RISK MANAGEMENT IN A FAIR VALUATION WORLD

By M. H. D. Kemp

[Presented to the Institute of Actuaries, 25 April 2005]

ABSTRACT

This paper considers the impact that the current trend towards fair valuation of assets and liabilities is likely to have on risk measurement and management practices within the financial services industry. The paper analyses the different sorts of risks faced by organisations such as asset managers, pension funds, banks and insurers and seeks to identify how their approach to the measurement and management of these sorts of risks might change as fair valuation becomes more entrenched. It argues that what it describes as traditional 'time series' based risk measurement is likely to be progressively displaced over time by a greater emphasis on what the paper refers to as 'derivative pricing' (or 'fair value' or 'market consistent') based risk modelling. It comments on the trend towards liability driven investment. The paper focuses on 'financial' risks (market, credit, liquidity and, more generally, asset/liability risk) rather than "operational" risks, whilst noting that the dividing line between the two can be open to interpretation. Insurance risk is seen as in some respects straddling both camps.

KEYWORDS

Fair Values; Fair Valuation Techniques; Risk Models; Risk Measurement; Risk Management; Value-at-Risk; Tracking Error; Time Series Based Risk Models; Derivative Pricing Based Risk Models; Performance Measurement; Market Risk; Credit Risk; Liquidity Risk; Liability Driven Investment; Collateralised Debt Obligations; Risk Budgeting; Risk Attribution; Quantitative Return Forecasting

CONTACT ADDRESS

M. H. D. Kemp, Threadneedle Asset Management Limited, 60 St Mary Axe, London EC3A 8JQ, U.K.

1. Introduction

1.1 *Rationale for Paper and its Main Conclusions*

1.1.1 There is a clear trend at present towards the *fair valuation* of assets and liabilities. This paper discusses the impact that this trend is likely to have on risk measurement and management practices within the financial services industry and some of the subtleties and challenges to which it may give rise. There are, in my opinion, powerful theoretical arguments that underpin the fair valuation concept (see Section 2), even if it has experienced some pushback from some sources. The paper assumes that the trend will continue.

1.1.2 The main aim of this paper is not to focus on, say, relatively detailed accounting implications of fair valuation for some specific type of financial services entity, as, for example, is covered for United Kingdom general insurers in Clark *et al.* (2003). Instead, it seeks to highlight the broader implications, as I see them, of fair valuation for the risk management approaches used by such entities.

1.1.3 The paper, like any other, betrays the author's own background. This is, at least more recently, primarily from within the asset management community. The paper therefore focuses more on assets than, liabilities, i.e. on investment risk, rather than non-investment risk, management tools, such as reinsurance, securitisation or outsourcing. Of course, it is not always helpful to treat assets and liabilities as two disjoint boxes, since one person's liabilities are often another person's assets. It should be noted that all views expressed in this paper are the author's alone, and do not necessarily accord with those of his employer.

1.1.4 Many themes are explored in the paper. Ones given particular prominence include:

Theme	Section/ Appendix
The underlying similarities between market, credit and liquidity risk, particularly if you take into account recent developments with CDOs and the like.	3, 9, 10
The distinction between the above sorts of risk (all of which involve an entity's interaction with external markets) and operational (including group) risk (which depends heavily on the entity's own internal structure). Insurance risk potentially straddles both camps.	3
The growth in 'liability driven investment' within the institutional investment scene (both pensions and insurance), and the relevance of derivatives (including credit derivatives) to this type of asset/liability risk management.	4, A
The philosophical and practical distinction between traditional 'time series' based risk modelling and 'derivative pricing' (aka 'fair value' or 'market consistent') based risk modelling. How this distinction has echos elsewhere, e.g. in the dichotomy between credit ratings and market implied default rates.	5 – 7, 9

The conceptual relevance of 'tranching' to solvency and other sorts of risk capital computations, and therefore to the likely way in which these computations will evolve over time (if these computations are to become more inherently market consistent in totality rather than just in part). How this favours the use of derivative pricing based risk modelling, with more traditional time series risk modelling perhaps ultimately being relegated to 'filling in the gaps' that market data do not readily provide.	9
The assertion that a 'fully market consistent' approach to setting capital requirements in effect involves answering, for some x, the question: "What capital does the company need (and in what form) to ensure that, if the company restructured itself into something akin to a CDO, the tranche relating to policyholder liabilities (or the equivalent for a non-insurer) would command a market spread (over the appropriate risk free rate) of less than x% p.a.?"	9.6
The need to take into account how the client's risk appetite differs from that of others investing in similar assets when interpreting the results of stochastic asset/liability projections and other similar exercises.	8
Whether 'long-term-ness' in insurance contracts is necessarily beneficial, either to the insurer or to the policyholder.	11
The challenges (and opportunities) that further change in the risk measurement and management arena will afford to actuaries.	12
The mathematical background to risk measurement (and quantitative return forecasting more generally), the geometrical analogy, and the challenges of high dimensionality.	5 – 7, C – E

