example, a firm

---

4 The word residual indicates that productivity explains the change in output that is not related to capital accumulation or other factors leading to an increase in output.

5 Herbst, M., 2009, “Jobless rate hits 10.2%,” Business Week, November 6. This article commented that “The ratio of job cuts to losses in gross domestic product, or GDP, in this recession has surpassed the historical norm.”

159
that buys its annual inventory needs in a single purchase has minimized the cost of placing orders for inventories but has maximized the cost of holding inventories throughout the year.

The intellectual underpinning of the Baumol/Harris model is the notion that total inventory costs rise as the stock of goods in inventory rises (i.e., due to higher carrying costs) and fall as inventories are ordered less frequently but in larger volumes. These two costs move in opposite directions since a reduction in ordering frequency saves money consumed in the administrative function but then increases the average volume of inventories which raises interest and other carrying costs. The trade-off between ordering and carrying costs is shown in Figure 1. The total cost curve in the figure combines the two components of inventory costs. For a company that knows its annual inventory needs with certainty, optimal order size (i.e., the number of units to order each time) is found at the point where the total cost curve reaches its minimum point. Assuming that new inventories are delivered instantaneously, these inventories are worked down to zero at which point the company reorders.6 The model is referred to as the economic ordering quantity model or EOQ model since it yields the economic or least cost ordering quantity.7

Before developing a model of short-run hiring, it is first necessary to consider why a company would tolerate or even desire inactive or extra workers.8 After all, excess workers impose a cost and provide no revenue to the firm. Extra workers are unneeded in a world with known and fixed demand levels, a stable economic environment in which the existing workforce is unlikely to retire or quit and new hires are made instantaneously without costs. In a world with frictions, including a nonzero employee turnover rate, extensive training (cost and time), and expensive and delayed hiring, companies need reserve employees now to smoothly and economically fill future employment needs in the future.

In an unstable environment, companies desire larger work forces than they currently need. The extra workers constitute a reserve supply of hired, trained, and available workers able to step in when product demand rises. The number of inactive workers desired by a company is assumed to be a fraction, \( \alpha \), of the active workforce, as seen in this equation: Inactive workers \( t = \alpha \) active workers \( t \) (1), where \( \alpha \) is the factor that describes the number of desired idle workers and \( t \) denotes year.

Firms set their \( \alpha \) to account for future labor turnover and anticipated growth in labor demand. All else equal, as firms’ expectations about future demand and growth increase (decrease), the value of \( \alpha \) should increase (decrease). The total number of employees equals the sum of active (productive) and inactive (nonproductive) workers. If active workers retire or quit at a steady rate and if firms hire only one inactive worker at a time then firms might have to constantly hire new workers throughout the year. Alternatively, if firms hire more workers than they need, for both replacement and growth needs, workers can be recruited and trained as a group, and transferred individually into the active workforce as turnover occurs.

When no output growth is expected and the turnover rate is steady, the desired number of inactive workers is constant over time. In that case, \( \alpha_t \) equals \( \alpha_{t-1} \), and there is no change in the number of desired inactive workers, as in this equation:

\[
\Delta \text{inactive workers} = \alpha_t \text{active workers}_t - \alpha_{t-1} \text{active workers}_{t-1} = 0 \quad (2).
\]

In that case, the number of new inactive workers hired during a year equals the number of workers who retire or quit which would then also equal \( \alpha_t \times (\text{active workers}) \). By contrast, when output growth is anticipated, a larger number of new inactive workers are hired during a year. The factor \( \alpha \) changes in that case. That is, the proportion of total employment that is inactive workers, \( \alpha_t \), increases with the expected growth in sales and as the demand for active workers increases. In addition, the inactive proportion of the workforce increases if the turnover rate of existing workers increases. Firms respond to a higher turnover rate by increasing their \( \alpha_t \)'s and their number of inactive workers hired.

When \( \alpha_t \) increases and is larger (smaller) than \( \alpha_{t-1} \), then the change in the number of inactive workers hired is positive (negative). In that case, a larger (smaller) number of new inactive workers is hired during a year. Thus, the number of inactive workers hired equals \( \alpha_{t-1} \times (\text{active workers}_{t-1}) + \alpha_t \times (\Delta \text{active workers}_t) \).

---

6 In a world with less certainty and the potential that orders are delayed, firm’s hold a safety stock of inventory.
7 The EOQ formula is found by minimizing the total cost curve. The formula is found in every financial management textbook; see for example, Ben-Horim, M., 1987, Essentials of corporate finance, Allyn and Bacon, page 469.
8 Gardner Ackley (1961), a member of President Johnson’s council of economic advisors, stressed that the aggregate labor market can not clear with all workers employed. At best, perhaps a 2% unemployment rate is achievable with these workers moving between jobs.
Fitting the demand for labor into the EOQ model

This paper argues that there are two types of workers: active and inactive. This bifurcation of labor demand can be modeled within the EOQ inventory framework to model labor demand so that it includes a short run demand for inactive workers. These inactive workers are demanded by firms to serve as a buffer stock of employees who are put to work when older workers leave or output expands. Firms must decide whether new inactive employees should be hired. They are assumed to choose the least cost alternative. The choices are a) all at once (one hiring cycle per year), b) one at a time (as many hiring cycles per year as the number of workers who retire or quit plus the number of new positions that open up), or c) some number of hires in between.

The hiring process creates three distinct costs for the firm. The first cost is the administrative cost of hiring which includes, design and placement of advertisements, screening the applicant pool, training new hires, and an adjustment period during which new hires learn their jobs.

The second hiring costs arises when there are too few inactive workers. In that case, the firm may incur an indirect (opportunity) cost that results from lost profits due to delayed or canceled output caused by missed production because there are too few workers. The third cost arises when the firm has inactive workers who must be paid even though they are not producing any output; this is analogous to a carrying cost for inventory.

The total cost of hiring includes administrative costs, indirect costs, and the cost of carrying inactive workers. Distinct economies of scale in hiring and training may encourage firms to engage in multiple hirings. Multiple hirings mean that more workers are hired than are immediately required. As a result, the firm develops a supply of inactive workers who are available for assignment quickly should the need arise. These extra workers also reduce a second hiring cost, lost profits arising when there are not sufficient workers. However, having inactive workers also raises the firm’s costs since inactive workers are not productive while they are getting paid awaiting an assignment.

As was true for inventories in the EOQ framework, a firm decides how many inactive workers to hire by trading off the cost of holding an inventory of extra workers (wages and benefits for inactive workers) against costs savings arising from hiring and training more new workers at one time. Average hiring costs are a decreasing function of the number of workers hired. Economies result in cost savings as more workers are employed. Carrying costs are an increasing function of the number of workers hired since the number of inactive workers rises as more workers are hired. The optimal number hired at one time increases as carrying cost decrease (i.e., when wages are lower) or hiring cost increase (i.e., the cost of advertising rises). But since the two costs move in opposite directions, the optimal number to hire at one time is found by trading off the two costs as shown in Figure 2. The number of workers to hire indicated at the minimum point on the total cost of hiring curve is the least cost number of new hires.

Layoffs in the EOQ labor demand model

Much of this paper has talked about the hiring process. This section discusses the lay off process when firms have too many workers. Companies rarely have exactly the number of workers they require. The EOQ labor demand model argues that firms maintain a supply of extra or inactive workers who can speedily step in to assist when output grows. A firm has more inactive workers when wages and benefits are lower, hiring costs decline, and as the company expects more growth in product output.

But what happens when, for example, expected output declines? In that case, the firm needs fewer inactive workers. In a declining output environment, the only inactive workers the firm needs are those required to replace workers who retire to change jobs and even they may not be required depending on the labor supply. Generally, there are fewer job changes in a recessionary period because the fall in output reduces the supply of alternative jobs [Diebold et al. (1997)], which leads to a reduction in the demand for inactive workers. The firm whose output is falling thus begins to lay off its inactive work force and if the output decline is sufficiently large or expected to be long term the firm may even begin to lay off a portion of its working (productive) employees. The reduction in the number of actively engaged workers corresponds to the dictates of the neoclassical model. The firm would fire any worker whose VMP is less than their wage. The critical fact is that the number of layoffs would then be larger than the decline in output suggests – in contrast to what is predicted with the labor hoarding model.

9 In times of uncertainty, companies may decide to change the mix of full-time (FT) to part-time (PT) workers in their reserve pool to give greater emphasis to PT workers. In doing so, the firm would reduce the associated cost of hiring, since PT workers are truly variable costs. Addressing the impact of reserve pool mix is an interesting question, but goes beyond the scope of this paper.
Both the hiring and firing decisions of firms depend on the demand for inactive workers. Layoffs result for two reasons. First, layoffs happen when output declines and the neoclassical model indicates that fewer workers are required. The second factor causing layoffs is when any of the factors determining the optimal size of the pool of inactive workers moves in the opposite direction; that is, layoffs increase when:

- Wages and benefits increase – since the carrying cost for holding an inventory of inactive workers rises.
- Hiring costs (either direct or indirect) decrease – since it is less expensive to hire workers more frequently.
- Expected future output is lower than anticipated – since fewer active workers are anticipated in the future.

While wage and benefit increases also result in layoffs in the neoclassical model discussed above, the two new factors causing layoffs within the EOQ type labor model, hiring costs and expected future output, do not lead to layoffs in the pure neoclassical model.

**Policy implications of the EOQ labor model**

The new EOQ model predicts more volatility in employment than would be the case under either the neoclassical labor model or the labor hoarding model. Comparisons made by the authors between economies based on natural resources versus those with a diverse industrial base suggests that during this worldwide period of economic decline and uncertainty, some economies have weathered the storm better than others. Australia, for example, has recorded moderate unemployment (5.3%) recently and moderate economic growth (2.7%) [Global Times (2010), EconGrapher (2010)]. By contrast, the U.S., in the second quarter of 2010, reported 9.7% unemployment [Portal Seven (2010)] and 1.6% growth in real GDP [Hilsenrath and Reddy (2010)]. Slower growth translates to lower expectations about future growth and consequently, as described above in the EOQ model, results in less hiring of new workers. The same trend holds when comparing growth and unemployment rates in individual states in the U.S. [Fernando and Jin (2010)]. During this time of severe economic distress, states with active natural resource and agriculture industries as well as highly educated populations have experienced relatively low unemployment rates (3.5% to 6.8%) compared to the national average (9.7%). Presumably these lower unemployment rates result from firms in those states that have higher expectations for output growth and a correspondingly higher \( \alpha \) (the adjustment parameter in the EOQ model).