1.2 *What do we Mean by 'Fair Value'?*

1.2.1 For the purposes of this paper we define the *fair value* of an asset or liability to mean its market value, if it is readily traded on a market at the point in time when the valuation is struck. The fair value of any other asset or liability is defined as a reasoned best estimate of what its market value would have been had it been traded at the relevant valuation point. Section 2 contains a consideration of some of the more tricky issues that arise in practice with such valuations, e.g. the issue of bid/offer spread, particularly if the position is large, and the need to hypothesise how a market would operate if there is not one currently (an issue that is particularly relevant to many types of liabilities).

1.2.2 This definition is essentially the same as the more standard definition of fair value as "the value at which an arms-length transaction involving willing, knowledgeable counterparties would take place". However, referring to market values does make clearer the link with 'market consistent' valuation principles. For example, it makes clear that determining the fair value of a non-traded asset or liability *is not the same as determining the valuer's own intrinsic assessment of its value*. Rather it involves *modelling how the market would value the asset or liability*, with the model, by implication, being calibrated in some suitable way back to market prices of instruments that are more readily traded.

1.2.3 A corollary is that the fair value of a liability does not depend on how the entity incurring the liability might hedge or otherwise manage the liability (except to the extent that honouring the liability may depend on the non-default of the entity bearing the liability).

1.2.4 An alternative term with normally the same meaning is the *mark-to-market* value of an asset or liability (although some people differentiate between this and *mark-to-model* in circumstances where a modelling element is required). When financial services regulators use the term *realistic* or *market consistent* valuations they also normally have a similar concept in mind.

1.3 *The Applicability of Fair Valuation Methodologies to Asset Managers and Other Similar Market Participants*

1.3.1 Certain parts of the financial services industry are already far down the fair valuation road. For example, asset managers nearly always report to clients using market valuations (or other sorts of 'fair' valuations for less liquid assets, such as property/real estate), but they are still affected by the trend towards fair valuation, either in their own right or because they need to be aware of its impact on their clients (and on the markets in which they might invest their clients' assets). Indeed, the term 'fair valuation' has acquired a particular meaning for asset managers in the light of the market timing and late trading scandals that have recently affected several United States fund management houses, see Investment Management Association (2004) and IMA & DATA (2004). It refers to the process of inferring a 'fair' price to place on units in a unitised fund at a point in time when some of the markets in which the fund invests are closed or in some other way the prices being used in the valuation are 'stale'.

1.3.2 This sort of fair valuation is actually quite a good example of fair valuation more generally, in terms of how it involves marking securities (or entire portfolios) 'to model'. In the U.K. context, the most obvious application is to U.S. equity retail funds with intra-day pricing points. For example, it might involve taking prices that were ruling at last night's U.S. close (c. 9 p.m. U.K. time) and imputing from them fair prices as at the fund's actual pricing point, say, 12 noon today (U.K. time), using movements in the interim in *market observables*, such as (in this case) the Globex S&P 500 future, or foreign listed variants of U.S. equities, exchange traded funds, etc. The aim is to stop arbitrageurs using the same sorts of calculations to exploit the otherwise stale nature of the fund's unit price to the detriment of other unit holders. It is fairly obvious that such a discrepancy might exist here, particularly since the S&P future is now traded almost around the clock. If you delve deeper, you discover that the prices of other sorts of securities can, in principle, also be exposed to such exploitation. For example, for some sectors of the bond market it is difficult to obtain prices other than at global close; more of them are marked 'to model' than you might perhaps expect.

1.3.3 Asset managers typically carry little investment risk on their own balance sheets (other than indirectly, because their own revenue stream, and hence business worth, are influenced by market movements). When they carry out trades, they are typically just *agents* acting on behalf of their clients, who are the *principals* involved in the trades. Of course, life is not always this simple. Some principals may transfer the investment risk, in whole or in part, to others (e.g. unit-linked life insurers legally own their unit-linked funds, but typically pass most or all of the market risks contained within these funds onto unit-linked policyholders). And sometimes investment managers may end up carrying more of the investment risk than they intended. Working out exactly which risk is borne by which entity within the overall value chain is not always trivial.

1.3.4 Agency/principal relationships do affect how players think about investment risk, see Section 3. For example, asset managers typically measure investment risk versus whatever *benchmark* they have been given (either implicitly or explicitly) by their client. In contrast, entities acting as principals may be more interested in their asset/liability risk, i.e. how different might be the movement or return (mark-to-market or otherwise) on their assets and liabilities, since it is this that flows through to the entity's own profit and loss account or solvency position.

1.3.5 In this context, an increased focus in recent years on *liability driven investment* within the defined benefit pension scheme community is noteworthy. This is explored further in Section 4 and in Appendix A. A similar nascent trend (perhaps more properly titled *capital management driven investment*) is starting to appear in the life insurance space, and arguably is already commonplace within the non-life and banking spheres.

1.4 *Risk Measurement and Management*

1.4.1 It is natural to try to encapsulate the measurement of investment risk using metrics that are relatively easy to understand, capable of being tracked through time, and able to be compared across different portfolios/ entities.

1.4.2 This paper argues that a single underlying framework conceptually exists for measuring essentially all types of *portfolio* (i.e. *financial*) risk. It is less clear that a similar all encompassing framework can be developed for *operational* risk. In Section 5, we consider the main sorts of metrics that can be used for this purpose. We focus on (forward looking) *tracking error*, *Value-at-Risk* and related metrics. We describe the similarities between different sorts of risk measures, and comment on when one might be more appropriate than another.

1.4.3 Nearly all such metrics ultimately rely on there being some hypothetical underlying joint probability distribution that simultaneously describes how individual assets and/or liabilities might move, both in isolation and in relation to each other. These sorts of risk models are described and analysed further in Section 6. We discuss the inherent mathematical limitations that any such risk model faces. These limitations apply, not just to risk forecasting, but also to return forecasting.

1.4.4 However, if we explore the interaction of fair valuation and derivative pricing with risk measurement, we discover that there are fewer inherent limitations than we might have first thought, see Section 7. We use these insights to develop risk models (and, indeed, a different way of thinking about risk) that can, in principle, overcome some of these limitations, although, in practice, it too runs into the problem of limited data sets.

1.4.5 The lack of sufficient information to be able to construct inherently reliable risk models has some potentially important implications for how one might try to manage (rather than merely measure) investment risk, see Section 8. In this section, we also explore some dichotomies between how different sorts of financial services entities think about *asset/liability management*.

1.5 *Market and Credit Risk*

1.5.1 In practice, credit risk is often differentiated from market risk. There are several good practical reasons for doing so; but it seems to me that, from a theoretical perspective, the distinction is less clear cut, particularly if you take into account relatively recent developments in the field of *collateralised debt obligations* (CDOs), *collateralised loan obligations* (CLOs) and the like, see Section 9.

1.5.2 CDO 'technology' can be used in seemingly endless ways to parcel out one set of risks (not always merely credit risk) in different ways to different market participants, potentially freeing up capital in an efficient manner for certain of these participants. CDOs are not the only financial innovation to have occurred over the last few decades. Of even greater importance has been the growth of the derivatives markets, and the associated financial theory underlying these instruments. Indeed, CDOs can be thought of as special cases of more general types of credit derivative, reinforcing the linkage between risk measurement/management and derivative pricing noted elsewhere in the paper.

1.6 *Liquidity, Insurance Risk, Operational and Group Risk*

There are four other sorts of financial risk that a financial services entity is typically deemed to be exposed to. Two of them, namely *liquidity risk* and *insurance risk*, are explored further in Sections 10 and 11. The other two we do not cover in any detail, for reasons explained in Section 3.

2. The Trend towards Fair Valuation of Assets and Liabilities

2.1 *Key External Regulatory Factors driving this Trend*

2.1.1 There are several external drivers favouring fair valuation, including:

(a) *Developments in international accounting standards.* Given the international nature of capital markets, standards setters are keen to move towards carrying assets and liabilities in balance sheets at fair value, because of the greater uniformity and standardisation that this should bring (particularly if the assets/liabilities are relatively easily traded financial instruments that are somewhat divorced from the rest of the organisation's business).