The public policy implications of this increased volatility include the need for government leaders to set realistic, optimistic expectations about future output changes as well as legislative decisions regarding unemployment benefit levels and duration. Managing expectations is important so that firms do not over-react to changes in the output market by making drastic and erratic changes in labor demand. Firms often assess the future likelihood of product or service demand and incorporate that expectation into future decisions about capital investment, product development, or labor demand [Kelleher and Zieminski (2010)]. That is, with realistic expectations clearly articulated by political leaders, firms can make more informed decisions about their need for workers. For a time during the current financial crisis/economic recession, the Obama administration was presenting an overly optimistic view of likely future outcomes [Port (2010), Burnett (2010)]. Realism is critical if listeners are to believe prognostications by politicians.

With increased volatility in labor markets, the question of whether current unemployment benefits are able to accommodate the likely breadth and length of unemployment is a concern. If a single or a few industries (akin to nonsystematic risk in financial markets) experience a downturn in demand, it is likely that the current length of unemployment benefits can accommodate the duration of unemployment as the displaced worker tries to find a subsequent job within the industry as it recovers or in another, healthier industry. However, if the economy as a whole (like systematic risk in financial markets) experiences a recession, as is the case in the U.S. and many other countries currently, the effects on the labor market are more extensive and volatile because firms lay off both active and inactive workers. Legislative bodies may need to realign the length and breadth of unemployment benefits to better match the likely period of unemployment experienced by workers.

Fiscal and monetary stimulus can go only so far in helping promote an economic recovery [Port (2010)]. In January, 2010, Obama called for tax breaks for small business owners to help them hire new workers [CBS News (2010)]. Decision makers at the firm level are likely to respond more to tax policy changes rather than fiscal or monetary policy initiatives. Giving firms a tax credit would help them hold on to their employees, some of whom they might otherwise lay off. Infusing stimulus dollars into the economy may or may not have an indirect effect on any given firm.

**Summary and policy implications**

The EOQ-based model of labor demand argues that during the rising portion of the economic cycle firms over-hire workers to provide an excess inventory of not yet needed workers relative to the demand that would be found in the neoclassical framework based on their production schedule. During the negative phase of the business cycle, the new model argues that firms fire or lay off more workers than would be expected in the labor hoarding view of the world. The policy implication of this work is that governments need to be more proactive, communicate effectively, and creatively use their tool kit of policy levers to fight unemployment as a recession ensues since otherwise firms will let more workers go than is expected. On the other hand, governments should modulate this policy as the economic recovery commences since firms accelerate their hiring in order to rebuild their inventory of excess workers.
References

- Ackley, G., 1961, Macroeconomic theory, Macmillan Co., Toronto
- Harris, F. W. 1913, "How many parts to make at once," Factory, The Magazine of Management, 102, 135-152
- Marx, K., 1844, Economic and philosopich manuscris and the communist manifesto, Prometheus Books; 1st edition (March 1988)
- Marx, K., 1867, Das kapital, Synergy International of the Americas, Ltd (September 1, 2007)
- Solow, R. M. 1964, "Draft of the presidential address to the Econometric Society on the short-run relation between employment and output
- Vablen, T., 1900, Preconceptions of economic science, Kessinger Publishing, LLC (June 17, 2004)

The Capco Institute Journal of Financial Transformation
Revisiting the Labor Hoarding Employment Demand Model: An Economic Order Quantity Approach

163

Jannis Bischof — University of Mannheim
Michael Ebert — University of Mannheim

Abstract
Accounting for financial instruments is based on a combination of fair value and amortized cost measurement. This paper examines how IAS 39’s mixed accounting model is reflected in measurement and presentation choices of international banks and how those choices will be altered by future regulation (IFRS 9 and Basel III). Potential problems arising from the approach taken by the IASB, such as earnings management or biases in investor perception, are identified and discussed.
The increasing use of complex financial instruments in the financial services industry was accompanied by major changes of the regulations governing financial institutions. One of these changes was the introduction of new accounting standards for financial instruments by the IASB. This paper describes the development of bank accounting following the introduction of IAS 39 in the late 1990s and examines the impact of future regulations when IFRS 9 and Basel III become effective.

The most fundamental effect of the initial adoption of IAS 39 was the introduction of fair value as a measurement base for particular financial instruments, such as financial derivatives and trading securities. At that time bank balance sheets in most countries were entirely based on amortized cost accounting [one exception was Denmark, see Bernard et al. (1995)]. As more countries introduced IFRS to their national reporting regime during the last decade, the use of fair values in banks’ financial statements steadily increased. However, the increasing influence of IAS 39 did not result in consistent reporting for financial instruments. IAS 39 is not a consistent accounting standard in the sense that it does not require uniform measurement of economically identical transactions [Wüstemann and Wüstemann (2010)]. Rather, the standard is based on a variety of accounting choices, most importantly the choice between fair value and amortized cost measurement. The resulting set of rules is commonly described as a ‘mixed accounting model’.

In this paper, we focus on two drawbacks arising from IAS 39’s mixed accounting model. First, we link the choice of measurement base to disclosed items in banks’ financial statements. Research in accounting and behavioral finance suggests that investors’ risk perception is influenced by both content and form of disclosure. Inconsistent disclosure resulting from inconsistent measurement, therefore, might facilitate biases in the investors’ perception of reporting banks. Second, the accounting literature provides ample evidence that managers opportunistically use the discretion offered by accounting standards. We review this literature and identify potential ways in which IAS 39 might be exploited for purposes of earnings management.

### The mixed accounting model for financial instruments under IAS 39

The accounting rules for financial instruments under IFRS are typically described as a mixed accounting model [Gebhardt et al. (2004); Walton (2004)]. This description refers to the three different measurement bases introduced by IAS 39: fair value through profit or loss, fair value through other comprehensive income (OCI), and amortized cost [see also Spooner (2007)]. Trading securities, which are held for short-term profit taking, and all financial derivatives, except those designated for hedge accounting, are measured at fair value through profit or loss. For all other financial assets and liabilities IAS 39 offers several accounting choices. First, fair value measurement through profit or loss can be applied on a voluntary basis if the fair value option is chosen for an instrument. The application of the fair value option is allowed in order to reduce accounting mismatches (i.e., to forgo the complex hedge accounting requirements and yet offset fair value changes in economic hedge relationships), to reflect internal reporting practices, or to avoid the separation of compound instruments. Second, amortized cost is the relevant base for the measurement of financial assets categorized as either loans and receivables (debt instruments not traded on active markets) or held to maturity (marketable debt securities with a fixed maturity). Amortized cost is also relevant for all financial liabilities not measured at fair value through profit or loss. Third, all financial assets except for trading assets and derivatives can be designated as available for sale (AFS) and measured at fair value through OCI. The AFS category of IAS 39 is applicable for both securities and loans. The categorization is required at the initial recognition of any financial asset and liability.

The persistence of the mixed accounting model in accounting for financial instruments has historical and political reasons. A fair value approach cannot be derived as such from the extant IFRS conceptual framework. Neither is fair value defined as a measurement base nor is the market valuation of an entity’s individual components defined as an accounting objective. Nevertheless, the IASC had proposed a full fair value measurement of financial instruments as early as 1991 [Exposure Drafts E40 and in 1994 E48, see Cairns (2006)]. A mandatory full fair value approach was proposed in the discussion paper, “Accounting for financial assets and financial liabilities,” issued by the IASC in March 1997. The draft standard on accounting for financial instruments announced by the ‘joint working group’ of standard setters (JWG) in 2000 adopted this proposal. This agreement was particularly justified by the criterion of relevance [Basis for conclusions, para. 1.8] which is one of financial statements’ qualitative characteristics as defined by, for example, the IFRS conceptual framework (para. 26). However, the only theoretical concept of relevance the JWG refers to is the concept of relevance for equity valuation [Basis for conclusions, para. 1.12]. A measurement base is of higher value.

<table>
<thead>
<tr>
<th></th>
<th>U.S. (largest banks)</th>
<th>U.S. (other banks)</th>
<th>Developed countries (IFRS)</th>
<th>Emerging and developing (IFRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV – assets</td>
<td>24.60%</td>
<td>16.20%</td>
<td>21.20%</td>
<td>13.20%</td>
</tr>
<tr>
<td>FV through OCI</td>
<td>15.00%</td>
<td>15.70%</td>
<td>7.20%</td>
<td>9.10%</td>
</tr>
<tr>
<td>FV through P/L</td>
<td>9.60%</td>
<td>0.55%</td>
<td>14.00%</td>
<td>4.10%</td>
</tr>
<tr>
<td>thereof: FV option</td>
<td>0.80%</td>
<td>0.20%</td>
<td>4.90%</td>
<td>1.60%</td>
</tr>
<tr>
<td>FV – liabilities</td>
<td>6.20%</td>
<td>0.30%</td>
<td>9.80%</td>
<td>2.80%</td>
</tr>
<tr>
<td>thereof: FV option</td>
<td>3.50%</td>
<td>0.15%</td>
<td>3.60%</td>
<td>1.10%</td>
</tr>
</tbody>
</table>

Table 1 – Extent of fair value measurement around the globe (financial year 2008, data is scaled by book value of total assets).
relevance than another if the resulting accounting figure is more highly associated with an entity’s stock price or its equity market value [Barth et al. (2001); Holthausen and Watts (2001)]. This association was empirically tested for fair values and amortized costs of almost all kinds of financial instruments [Barth (1994); Ahmed et al. (2006)]. Studies suggest a higher value relevance of fair value measurement for all instruments except for certain off-balance sheet transactions such as loan commitments, of which the market price can hardly be estimated. A thorough review of those studies is presented by Linsmeier et al. (1998).

The controversy about the JWG proposal was mainly due to two economic consequences of the approach that were not directly addressed by the standard setters. First, banking institutions feared high implementation costs as the internal measurement of financial instruments not held for trading was regularly not based on fair value estimations and fair values were thus not readily available in the absence of quotations on active markets [Gehhardt et al. (2004)]. Second, banking supervisors feared an increase in earnings volatility resulting in financial instability [European Central Bank (2004); Novoa et al. (2010); Walton (2004)]. The latter point was of particular importance in the debate about the role of fair value accounting during the recent financial crisis [André et al. (2009); Khan (2009); Laux and Leuz (2010)]. In conclusion the mixed accounting model can be considered a political compromise.

Evidence from reporting practices by banks around the globe suggests that in spite of the extended fair value option, a majority of financial instruments are measured at amortized cost. In fact, fair value measurement is still the exception for most financial assets. Analysis of a sample of 507 banks from the U.S. and 783 banks from IFRS-adopting countries reveals that only large U.S. banks and banks from developed countries outside the U.S. apply fair value measurement for a substantial fraction of assets [Bischof et al. (2010)]. For the 32 largest U.S. banks, fair value assets (liabilities) make up 24.6% (6.2%) of total assets (liabilities) on average. However, only 9.6% of total assets are measured at fair value through profit or loss. Thus, fair value changes of most assets are only reflected in OCI and, consequently do not affect volatility of net income. Figures are slightly different for IFRS adopting banks from developed countries (according to the IMF classification). On average, 21.2% (9.8%) of total assets (liabilities) are measured at fair value with only 7.2% of assets being measured at fair value through OCI. The larger fraction of assets measured at fair value through profit or loss results from the more extensive use of the fair value option by IFRS banks. IAS 39 introduced this option in 2004, whereas it was not available under U.S.-GAAP before the financial crisis and is only rarely used by U.S. banks today. Fair value accounting has an even smaller impact on the reporting practice of small U.S. banks as well as banks from emerging and developing countries (according to the IMF classification). Especially for small U.S. banks, the fraction of assets (liabilities) measured at fair value through profit or loss, which takes a value of 0.5% (0.3%), is absolutely immaterial [see also SEC (2008)]. Overall, the mixed accounting model can be considered a model which is primarily based on amortized cost accounting rather than on fair value accounting.

**Consequences for bank balance sheets**

**Balance sheet presentation**

Normally, an IFRS-adopting bank has three general options for presenting financial instruments on its financial statements. The first option, though not widely applied in IFRS financial statements of European banks (Figure 1), is a presentation by investment purpose that distinguishes, for example, between a hedging and a trading instrument or between a long-term and a short-term investment. A representative example is the balance sheet of the German Commerzbank (Table 2). The second option is a presentation by product type that distinguishes, for example, between stocks, bonds, and derivatives. This presentation format was advocated by the joint working group of standard setters which aimed particularly at a distinction between derivative and non-derivative instruments. It later recommended the application of this format in the draft standard on accounting for financial instruments [Basis for conclusions, para. 5.1-5.5]. The detail of information about derivatives usage provided by banks indeed seems to have improved in the 1990s, at least in the U.S. [Edwards and Eller (1996)]. However, there is some convincing evidence that a distinction of financial instruments based exclusively on their type will result in a biased risk perception by investors [Koonce, Lipe, and McAnally (2008, 2005); Koonce, McAnally, and Mercer (2005)]. This may be one reason why less than one-fourth of banks apply this format in their IFRS financial statements (Figure 1) and why disclosure of derivatives usage by banks is still considered to be incomplete [Woods and Marginson (2004)]. The choice of this format is common in Nordic countries such as Denmark, Norway, or Sweden. A representative example is the Swedbank balance sheet (see Table 2). The third possible format, used in Denmark, Norway, and Sweden is the presentation by investment purpose (Figure 2). A thorough review of the Danish financial reporting practice by banks is presented by Linsmeier et al. (1998).
by a majority of banks, is a presentation by measurement category. IFRS 7 allows a bank to use those measurement categories as line items on the financial statement. A representative example is the balance sheet of the Spanish Santander Group. It is for this last option that the choice of an instrument’s measurement base does not only affect the company’s income but also presentation and disclosure.

The overview in Table 2 links balance sheet items to the underlying IAS 39 measurement categories. The findings demonstrate that it is virtually impossible for a user of a financial statement to understand a bank’s application of fair value measurement unless the measurement bases are directly presented on the balance sheet. Panels B and C show that, if another format is chosen, several line items aggregate instruments with differing measurement bases. For example, Swedbank measures some bonds at fair value, while other interest-bearing securities are measured at amortized cost (held-to-maturity). Similarly, Commerzbank’s investment securities are measured at fair value through profit or loss (fair value option), at fair value through OCI (available-for-sale), and at amortized cost (loans and receivables). This finding is in accordance with the observations of Klumpes and Welch (2010) for U.K. banks. Because the choice of a measurement base for non-derivative financial instruments is left to management’s discretion under the mixed accounting model, even companies which hold financial instruments of identical economic characteristics could present financial statements that differ both in the measurement and in the labels of the individual line items. In the financial services industry, economic identity can easily be established by exploiting the replicability characteristic of financial derivatives. For example, a company engaged in a non-contingent derivative financial contract is obliged to categorize this contract as a trading instrument and to measure it at fair value through profit or loss even if it was actually acquired for hedging purposes. In order to circumvent this obligation, a company might enter into non-derivative lending and borrowing contracts that exactly replicate the future cash flows of the non-contingent derivative. If a company opts for letter option, there is an accounting choice between three different measurement categories. A company can apply the fair value option, it can use the available for sale category, or it can classify the instruments as loans and receivables. If the presentation by measurement categories in accordance with IFRS 7 is chosen, the choice of measurement category determines the label of the balance sheet item presenting the financial contract.

| Panel A – Presentation of assets by measurement category [Santander (2009)] |
|-----------------|--------|--------|---------|---------|---------|----------|
| Cash and balances with central banks | | | | x | | |
| Financial assets held for trading | x | | | | | |
| thereof: trading derivatives | | | | | | |
| Other financial assets at fair value through profit or loss | | | x | | | |
| Available-for-sale financial assets | | | | | | x |
| Loans and receivables | | | | | | x |
| Held-to-maturity investments | | | | | | x |
| Hedging derivatives | | | | | | |
| Other assets | | | | | | |

| Panel B – Presentation of assets by product type [Swedbank (2009)] |
|-----------------|--------|--------|---------|---------|---------|----------|
| Cash and balances with central banks | | | | x | | |
| Treasury bills and other bills | x | | | | | |
| Loans to credit institutions | | x | | | | |
| Loans to the public | x | | x | | | |
| Bonds and other interest-bearing securities | | x | | | | |
| Shares and participating interests | x | x | | | | x |
| Derivatives | | | | | | x |
| Other assets | | | | | | |

| Panel C – Presentation of assets by investment purpose [Commerzbank (2009)] |
|-----------------|--------|--------|---------|---------|---------|----------|
| Cash reserve | | | | | | x |
| Claims on banks | x | x | x | | | |
| Claims on customers | x | x | x | | | |
| Derivatives for hedging purposes | | | | | | x |
| Assets held for trading purposes | x | x | | | | |
| Financial investments | x | x | | | | x |
| Other assets | | | | | | |

Table 2 – Representative balance sheet formats by European banks
Experimental research in accounting has used labeling effects to explain differing receptions of financial statement information, depending on the presentation of the underlying event. Hopkins (1996), for example, provides evidence that the financial statement classification of hybrid financial instruments affects the stock price judgments of financial analysts by priming either equity- or debt-related characteristics of the hybrid instrument. Maines and McDaniel (2000) demonstrate in a more general setting that investors’ use of comprehensive income information largely depends on how income is presented, i.e., the presentation format. Koonce et al. (2005) find that labels, attached to financial instruments with identical underlying net cash flows and risk, influence the risk associated with each instrument. In particular, they find that labels indicating derivatives increase risk perception. They explain this effect with the extensive media coverage of catastrophic losses from financial derivatives. The label ‘swap,’ which they use in their study, makes these negative associations available to non-professional investors and leads to an increased risk perception [Koonce et al. (2005)]. This effect is mitigated when additional labels explicitly indicate a hedging purpose as opposed to a speculative purpose of the swap. These results conform to Weber et al. (2005) who find that risk perception in general is significantly affected by an asset’s name. Bodnar and Gebhardt (1999) report survey evidence that managers are aware of investors’ and analysts’ negative associations when confronted with an entity’s use of derivatives. Overall, accounting research suggests that presentation formats or balance sheet items provide labels, which can cause association-based errors.

Considering these findings, it is very likely that the different balance sheet items resulting from the application of IAS 39’s mixed accounting model affect investors’ perceptions of reporting entities. In an experimental study aimed at testing the effect of balance sheet categories on risk perception, we find evidence that non-professional investors perceive ‘financial assets held for trading’ and ‘financial assets at fair value through profit or loss’ as more risky than other categories [Bischof and Ebert (2009)]. In particular, participants in these experiments link fair value measurement to investments in derivatives. Thus, negative associations with the use of derivatives carry over to balance sheet items. It is important to note that the biases in risk perception do not disappear when additional footnotes provide information about the nature of the underlying financial assets [Koonce et al. (2005)].

Earnings and capital management

There is a vast amount of empirical evidence in the accounting literature concerning earnings management [Roychowdhury (2006); Graham et al. (2005); Nelson et al. (2003); Healy and Wahlen (1999)]. The choice between fair value measurement and amortized cost accounting affects net income and shareholders’ equity (including revaluation reserves from unrealized gains and losses on AFS securities). Due to the link between financial accounting and the determination of regulatory capital, this choice also has an impact on capital adequacy ratios [Barth and Landsman (2010); Bushman and Landsman (2010)]. Thus, management incentives for choosing fair value measurement may result from either earnings targets or from regulatory capital restrictions that are tied to accounting measures (or both). The banking literature has documented evidence for the relevance of both motivations. Beatty et al. (2002) show that the avoidance of earnings declines drives accounting choices of publicly listed commercial banks. Shen and Chih (2005) demonstrate the importance of traditional earnings targets, such as the zero earnings threshold, for commercial banks. At country level, the quality of supervisory institutions, which externally monitor bank disclosures, plays a critical role in the restriction of excessive earnings management [Shen and Chih (2005); Bushman and Williams (2009); Fonseca and González (2008)]. Particularly, in a mixed accounting model the desire to mask poor performance may steer a bank management’s decision to adopt a specific measurement base which has a positive effect on the bank’s share price (assuming functional fixation of market participants). Additionally, boosting reported earnings might benefit management’s compensation. Yet another incentive to choose particular measurement bases may stem from the desire to maintain regulatory capital restrictions and avoid the risk of regulatory costs. The latter could be triggered by violations of such restrictions and subsequent supervisory actions. Ramesh and Revsine (2001) provide evidence that accounting choices at the first-time adoption of SFAS 106 and 109 are associated with a commercial bank’s capital strength, and there is broad evidence that commercial banks use ample accounting discretion to manage capital ratios [Beatty et al. (1995)]. Taken together, the evidence suggests that the accounting choices offered by IAS 39’s mixed accounting model are likely to facilitate opportunistic reporting behavior rather than to produce decision-useful information.