(b) *Developments in international regulatory thinking regarding how financial services entities ought to be regulated.* Globalisation has led to a desire for harmonisation amongst different regulators. An example is the Basel II agreement on banking supervision, with its three pillar approach, pillar one being suitable capital adequacy rules, pillar two being the interaction between the firm and the regulator, and pillar three being extra disciplines imposed by the marketplace. The basic approach seems to have won wide acceptance across the globe. Indeed, it has spawned a similar overarching *Solvency II* project within the European Union for insurance company regulation. Whilst governments might, in theory, have an incentive to encourage organisations to domicile within their own domains via lax regulation, the relevant *regulators* (called *supervisors* in some jurisdictions) have the opposite incentive. Who wants to be the regulator that lands the next BCCI on its plate? Fair valuation techniques have some obvious attractions for regulators, see below.

2.1.2 Of course, there are also drivers in the opposite direction. Some national insurance industries have lobbied hard against fair valuation, and few people seem prepared to get banks to mark to market their retail books (e.g. mortgages, savings accounts, although see Sections 2.5 and 2.6). Two other concerns seem to have been that:

(a) *Introduction of fair valuation of assets and liabilities creates greater volatility in profits.* Of course, arguably the volatility is there anyway (just not readily apparent), or, maybe, the worry is that fair values will just be overly complex to calculate (and understand).

(b) *Introduction of new capital adequacy rules with which developments in fair valuation are linked may penalise certain sections of industry*. For example, new risk weighting rules may make it more onerous for banks to lend money to, say, middling-sized corporates. Of course, if banks had sufficiently sophisticated risk management systems, then they might not focus on any specific regulator defined capital adequacy rules, but would, instead, work out the 'true' risks inherent in such lending. If lending policies to such corporates had previously been too lax (or too stringent), then new capital adequacy rules should not stop the banking world eventually honing in on the right balance between risk and return.

2.1.3 There are several sectors of the financial services industry where fair valuation of both assets and liabilities is already the norm rather than the exception. One example is asset management (which also happens to be a reasonably global business). The value placed on units in an open ended unitised fund such as a *unit trust* or *open ended investment company* (OEIC) is normally calculated by taking the market value of the fund's assets less liabilities and dividing by the number of units in existence. Trading desks within banks also typically mark-to-market their assets and liabilities on, say, a daily basis (although the same is not necessarily the case for the loans which their loan departments hold).

2.1.4 There are other parts of the financial services industry where fair valuation is less entrenched, e.g. pension schemes; but even here, papers like Cowling *et al.* (2004) suggest that, in the U.K., fair valuation methodologies will, in time, become the norm.

2.2 *Examples*

2.2.1 The U.K.'s Financial Services Authority (FSA) regulates a large part of the U.K.'s financial services industry, having taken over responsibility from several predecessor organisations (e.g. the Bank of England for banks, IMRO for asset managers, the DTI for insurers) when the U.K. adopted a *unitary* regulatory framework.

2.2.2 The FSA has recently been introducing a new regulatory framework for U.K. life and non-life (i.e. property/casualty) insurers. The approach owes much to the one that it has already adopted for the banking sector.

2.2.3 In broad terms, the FSA's overall framework for the whole financial services industry might be characterised as permitting more sophisticated players to use their own internally developed models (subject to vetting by the regulator), with less sophisticated players having to fall back on more broad brush calculation methodologies. Over time, we might expect the more broad brush calculations to involve higher capital requirements in the majority of cases, to provide an appropriate incentive to enhance the sophistication of internal risk systems.

2.2.4 For the U.K. insurance industry, the FSA's framework involves a greater focus than previously on 'realistic' reporting and capital adequacy computations. U.K. insurers' assets have, in effect, for many years been carried at market value. So, the key changes are:

(a) liabilities are (at least for large with-profits funds) 'realistically' valued, i.e. valued in a market consistent fashion, as if they were traded in an open market and/or hedged by purchasing broadly equivalent instruments from third parties;

(b) adequate capital is held to protect against adverse movements between the assets and the liabilities (subject to any overriding minima imposed by, say, E.C. Directives); and

(c) there is a greater focus on systems and processes to measure and manage risk.

2.2.5 The immediate contribution from fair valuation is obvious — the liabilities now have to be valued using fair valuation methodologies. Longer term, it is the interaction between fair valuation and capital adequacy that is most likely to alter the shape of the industry and its thought processes.