The opportunity to engage in earnings management by exploiting fair value measurement hinges on the liquidity of the underlying assets and liabilities. SFAS 157 and IFRS 7 distinguish between three levels of fair value measurements:

- **Level 1** – unadjusted quoted prices in active markets for identical assets (pure mark-to-market).
- **Level 2** – valuation models using inputs that are based on observable market data, either directly (as prices) or indirectly (derived from prices).
- **Level 3** – valuation models using inputs that are not based on observable market data (pure mark-to-model).

There are opposite incentives for companies to apply fair value accounting if assets are illiquid. On the one hand, mark-to-model accounting provides management with more discretion to manage accounting numbers upwards and to avoid write-offs [Laux and Leuz (2010)]. On the other
hand, recent empirical evidence suggests that share prices experience substantial discounts for an investment in level 3 assets and liabilities [Song et al. (2010)]. When financial markets dried up during the financial crisis, fair value measurement at level 3 gained importance for both assets and liabilities. Data from 36 U.S. banks shows that assets and liabilities gradually shifted from level 1 to levels 2 and 3 of the fair value hierarchy during the 2007-2008 financial crisis (Figures 2 and 3).

### Potential consequences of IFRS 9 and Basel III adoption

**IFRS 9**

Bank regulators and politicians heavily criticized fair value accounting for having accelerated the 2008 banking crisis. The G-20 leaders requested the IASB and the FASB to significantly improve accounting standards at the London Summit in March 2009 and the Pittsburgh Summit in September 2009. As a result, the IASB proposed a new accounting standard (IFRS 9) set to replace IAS 39. IFRS 9 addresses several of the concerns raised by banking regulators [Basel Committee on Banking Supervision (2009)]:

- Arbitrary rules such as the held to maturity tainting rule are eliminated.
- Hedge accounting rules, particularly the testing for effectiveness, are substantially simplified and are more directly related to a bank’s risk management practice.
- Loan losses will be recognized earlier under the proposed expected loss regime.
- Reclassifications from the fair value to the amortized cost category are permitted if a bank changes its business model.
- Disclosure formats are more standardized.

Most importantly, however, the newly proposed rules manifest a mixed accounting model rather than introduce a major change of accounting regime. The two different measurement bases will continue to coexist on bank balance sheets. Amortized cost remains relevant for debt instruments (both loans and securities) not held for short-term profit-taking. This corresponds largely with the classification of instruments as loans and receivables or as held-to-maturity under former IAS 39. All other assets (most importantly financial derivatives, equity instruments, and other trading securities) shall be measured at fair value through profit or loss. The only exception is introduced for equity investments, the gains and losses of which can be recognized in OCI rather than in profit or loss. The latter option is equivalent to accounting for equity investments under IAS 39’s available-for-sale category. The choice of the fair value option is still possible for both assets and liabilities in order to reduce inconsistencies potentially arising from accounting mismatches.

In consequence, the new rules will most likely not alter accounting practice of banks fundamentally. Measurement at fair value through OCI will lose some importance since it is not eligible for debt instruments any more. However, the traditional balance sheet structure with amortized cost being by far the most important measurement base for assets and liabilities will not be affected by IFRS 9 adoption. For this reason, the shortcomings of the current mixed accounting model for financial instruments which we have identified above seem not to be adequately addressed by the new approach. The FASB has most recently proposed a draft U.S. standard for financial instruments (Topics 815 and 825) which would result in financial accounting of banks undergoing a more radical change by implementing a full fair value model.

**Basel III**

As outlined above, capital regulation provides incentives for banks’ accounting choices. Consequently, the adoption of the new regulatory framework issued in December 2010 by the Basel Committee on Banking Supervision is likely to affect the practical application of the mixed accounting model. The capital management incentives stem from the link between financial reporting and banking regulation. For example, accounting numbers determined in accordance with IFRS are the basis for
the calculation of tier 1 and tier 2 capital. Prudential filters adjust balance sheet figures for regulatory purposes. For example:

- **Cash flow hedge reserve** – positive amounts should be deducted and negative amounts should be added back to tier 1 capital.
- **Cumulative gains and losses due to changes in own credit risk** – if the fair value option is applied for financial liabilities, all unrealized gains and losses that have resulted from changes in the fair value of liabilities that are due to changes in the bank’s own credit risk are derecognized from tier 1 capital.
- **Securitization transaction** – gains on sales related to a securitization transaction are derecognized from tier 1 capital.

These prudential filters affect the economic consequences of a bank’s choice, for example, to apply hedge accounting, to measure financial liabilities at fair value, or to engage in securitization transaction. In the end, regulatory capital remains to be vulnerable to management through reporting choices such as the choice of a measurement base.

**Conclusion**

Despite the controversial debate about fair value measurement, the mixed accounting model for financial instruments is primarily based on amortized cost rather than on fair value. Only a few banks with major investment banking activities (i.e., Deutsche Bank in Europe, Goldman Sachs in the U.S.) apply fair value accounting for a majority of assets. On average, fair value measurement is relevant for only 25% of total assets (book value) of large U.S. banks, 21% of total assets of IFRS-adopting banks from developed countries, 16% of total assets of small U.S. banks, and 13% of total assets of banks from emerging and developing economies.

The diversity of measurement bases has not only implications for the measurement of income and equity, but also for presentation and disclosure of financial assets. The link between measurement and disclosure results from the IFRS 7 requirement to present financial assets and liabilities by classes: such a class can be defined in accordance with the measurement categories. As a consequence, economically identical transactions potentially result in differences of both net income and balance sheet presentation if two banks choose different measurement categories. Basically it is this inconsistency for which IAS 39 is criticized from a theoretical perspective, because the inconsistent accounting treatment might produce biases in investors’ risk perception and facilitate opportunistic earnings management. It does not seem like the drawbacks will be substantially mitigated when the current reform of IAS 39 is completed and IFRS 9 is eventually adopted, because the newly proposed rules are more a cosmetic than a fundamental change of accounting for financial instruments.

**References**

- Basel Committee on Banking Supervision, 2009, “Guiding principles for the replacement of IAS 39”
Indexation as Primary Target for Pension Funds: Implications for Portfolio Management

Abstract
The current financial crisis has strongly affected the financial status (expressed by the funding ratio) of most pension funds and their ability to grant full indexation of liabilities to the inflation rate. The indexation benefits represent a priority for a participant of a pension fund bearing purchasing power risk. Differently from return-oriented optimization, we define an objective function based on the indexation decision, which is conditional on the financial status of the fund. This paper focuses on the definition of an ALM model aimed to maximize the indexation granted by a defined benefit pension fund to its participants by introducing real assets in the portfolio, but also imposing risk-based regulation. The model is applied to the portfolio of the ABN AMRO pension fund, which aims to fully index its liabilities with respect to the Dutch inflation. The results suggest the initial funding ratio strongly affects the ability of the fund to set a fully-indexed investment strategy over longer time horizons, while changes in the risk aversion parameter have a limited influence. A crucial role is played by property, which is preferred to equity, while commodities have risk diversification properties exploitable only in the long run for higher funding ratio, otherwise the regulatory framework takes the form of a barrier to invest in commodities.

Angela Gallo — Department of Business Administration, University of Salerno
The recent turmoil in the financial market sets even more challenges in terms of performance for pension funds, among the major investors in the stock markets. These challenges must be added to the difficulties already faced by these investors during the past few decades (in particular during the pension crisis of the 2000-2003), because of the strong reduction in equity premiums, the decline in long-term bond rates, the aging of the population, the stricter supervision adopted by the regulators, and the accounting innovation in terms of fair valuation of the liabilities. These days, interest rate risk, equity risk, longevity risk, and inflation risk have to be taken into account in the definition of the investment policies as crucial risk-drivers for solvability. As for a defined-contribution pension fund (DC), the impact of the financial crisis depends critically on the pension fund’s asset allocation and its member’s age, for a defined-benefit pension fund (DB), the main concern is the reduction of the funding level. The retirement income provided by defined benefit pension plan is in principle unaffected by changes in investment return, but lower asset prices worsen their financial solvency. In the last year, many of the DB pension funds in the OECD countries reported lower funding levels and in some cases large funding gaps [OECD (2009)]. Whereas the impact of the financial crisis is not such as to harm the solvability of a DB pension fund, the reduction of the funding levels has as a main consequence a reduction in the indexation granted to pension fund participants until funding level recovers. The indexation represents a correction of the pension rights aimed at compensating the loss in terms of purchasing power due to price or wage inflation and therefore offers a hedge against the purchasing power risk faced by pension fund participants. For the last few decades, the full indexation of liabilities has been an undisputed guarantee offered to the participants of a pension fund, but it has become less sustainable for many DB pension funds since the 2000-2003 stock market collapse. Most of them opted for voluntary and conditional/limited indexation policies. Depending on the financial position of the fund, the compensation can also be null or only partial when the funding ratio falls below required level. In the U.K., indexation is typically restricted to the range of 0%-5% per year (limited indexation). In the Netherlands, pension funds mostly opted for a solution consisting of conditional indexation to Dutch price/wage inflation. The decision to grant indexation depends on the nominal funding ratio, defined as the ratio of assets to liabilities. If the funding ratio falls below a specific threshold level, indexation is limited or skipped altogether, assuming the features of an option [de Jong (2008)]. However, even if not explicitly stated in the pension contracts, most of the Dutch pension funds state that the maximum price or wage indexation is aimed for [Bikker and Vlaar (2007)].

From a participant’s perspective, the conditional indexation policy implies that the ‘indexation risk’ (or purchasing power risk) partly translates from the pension fund to its participants. This solution has been strongly rejected by pension fund participants, given the worldwide recognized assumption that pensioners aim to keep constant their standard of living after retirement [Modigliani (1986)]. Inflation risk can strongly impact the pension rights accrued during the working years, resulting in a loss of the purchasing power of their savings at the retirement. To get an idea of the impact of inflation over a long time horizon, take the following example: over a ten-year period, an average price inflation of 3.21% will correspond to a loss of €271 in the purchasing power of a pension right of €1000. Since most pension fund investment horizons are around 40 years, it is clear that indexation policy is very important. Even if the recent financial crisis has reduced the acceleration of inflation rate, we could expect it to rise in the future due to the anticipated increase in demand for food and energy resources. However, inflation risk is not the only risk-driver of indexation risk. As conditional indexation is often defined in terms of the nominal funding ratio, the indexation risk is a combination of inflation, interest rate, and market risks, affecting both the market value of the liabilities and of the assets. Both decisions on the assets side (investment strategy) and liabilities side (type of pension system) affect the capability of the fund to grant indexation. Decisions concerning the liabilities side used to remain mostly unchanged during the life of the fund, so that asset allocation decisions were actually made with the purpose of managing indexation risk. Given the importance of the purchasing power of pension rights and the recognized social and political role played by pension funds, we suggest that the new definition of the ‘pension deal’ in terms of risk-sharing also implies a new definition of the criteria underlying the asset allocation of the funds. Moreover, in the recent financial crisis, we have seen pension funds become preoccupied with reducing their exposure to highly risky investments and forced to sell part of their equity holdings, even at a loss, to meet regulatory standards. This is the result of the current return-oriented pension management. It has the drawback of being highly procyclical: during economic expansion the pension fund is willing to bear more risk to obtain higher return, but when there is a downturn it leads to severe losses and consequentially to the reduction of the indexation at the expense of the pension fund participants.

In this paper, we will focus on a pension fund’s asset allocation model, with the clear objective of maximizing the indexation of the liabilities, since it is the full indexation of the pension rights, and not the maximization of returns, that is the main priority. The fund’s ability to reach this target will be examined via the introduction of real assets, such as property and commodities, into the portfolio. These assets have been shown to offer valuable inflation hedging properties. We will also impose constraints to maximization according to the annual risk-based regulatory standards. There are two reasons for their inclusion. Firstly, most pension funds have to also take into consideration the rules imposed by the supervisory authorities in their internal analysis. Secondly, we want to test if the adoption of risk-based standards can harm the exploitability of the inflation hedging properties of real assets.

Through a definition of a simulation/optimization model in an asset and
liability management (ALM) context, we define an objective function represented by the indexation decision, conditional on the nominal funding ratio. We use the traditional mean-variance framework [Markowitz (1952)] combined with a simulation model as in Boender (1997). We optimize the pension fund’s portfolio following the ‘liability driven investment’ (LDI) technique promoted by a number of investment banks, such as Morgan Stanley, over the past few years. The LDI model divides the portfolio in two parts. The first part, the matching portfolio, must be able to meet the nominal liabilities over time, adopting duration matching strategies. We set the matching portfolio to earn a return equal to the nominal growth of the liabilities. We consider this portfolio as an ideal asset perfectly correlated with the liabilities. The second part (the risk-return portfolio) is composed of return-seeking assets, which are represented by equities, property, and commodities. Maximizing the objective function will give us the optimal asset allocation in terms of the amount of resources to be invested in the matching portfolio and in the risk-return portfolio, but also the type of return-seeking assets to be included in the latter portfolio. The optimal portfolios are those that are able to maximize the purchasing power of the pension rights of the participants. The analysis compares the optimal portfolios for different investment horizons, risk-aversion levels, and initial funding ratios and quantifies the indexation loss associated with each portfolio.

The model is applied to the real case of the ABN AMRO BANK pension fund. ABN AMRO kindly provided us with the scenarios of the relevant economic time series, the nominal payments they have to face in the future, and the conditional indexation rule. The pension fund only guarantees the nominal payments, but is willing to provide full indexation of their future payments. We assume a ‘liquidation perspective,’ in that the pension participants and the invested assets are fixed at 2009 and will not be increased by new contributions. We start from assuming an initial funding ratio level of 110 (equivalent to 10% surplus of assets on liabilities). Afterwards, wealthier positions are considered by setting the initial funding ratio at 120 and 130. We expect that the richer the fund is, the more the indexation policy is sustainable and its capability to meet solvency constraints. On the other hand, when the funding ratio is relatively low, we expect the fund to take even more risk to meet its nominal obligations (and constraints) and the provision of the indexation. The optimal portfolio will also depend on the levels of the risk aversion parameter.

In our model, this parameter represents a penalty to the volatility of the indexation decision, implicitly corresponding to the possibility of indexation cuts. We expect that a higher risk-aversion parameter could lead to the selection of a safer asset mix when the funding ratio is relatively high, and vice versa. As in the traditional analysis, this risk aversion parameter could be considered as a proxy for the fund’s flexibility to react to other variables such as extra-contribution by the participants or financial support of sponsors. Finally, a third dimension is represented by the investment horizon of the optimization. A pension fund is typically considered to be a long-term investor due to the long maturity of its liabilities. However, the Dutch risk-based supervision on pension funds (FTK), similar to other countries that adopt risk-based regulation, imposes solvency constraints on the one-year probability of underfunding (probability of funding ratio below 100). Consequently, both the short- and the longer-term time horizons have to be taken into account simultaneously. We investigate the 3, 5, and 10 year time horizons. These time horizons do not correspond to the long-term horizon of the liabilities, which is about 40 years, but the definition of the asset allocation tends to be much shorter. This kind of analysis helps us get a better understanding of the inflation hedging properties of the assets in the portfolio at different time horizons and how they can be exploited under regulatory constraints.

The ALM literature initially focused on mean-variance single-period optimization analysis, focusing either on the optimization of the surplus (difference between asset and liability values) or the ‘universal’ measure represented by the funding ratio return [Leibowitz et al. (1994)]. Successively, the analysis was extended to consider the long-term nature of the pension fund, with the imposition of adequate short-term risk constraints to the maximization of the funding ratio. Recently, ALM studies mostly apply operations research models to optimize funding and investment policies under uncertainty [Ziemba and Mulvey (1998)]. Using stochastic programming techniques, they assume as objective function the end-of-period wealth of the funds, or the minimization of the risk of underfunding, and impose as constraints several requirements with respect to solvency, contribution rate, and indexation policy. Moreover, several models have been developed using chance constraints to limit the probability of underfunding for the following years [Dert (1995)] or assuming measures of underfunding risk such as the conditional value at risk [Bogentoft et al. (2001)]. In this field, Drijver (2005) was the first person to formalize indexation decisions, although in a rough way, considering the conditional indexation policy in the objective function as a penalty associated with not giving full indexation. The main difference of this model relative to our analysis is that they assume unconditional indexation to be one of the constraints that the pension fund has to meet. In our model we consider the indexation decision, conditional on the funding ratio, as the objective function, setting a direct link between the definition of the optimal portfolio and the purchasing power of the participants. This represents the novelty of our approach. Moreover, we adopt a simulation-optimization model as in Boender (1997), avoiding the complexity of these previous works related to the need for analytical solutions. Our analysis also differs with those mentioned above because we consider the inclusion of real assets in the pension fund portfolio next to traditional assets such as bond and equity. In the aforementioned papers, the debate was mainly focused on investing between bonds and equities [Benartzi and Thaler (1995)]. Most of the literature on pension funds focuses on portfolios composed of bonds, equities, and cash. They typically find that there is a preference for equities in the long run, mainly due to its mean
reversion effect, which many believe makes it a safer investment than bonds [Campbell and Viceira (2005)]. However, the equity premiums have now been dramatically reduced, while bonds are important for hedging the interest rate risk arising from the new market-to-market valuation of the liabilities. Moreover, empirical evidence [Fama and Schwert (1997)] suggests a negative relationship between expected stock returns and expected inflation. This result seems to be consistent with the view [Fama (1981)] that high inflation has a negative impact on economic activity and thus on stock returns. On the other hand, stock dividends are positively impacted by higher future inflation [Campbell and Shiller (1988)] and this means that they can offer partial inflation hedging protection in the long run. Dert (1995), however, found a negative correlation between stock returns and Dutch price inflation. Regarding the inflation property of the bonds, by definition, we expect a positive long-term correlation between bonds returns and changes in inflation, while in the short run we expect lower or negative correlation, due to deviations between realized and expected inflation.

Among real assets, commodities are generally considered to be leading indicators of inflation. In fact, they are among the main drivers for increases in inflation, especially agriculture, minerals, and energy. As shown in Gorton and Rouwenhorst (2006), commodities futures show good hedging-inflation properties in long and short run, having a positive correlation with inflation, which increases with the holding period and that is greater when annual or 5-year frequencies are considered. Real estate investments also allow for enhanced inflation protection as showed in Fama and Schwert (1997), and this effect is particularly significant over longer time horizons. Moreover, as argued in Froot (1995) and Hoevenaars et al. (2008), real estate investments behave quite similarly to stocks, showing good inflation properties, except that they are not as beneficial when it comes to risk diversification of the portfolio. Recent works also consider the inclusion of derivatives instruments in the portfolio of a pension fund to hedge both interest rate and inflation risks. In particular, investments in interest rate swaps and inflation swaps are increasingly attracting pension fund managers. Interest rate swaps are suitable alternatives for investing in nominal bonds to manage interest rate risk, due to the higher liquidity of the corresponding market. Inflation swaps are considered to be suitable alternatives for inflation hedging strategies. Moreover, they are viewed to be better investments than inflation-linked bonds because they can offer a better return performance. However, there is still reluctance about both of these instruments because at the moment the capacity of the inflation-linked security markets is not sufficient to meet the collective demands of institutional and private investors. As for the over-the-counter markets, they suffer from a perceived increase in counterpart risk. For these reasons, our analysis will exclude the derivatives since these instruments are still not common in the pension funds’ portfolios and the literature about their use is still at an early stage.

Closer to our work is the paper by Hoevenaars et al. (2008), who analyze a diversified portfolio in an asset and liability context. They construct an optimal mean-variance portfolio with respect to inflation-driven liabilities based on a model implying forward looking variance and expected returns. They find that alternative asset classes add value to the portfolio: commodities are good risk diversifiers, even for long-term investments; stocks are inflation hedgers in long run; hedge funds can offer return enhancement, and listed real estate behave closely like stocks. Our paper differs from that of Hoevenaars et al. (2008) in that we optimize using a different objective function, and more importantly, we use a scenario-based approach which corresponds to a simplified version of the optimization model actually adopted by pension funds managers in their internal decision-making process. Moreover, we choose to impose current risk-based regulatory constraints, which can heavily impact asset allocation decisions and the pension funds’ abilities to exploit inflation hedging properties in the long term.

Our analysis can be somehow considered partial. A real pension fund is characterized by multiple competing objectives defined as risk-budgeting [Boender and Vos (2000)], while our stylized pension fund solely aims for maximal indexation with respect to the short term regulatory rules. That is to say, it does not take into account, for instance, the contribution policy. However, as in Siegmann (2007), we can invoke the 1-1 relation of the indexation policy (conditional on the financial position of the fund) with the funding ratio. If the funding is high, the constraints are satisfied and also the contribution level can be lowered (and vice versa). The results suggest that the sustainability of the indexation-based portfolios is easily affordable over a short time horizon, even if the full indexation can be reached only at higher level of the initial funding ratio. The initial funding ratio strongly impacts the capability of the fund to set an indexation-based investment strategy over the longer time horizon. This can be easily explained considering the cumulative effect of the indexation policy: once the indexation is granted, it is permanently part of the nominal liabilities which will be eventually indexed the next year and so on. This cumulative effect requires higher returns on the portfolio to cover the increasing ‘real’ growth of the liabilities. Another important evidence is the limited impact of different risk aversion parameters in the definition of the compositions of the optimal portfolio. Concerning the composition, there is a convergence in the results towards a portfolio composed of the matching portfolio (around 88-90%), property (8-9%) and a residual part in equity (1-2%). There is no significant role for the typical inflation hedger assets, such as commodities and equities. Property represents a better investment opportunity than equity over every time horizon. This composition changes when riskier strategies are needed to reach higher levels of indexation. In this case, there is a significant shift of resources from the matching portfolio to property. Commodities are included in the portfolio only over the longest time horizon and when the fund has a solid initial financial position. These results reveal that for a Dutch pension fund,
Indexation as Primary Target for Pension Funds: Implications for Portfolio Management

The Capco Institute Journal of Financial Transformation

Indexation in the pension industry, we define the indexation rule as follows: Consistent with the actual implementation of the conditional indexation of the amount of resources it has to cover the related nominal liabilities. The ultimo funding ratio – represents the financial status of the fund in terms of the end of the year t (FR). It is computed using the annual market values for assets and liabilities exposed to relevant market risk factors. The indexation policy depends only on interest and inflation rate dynamics, while the asset flows are independent of the participants' actions. In our stylized fund, the number of indexation is fixed and the invested collected assets will not be reduced by contributions (run-off hypothesis); they will only change due to changes in portfolio returns. Consequently, the flows from the liabilities only face interest and inflation rate dynamics, while the asset flows are exposed to relevant market risk factors. The indexation policy depends on the financial status of the fund expressed by the funding ratio at the end of the year t (FR). It is computed using the annual market values for both assets (A\textsubscript{t}) and liabilities (L\textsubscript{t}): FR\textsubscript{t} = A\textsubscript{t} / L\textsubscript{t} (1), where (FR\textsubscript{t}) – the ultimo funding ratio – represents the financial status of the fund in terms of the amount of resources it has to cover the related nominal liabilities at the end of the year.

Consistent with the actual implementation of the conditional indexation policy in the pension industry, we define the indexation rule as follows:

- If the funding ratio is greater than the required funding ratio, full indexation is granted and previously missed indexation is recovered. The required funding ratio is defined by the pension law and depends on the strategic asset allocation (SAA) of the fund and on its duration mismatch between pension assets and liabilities. For simplicity we assume the required funding ratio to be equal to 115, which approximates the average required funding ratio for the Dutch pension funds market.

- If the funding ratio is smaller than 105 (minimum solvency funding ratio) the nominal liabilities at time t corresponds to the nominal liabilities at time t-1 (no indexation).

- If the funding ratio is between 105 and 115, partial indexation is granted.

To model this indexation rule, we define δ\textsubscript{T} as the indexation decision at time T, which will assume the value of 1 for full indexation, 0 for no indexation, and a value between 0 and 1 if partial indexation is granted, depending on the financial status of the fund in terms of ultimo funding ratio (FR\textsubscript{T}). We want to maximize the expected value of δ\textsubscript{T} (delta) at a certain horizon T corresponding to relevant investment horizons. The maximization of delta is based on the definition of the amount of the resources to invest in each asset class j (no short selling) included in the portfolio. The model is static: over the time horizon T, the asset allocation is kept constant, that is to say, there are no policy changes between 0 and T. As in a mean variance framework, the maximization of the expected value of delta is associated with a penalty consisting of the variance of the delta. Higher volatility of delta penalizes the utility associated with the indexation. As it has also been criticized in Leibowitz et al. (1994), where the objective function is represented by the funding ratio return, the mean-variance model has the drawback that it does not consider that a pension fund is more sensitive to downside risk measures than to symmetric measures of risk (such as the variance). Also for our objective function, this consideration is valuable. The pension fund is sensitive only to the risk of not being able to grant the indexation (indexation cuts). However, due to the complexity of the mean-shortfall model, and to let our model be numerically tractable in a simple way, we use this symmetric measure of risk. The formulation of the optimization problem is given by: Max_{\omega, \delta} E(\eta) - \gamma \cdot V(\delta) (2); P(\text{FR}_{t+1} > 105 | FR_t) > 0.975 (3), where γ (gamma) is the risk-aversion parameter of the pension fund.

As a constraint of our analysis, we consider the condition on the solvency as promoted by FTK. However, even though we refer to the Dutch regulatory framework, there is no loss of generality in our model since recently more and more countries worldwide are evaluating the opportunity to implement more sophisticated risk-based standards following the Dutch experience [Brunner et al. (2008)]. FTK sets a constraint on the minimum required solvency in the short term. It imposes that every year the funding ratio should be such that the probability of underfunding in the next year is smaller than or equal to 2.5%. We use a scenario-based ALM model as in Boender (1997) to implement the optimization described above. It is a basic version of the well-known model used in the pension funds industry to support their actual decision-making. In our work, the optimization model is implemented using a scenario-based ALM model, which allows for a more realistic simulation of the pension fund's financial situation and the potential impact of indexation decisions.
will be based on a range of possible future developments (scenarios) of the deltas, depending on the range of possible future developments of all the other economic variables such as the interest rate yield curve (and consequentially of the present value of the liabilities), the asset class returns (and consequentially of the market value of assets), which define the funding ratio values, and the inflation rates, which define the indexation levels. The expected value of delta and the variance of delta in our objective function are computed for each combination scenario-time. We will run an optimization at given time horizons $T$ (3, 5, 10 years), to determine which portfolio weights allow for the maximization of the objective function, under the satisfaction of the solvency constraint.

**Indexing the market values of assets and liabilities**

As mentioned before, the indexation decision is conditional on the nominal funding ratio. To compute this funding ratio we need to define the market value of asset and liabilities. We set the time 0 as the moment from which the pension fund is formally closed to new participants and the old ones do not make any further contributions. Every year the pension fund only has annual nominal cash flows (CFs) to be paid to the participants at the end of each subsequent year until the definitive closing date ($n$). The present value at time $t$ of all these future nominal obligations is computed market-to-market as: $L^U_{t} = \delta_{t} \sum_{n=0}^{\infty} CF_{t+n}(1+i^k)^n$ (4), where $k$ is the maturity of each residual cash flow and $i^k$ is the spot rate associated to the corresponding node on the interest rate yield curve. The notation $L^U_{t}$ stands for the ultimo value of the liabilities and accounts for the fact that the present value is calculated on the basis of a yield curve estimated at time $t$. The interest rate yield curve is generated by the well-known Nelson and Siegel (1987) model, fitted via a least-square curve estimated at time $t$. The interest rate yield curve for the liabilities, which also takes into account the eventual indexation decision: $Lt = \sum_{n=0}^{\infty} (1+i^k)^n$ (5), where $i^k$ is the inflation rate recorded at time $t$ and $\delta_t$ is the spot rate at $t$.

The nominal ultimo value is used to construct the ultimo funding ratio used to make the indexation decision at the end of the year. Depending on the value of the funding ratio at time $t$, the indexation decision is taken and the ultimo value of liabilities is calculated. The formula to obtain the ultimo indexed value of liabilities is as follows: $L^U_{t} = L^U_{t} \cdot \delta_t \cdot (1+\pi_t)$ (6), where $\pi_t$ is the inflation rate recorded at time $t$ and $\delta_t$ is the variable which allows us to consider a more complex indexation decision, that is to say, a decision which also take into account recovery and partial indexation. It is defined as: $\delta_t = 1/(1+\pi_t + F(x)[1/\delta_{t-1} - 1/(1+\pi_t)])$ (7), where $F(x)$ is modeled as a logistic function and introduces the conditionality of the indexation decision on the ultimo funding ratio (see Figure 2 below): $F(x) = 1/(1+e^{c-x/y})$; $c=1; x = FR^U - 110$ (7). Delta tilde is properly linked to delta, our objective function, as follows: $\delta_t = \delta_{t-1} \cdot 1/(1+\pi_t)$ (8). This mathematical framework allows us to replicate the dynamics of the liabilities when a conditional indexation policy is adopted. To show how this model works, let us consider an ultimo funding ratio smaller than 105 at the end of the year $t$. Since we set the reference ultimo funding ratio at 110 in the logistic function, $F(x)$ is equal to zero in formula (7). In this case, $\delta_t$ becomes equal to $1/(1+\pi_t)$, and in the formula (5) the indexation is canceled out. Then, $\delta_t$ in formula (8) will assume a value equal to $1/(1+\pi_t)$ if full indexation was granted in the previous year ($\delta_{t-1} = 1$), otherwise it assumes a value equal to $\delta_{t-1}/(1+\pi_t)$, including the information about the missing indexation in the current year and in the previous year.

Similarly, if the ultimo funding ratio is above 115 the value of the logistic function is 1 and the $\delta_t$ is equal to 10/(1+\pi_t). If $\delta_{t-1}$ is equal to 1 (full indexation in all the previous periods), the full indexation is granted because $\delta_t$ assumes value 1 in formula (6). However if $\delta_{t-1}$ is below 1, as in the cases above (funding ratio in the previous year is below 105), $\delta_t$ assumes value equal to $(1+\pi_{t-1})$, which corresponds to also recovering also the indexation missed in the previous period; otherwise it is equal to the product of all the missing indexations $\prod_{i=0}^{t-1}(1+\pi_{t-1})$, where $i$ represents all the previous periods with missing indexation. Finally, for values of the ultimo funding ratio between these two thresholds, the logistic function assumes a value of between 0 and 1. In these cases we have partial indexation granted and hence, partial missing indexation is to be recovered. For example, when the ultimo funding ratio is equal to 110, the logistic function assumes a value of 0.5 and only half the indexation will be granted. Delta and delta tilde will take into account that only partial indexation has to be recovered. Once the ultimo indexed value of liabilities is determined, by subtracting the corresponding cash flows to be paid at the end of the year (also updated by indexation decision), we compute a primo value for the liabilities, which also takes into account the eventual indexation decision: $L^U_{t} = L^U_{t} \cdot \delta_t - [CF_t \cdot (1+\pi_t)]$ (9). This value represents the end of the year post indexation and payments and corresponds to the initial value of the liabilities for the next year (but discounted with the interest rate yield curve at time $t$).

On the other side of the intermediation portfolio, for each time $t$, according to the liability driven investment (LDI) paradigm, the asset portfolio ($A_t$) is divided into two sections: the matching portfolio ($A_{M,t}$) and the risk-return portfolio ($A_{RR,t}$). The matching portfolio is assumed to earn exactly the liability nominal growth (liability return $\pi_t$) to match nominal liabilities as a result of a perfect immunization strategy. The liability return represents the variation in the ultimo value of the liabilities from one year to another only due to interest rate yield curves and cash flows dynamics (only in nominal terms). The risk-return portfolio consists of different asset classes such as equities and alternative assets. It is meant to provide enough resources to grant indexation by means of the inflation hedging properties of these assets. The amount invested in each portfolio is enough resources to grant indexation by means of the inflation hedging properties of these assets. The amount invested in each portfolio is estimated as: $w_{M} = \Delta A_{M,t} / A_{t}$ and of the risk-return portfolio to the total value $w_{RR,t} = A_{RR,t} / A_{t}$.

Consistent with the liabilities framework, we define two different values of the assets. The first one, defined as the ultimo asset value ($A_{U,t}$), is the reference value for the computation of the nominal ultimo funding ratio.
on which the indexation will depend on. It is computed as: $A_{t}^{U} = \tilde{A}_{t-1}^{P} \cdot (1+\pi_{t-1})$ (10). It expresses the value of the invested assets before the indexation decision is taken and the payment of the cash flow for the corresponding year is made, while is the primo value for each portfolio. Similar to the primo value of the liabilities, it is computed as: $A_{t-1}^{P} = A_{t-1}^{U} - [\tilde{C}_{t-1} \cdot \tilde{b}_{t-1} \cdot (1+i_{t-1})]$ (11). The primo values of the assets and liabilities are used for defining the constraints at each year relative to the next and it is obtained excluding the cash flow (eventually updated to indexation) that has to be paid in the corresponding year.

The outcomes of our optimization model are about the definition of the ‘optimal’ asset allocation $(w_{M}, w_{r-r})$ of the resources between the two portfolios and within the risk-return portfolio, able to maximize the indexation decision.

**Data**

ABN AMRO Pension Fund provided us with a unique dataset composed of annual data of assets returns, interest rates, and price inflation Nelson-Siegel parameters as endogenous variables generated by a Vector Autoregressive Model (VAR). Based on this dataset, we generated a total number of $q$ scenarios equal to 2500 for all the variables in our model for the period 2009-2022 on an annual basis to determine the value of delta in each combination scenario-time. On the asset side, the asset returns are generated for commodities (GSACHI Index), Property (ROZ/IPD Dutch Property Index), equity growth (MSCIWI), equity value (MSCISWI hedged), and emerging market equities (MSCI Emerging Markets Index).

On the liability side, we make use of an original dataset composed of all the residual cash flows from 2008 to 2022 on the assumption that in 2022 the ABN AMRO Pension Fund would close. This was estimated using actuarial simulations that are properly linked to the other simulated economic times series. The present value of the liabilities generated by the interest rate yield curve has an expected long-term annual growth of 5.71%, while the standard deviation is around 12%. This annual growth is defined as liabilities return. In general terms, it means that to reach the full indexation of the liabilities in this ALM context, the invested assets available at beginning of 2009 must be ideally allocated in such a way as to earn on average the annual nominal liabilities return plus an average inflation rate of around 2% without the risk of underfunding being too high.

**Empirical results: the ABN AMRO pension fund**

This section presents the main results from the implementation of our indexation-based optimization/simulation model to the ABN AMRO Pension Fund dataset. The analysis is developed along three dimensions: the risk aversion level, the initial funding ratio, and the investment horizon.

The composition of the optimal portfolios shows how resources are allocated between the matching portfolio and the risk-return portfolio to reach the highest level of indexation, which real assets are included, and how the composition changes at different time horizons and risk aversion levels, given the initial financial position of the fund. The optimization does not allow investing in short positions and is subject to the satisfaction of the solvency constraints for all the years included in the investment horizon. For each portfolio utility, expected delta, standard deviation of delta, indexation loss, and the composition of the portfolio are calculated for 3, 5 and 10 year time horizons. The utility value helps us to identify which optimal portfolios offer the best trade offs according to the mean-variance criteria. The expected delta gives the average value of delta across scenarios at a given time horizon and risk aversion level. If delta is equal to 1, the full indexation has been granted in all the previous periods, otherwise if delta is smaller than 1, the portfolio is able to ensure only partial indexation of the pension rights. The ‘distance’ from the full indexation can be defined as indexation loss associated with each optimal portfolio. Given the formulation of our model, it can be defined as $1-\delta/\delta_{F}$. For example, a value of expected delta equal to 0.98 approximately represents a loss of 2% in terms of missed indexation over the investment period. Standard deviation of delta gives a measure of the risk associated with expected delta, and consequently it is a measure of the implied risk of the investment strategy. Delta depends on the nominal funding ratio, whose volatility changes according to both liability and portfolio volatilities. Since nominal liability volatility is the same for all the portfolios, higher standard deviation of delta are due to higher volatility of the optimized portfolio. A general overview of optimal portfolios shows that the sustainability of the indexation-based optimization is easily affordable over a short time horizon, even if the full indexation can be reached only at a higher level of the initial funding ratio (120 and 130). Over longer time horizons, when the funding ratio is 110, which corresponds to a weak (but still solvent) financial position of the fund, the optimization is not able to find feasible solutions to satisfy all the solvency constraints. The initial funding ratio strongly impacts the ability of the fund to set an investment strategy over a longer time horizon. These results can be explained considering the cumulative effect of the indexation policy. Once the indexation is granted, it is permanently part of the nominal liabilities which will be eventually indexed the next year and so on. It means that if indexation is granted, a greater amount of resources is needed to match the (new) nominal liabilities in the following years, even if the solvency constraints prevent us from assuming excessive risk. A solid initial financial position is better for sustaining indexation over longer time horizons. Important evidence is the limited impact of different risk aversion parameters in the definition of the compositions of the optimal portfolio. In most cases, for different values of gamma we observe changes in the compositions of around 0.5%. For this reason, this section will solely focus on the analysis of the results for gamma equal to 10. Once the optimal portfolios are defined, we extend our analysis and impose two different constraints relative to the weight of the matching portfolio and property in the optimization. The first constrain concerns the impossibility of investing a high percentage of the available resources
in the matching portfolio, due to the imperfections of the long-term bond markets. This constraint conventionally limits the weight of the matching portfolio to be equal to or smaller than 63%. These portfolios, defined as ’MP-restricted PF,’ can be implemented only by richer pension funds (with funding ratios of greater than 120) and, due to the limited investment opportunity set, are less efficient than the optimal portfolios. However, these portfolios can actually be replicated in the financial markets. The second constraint concerns investments in property. As mentioned before, this asset is a valuable asset because of its low volatility and high return. However, it is by definition an illiquid asset. Most of pension funds set a limit on investments that cannot be easily converted into cash, given the annual liquidity pressures of the cash flow payments. For this reason, we optimize imposing the weight of property to be equal or smaller than 15%. When both the constraints are added to the solvency constraints, feasible solutions are available only when the initial funding ratio is set equal to 130.

We begin discussing the optimal portfolios when the initial funding ratio is 120 and gamma is 10. The highest utility is associated with the shortest time horizon and decreases for longer time horizons due to the cumulative effect of the indexation. Figure 1 shows how the composition of these three portfolios changes over time. The portfolios are composed of the matching portfolio and the risk-return portfolio by property and a small contribution of equities (in particular equity growth). From these results we find that property, in contradiction to the findings of previous studies, is able to make a substantial contribution to the definition of the optimal portfolio and is preferable to equity in the short, medium, and long term. Another result, which is in contrast with the literature, is the absence of the commodities, even over a short time horizon. Their high volatility could be a threat to meeting the solvency constraints and lead to the exclusion of these assets. An interesting result concerns the distribution of the weights over the medium term. At 3 and 10 years the portfolio invests a high percentage in the matching portfolio, about 89 to 90%, 8 to 9% in property, and a residual 1 to 2% in equities. Over a 5-year horizon the composition is quite different. There is a substantial shift of resources from the matching portfolio (~22.5%) to property and equities, which increase respectively by 20.8% and 1.7%. Since the mean, the standard deviation, and the cross correlation do not show significant changes from the short to the medium term, a better insight could be derive from observing the distributions of the nominal funding ratio and delta. Table 1 shows the descriptive statistics and distributions of delta and nominal funding ratio for each time horizon. Over a 3-year time horizon, the expected delta is 1 (on average the portfolio ensures the full indexation in all the periods) associated with an extremely low standard deviation. The Figure a) shows that the distribution is within a range of high delta values, between 0.98 and 1. It has a negative asymmetry, meaning that the mass of the distribution is concentrated on the right side of the figure. It has relatively few lower values. We also consider the probability of delta greater than 0.98 and 1. These measures of probability could be considered as downside risk measures. The first one gives the probability of not losing more than 2% of indexation, while the second to have full indexation. These values for a 3-year time horizon are equal to 100% and 78%. It also means that in 1948 out of 2500 scenarios the full indexation is granted. These measures of downside risk can be helpful in the valuation of the portfolios, whereas the expected delta only gives an averaged value of the indexation. Figure b) presents the nominal funding ratio distribution for 3-year time horizons. The distribution is close to the normal distribution shape as suggested by the low values of the asymmetry and kurtosis. The mean is higher than the minimum required level for the full indexation as defined by the indexation rule, ensuring high level of indexation. Downside risk measures are computed also for the nominal funding ratio. The probability of the funding ratio being below 105 gives the probability of underfunding of the pension fund, which corresponds to the number of scenarios where no indexation is granted over the total number of scenarios. The probability of a funding ratio that is greater than 115 gives the same information as the probability of delta being greater than 1, as defined by the indexation rule. Over a 3-year time horizon these probabilities are respectively 0 and 78%.

These statistics reveal the sustainability of this investment strategy aimed at the maximization of the indexation decision in the short term. Over the medium term the optimal portfolio implies an indexation loss of about 0.03%, as well as a higher dispersion of delta, and in particular of the nominal funding ratio (Figure d). Also the downside risk measure suggests that the optimal portfolio is riskier over the 5-year horizon. The indexation maximization is obtained by adopting a riskier strategy, which could explain the shift to property and the highly risky equity. Property is less risky than the matching portfolio, but since the latter is by definition perfectly correlated with the liabilities return, it is less effective for hedging liabilities. Figure 2 shows the evolution over time of the nominal funding ratio for the three optimal portfolios and of the relative solvency constraints, assuming their compositions are kept constant (as in Figure 1) over the 10-year time horizon. For the first 3-years, all the three portfolios are able to grant full indexation and to meet the solvency requirements.
However, at T=3 the optimal portfolio is unable to offer full indexation over five years. Higher portfolio returns are needed to reach higher indexation, obtained by the introduction of equities and property and then, at higher risk. At longest horizon the Optimal Portfolio T=5 cannot be implemented because it does not satisfy all the solvency constraints. The best investment strategy is similar to the optimal portfolio at T=3.

Over the 10-year time horizon the cumulative impact of the indexation significantly impacts the liabilities, which need to be matched by a stronger investment in the matching portfolio. However, the indexation is fully obtained only in the first three years and is halved at the end of the period (T=10). This implies an indexation loss of 4.3% (delta is equal to 0.95), approximately corresponding to two years of missing indexation when the inflation rate is constant and equal to 2%. The probability of the delta being greater than 0.98 is only 26%, while the probability of underfunding rises to 2%, close to the regulatory constraint of 2.5%. These statistics express the riskiness relative to the implementation of an indexation-based optimization over a longer time horizon (with a static asset allocation) and convenience for a 3-year investment horizon. From this perspective it is important to examine how the composition of the optimal portfolio changes after 3 years, given a different initial funding ratio. It is important to note that the higher utility is associated with the assumption of a higher funding ratio. Figure 3 shows the compositions of the optimal portfolios for different initial funding ratios at the 3-year time horizon.

All portfolios are made up of the same type of assets: matching portfolio, property, and equities. At initial funding ratios of 120 and 130 the weighting of the portfolios is very similar and the portfolios reach the full indexation with a low risk. When the initial funding ratio is equal to 110, the delta is 0.976 and the volatility is significantly higher. As discussed above, we find that when the pension fund needs to invest more aggressively to reach higher level of indexation or because it starts from a weak financial position, the proportion that it invests in property is increased, as compared with the other optimal portfolios.

Commodities do not seem to play any role in the short term. The inclusion of this asset class is reported only over the 10-year time horizon, when the pension fund has a solid initial financial position (FR=130). Figure 4 shows the compositions of the optimal portfolios for the three investment
horizons under investigation, when gamma is set equal to 10 and the initial funding ratio is 130. An interesting finding is that the composition of the optimal portfolios at 3- and 5-year horizons is similar to the composition seen previously, composed of the matching portfolio, property, and equities. In this case, given the stronger initial position, the full indexation is reached also at the 5-year time horizon, associated with very low standard deviation of delta. At the 10-year time horizon, the composition is different. Once again there is a shift of resources from the matching portfolio to property, but also to commodities (3.1%) and equities. This could be explained by a riskier investment strategy. These results confirm the crucial role played by property, the secondary role played by equities, and that Commodities have risk diversification properties that are exploitable only in the long term when solvency constrain are imposed.

From the analysis developed so far, it emerges that there is a convergence in the compositions of the optimal portfolios that are able to ensure full indexation at different time horizons or initial funding ratios. This composition invests around 88 to 90% in the matching portfolio, 8.8 to 10% in property, and residual resources in equities. This composition changes only when riskier investment strategies are needed to reach higher levels of indexation. However, this composition cannot be easily replicated in the financial markets. For this reason, we restrict the matching portfolio to be equal to or smaller than 63%. Feasible solutions are available only at higher initial funding ratios and over the short and medium term. These portfolios reach slightly lower or equal levels of utility than the optimal portfolios, but associated with higher standard deviations. The compositions of these restricted portfolios are shown in Figure 5. What cannot be invested in the matching portfolio is invested, almost exclusively, in property. Investments in equities remain almost unchanged, with the exception of a slightly higher level of investments in emerging market equity over the short term.

Once again, these restricted optimal portfolios present a shortcoming relative to their composition, due to the illiquid nature of the property. A pension fund hardly invests such a large amount of its resources in illiquid investments. Figure 6 shows what happens to the composition of the portfolio if we add a new constraint and restrict the weight of property to be equal to or smaller than 15%. Feasible solutions are only available in the short and medium term when the funding ratio is 130. We observe a relevant investment in passive equity and a larger investment in commodities. The contributions of these assets in the short and medium term exist and are valuable when property is not available.

**Conclusion**

This paper has developed an indexation-based optimization model which considers that indexation should be the primary target for a pension fund. Given this new pension deal offered by the pension fund, the indexation has to be considered as the objective function of the optimization of the portfolio. Since compensation of the losses in the purchasing power of the liabilities are no longer guaranteed by the pension fund, a specific model is needed to aim for its maximization. The model has been applied to the real case of ABN AMRO Pension fund and considers also the inclusion of real assets in the portfolio of the fund. The composition of optimal portfolios has been examined for different initial funding ratios, risk aversion levels, and investment horizons. The influence of different risk aversion levels in the definition of the compositions of the optimal portfolio is limited. The sustainability of the indexation-based portfolios is easily affordable over the short term, even if the full indexation can be reached only at higher levels of the initial funding ratio. The initial funding ratio strongly impacts the ability of the fund to set an investment strategy over the longer term. This can be easily explained considering the cumulative effect of the indexation policy.

Concerning the composition, there is a convergence in the results towards a portfolio composed of the matching portfolio (around 88-90%), property (8-9%), and equities (1-2%). Commodities and equities, which are typically viewed as inflationary hedging instruments, play a very limited role in these portfolios. Property represents a better investment opportunity than equities during all time horizons. These compositions change when riskier strategies are needed to reach higher levels of indexation. In this
case there is a significant shift of resources from the matching portfolio to property. Commodities are included in the portfolio only over the longest time horizons and when the fund has a solid initial financial position to overcome the risk-based regulatory constraints. These results partially contrast with the main findings in previous studies of the subject. That is to say, that equities are a preferable asset to property and that commodities are good risk-diversifiers during each time horizon. However, when we restrict the optimization, imposing constraints to investments in the matching portfolio (due to the imperfections of the long term bond market) and in property (due to its illiquid nature), commodities and in particular passive equities play a crucial role in the short and medium term.

References

• Froot, K. A., 1995, Hedging portfolios with real assets,” Journal of Portfolio Management, Summer, 60-77
Guidelines for Manuscript Submissions

Guidelines for authors
In order to aid our readership, we have established some guidelines to ensure that published papers meet the highest standards of thought leadership and practicality. The articles should, therefore, meet the following criteria:

1. Does this article make a significant contribution to this field of research?
2. Can the ideas presented in the article be applied to current business models? If not, is there a roadmap on how to get there?
3. Can your assertions be supported by empirical data?
4. Is my article purely abstract? If so, does it picture a world that can exist in the future?
5. Can your propositions be backed by a source of authority, preferably yours?
6. Would senior executives find this paper interesting?

Subjects of interest
All articles must be relevant and interesting to senior executives of the leading financial services organizations. They should assist in strategy formulations. The topics that are of interest to our readership include:

- Impact of e-finance on financial markets & institutions
- Marketing & branding
- Organizational behavior & structure
- Competitive landscape
- Operational & strategic issues
- Capital acquisition & allocation
- Structural readjustment
- Innovation & new sources of liquidity
- Leadership
- Financial regulations
- Financial technology

Manuscript guidelines
All manuscript submissions must be in English.

Manuscripts should not be longer than 7,000 words each. The maximum number of A4 pages allowed is 14, including all footnotes, references, charts and tables.

All manuscripts should be submitted by e-mail directly to the editor@capco.com in the PC version of Microsoft Word. They should all use Times New Roman font, and font size 10. Accompanying the Word document should be a PDF version that accurately reproduces all formulas, graphs and illustrations.

Please note that formatting in italics or underline will not be reproduced, and that bold is used only in subtitles, tables and graphs.

Where tables or graphs are used in the manuscript, the respective data should also be provided within a Microsoft Excel spreadsheet format.

The first page must provide the full name(s), title(s), organizational affiliation of the author(s), and contact details of the author(s). Contact details should include address, phone number, fax number, and e-mail address.

Footnotes should be double-spaced and be kept to a minimum. They should be numbered consecutively throughout the text with superscript Arabic numerals.

For monographs

For books

For contributions to collective works

For periodicals

For unpublished material
Request for Papers
Deadline June 16th, 2011

The world of finance has undergone tremendous change in recent years. Physical barriers have come down and organizations are finding it harder to maintain competitive advantage within today's truly global market place. This paradigm shift has forced managers to identify new ways to manage their operations and finances. The managers of tomorrow will, therefore, need completely different skill sets to succeed.

It is in response to this growing need that Capco is pleased to publish the ‘Journal of financial transformation.’ A journal dedicated to the advancement of leading thinking in the field of applied finance.

The Journal, which provides a unique linkage between scholarly research and business experience, aims to be the main source of thought leadership in this discipline for senior executives, management consultants, academics, researchers, and students. This objective can only be achieved through relentless pursuit of scholarly integrity and advancement. It is for this reason that we have invited some of the world's most renowned experts from academia and business to join our editorial board. It is their responsibility to ensure that we succeed in establishing a truly independent forum for leading thinking in this new discipline.

You can also contribute to the advancement of this field by submitting your thought leadership to the Journal.

We hope that you will join us on our journey of discovery and help shape the future of finance.

Prof. Shahin Shojai
Editor@capco.com

For more info, see opposite page
Minimise risk, optimise success

With a Masters degree from Cass Business School, you will gain the knowledge and skills to stand out in the real world.

MSc in Insurance and Risk Management

Cass is one of the world’s leading academic centres in the insurance field. What’s more, graduates from the MSc in Insurance and Risk Management gain exemption from approximately 70% of the examinations required to achieve the Advanced Diploma of the Chartered Insurance Institute (ACII).

For applicants to the Insurance and Risk Management MSc who already hold a CII Advanced Diploma, there is a fast-track January start, giving exemption from the first term of the degree.

To find out more about our regular information sessions, the next is 10 April 2008, visit www.cass.city.ac.uk/masters and click on 'sessions at Cass' or 'International & UK'.

Alternatively call admissions on:
+44 (0)20 7040 8611