2.2.6 There are also substantial changes afoot in the U.K. defined benefit pension fund industry (a part of the financial services industry not currently regulated by the FSA). New pension fund accounting standards and developments within actuarial thinking have led to a greater focus on attempting to identify what tradable assets might hypothetically best 'match' or hedge, the scheme's liabilities, and then valuing the liabilities by reference to the market value of these 'matching' assets.

2.2.7 The creation of a centralised Pensions Protection Fund (PPF) may further hasten these changes, assuming that the PPF levies contributions on some risk adjusted basis (and not merely on, say, the size of the scheme's assets or liabilities), see Section 4. This would give pension funds an added incentive to manage such risks. It, of course, also requires some sort of objective measure of these risks and of the value of the assets and liabilities driving them, which most likely will involve fair valuation techniques.

2.2.8 Elsewhere in the E.U., there is a similar trend towards unitary regulators and, because of it, towards fair valuation. Indeed, in some continental European countries, the unitary regulation even includes pension funds. Pension funds have historically been seen in Continental Europe more as variants of insurers and less as specialist entities in their own right within the financial services arena. An example is Holland. The Dutch pensions regulator, PVK, wrote to pension funds in September 2002 requiring them to get to a 105% solvency level in one year. On 1 January 2006 a new regulatory framework for Dutch pension funds comes into effect, requiring the use of 'fair values' for liabilities, see Hurst (2004). The PVK also regulates banks and insurance companies. The Belgian regulator (also a unitary regulator) is also introducing fair valuation regulatory approaches for the insurance companies that it regulates.

2.2.9 The Danes moved to a fair valuation approach for insurance liabilities a year or two ago. Danish life insurers can discount at a flat interest rate, reduced by 5% to provide a margin of prudence, or they can discount using a yield curve published daily by the regulator without the margin. The supervisor publishes daily a yield curve for this purpose. After some discussion, it was agreed to base this yield curve on swap rates rather than the yields ruling on government debt, a topic we discuss further in Section 10. Similar sorts of discussions are likely to be had by each regulator introducing fair values; it is relatively easy to specify the broad framework, less easy to get everyone's agreement to the fine print.

2.2.10 In contrast, the French insurance regulator is apparently less enthusiastic about fair valuation methodologies, perhaps because of a worry that it might reduce the amount of solvency capital held by the entities it regulates. The French do not have a unitary regulator that encompasses both banking and insurance.

2.3 *The Underlying Theoretical Attraction of the Fair Valuation Concept*

2.3.1 I personally favour fair valuation not just because there are external regulatory pressures in its favour, but also because it seems to me to be inherently logical. From the perspective of the entity itself, the use of fair valuations has some underlying rationale (as long, perhaps, as it does not hinder your competitive position to use such methodologies). These include:

(a) *It is conceptually the most appropriate way to value assets and liabilities for solvency purposes*. If you conceptually had to close the business down and sell off all your assets and liabilities, then their value would be what you could get for them in the market place (albeit you might have some flexibility over timing, to avoid being a forced seller). Lack of a traded market for the assets and liabilities in question obviously makes the calculations more challenging, but ignoring the issue would not help you to negotiate suitable prices for the assets and liabilities if ever you *really* had to sell them.

(b) *Fair values are widely seen as more 'objective' than any other sorts of valuation*. Conceptually, they should require less in the way of subjective input than other methodologies.

(c) *If you carry assets and liabilities (or more precisely their difference) at any other valuation, then you are implicitly taking a view that you can generate added (or subtracted) value by the way that you manage them versus what the market believes it could achieve*. It would seem prudent, from a risk management perspective, to assume that you will not add value by exploiting some perceived skill you might think you have, but which you do not actually have. Conversely, it would seem overly conservative to assume that you will systematically subtract more value than others in a similar position to you. Of course, tax can complicate the picture (although even here, if the difference is significant, then there may be a risk that the tax authorities take a different view to your tax advisors on the matter in question!).

(d) *Several other organisations, e.g. ratings agencies, analysts and regulators would all like the same information.* Many organisations are interested in the likelihood of default of the companies that they are reviewing. They should favour approaches to the valuation of assets and liabilities that aid comparability (both within and across business types). Fair valuation has a strong appeal to them. Arguably, if they cannot access fair valuations directly, then they will attempt to create their own harmonised views across different entities. It ought, logically, to be economically efficient for entities themselves to provide such harmonised information without having others attempting to second guess what the numbers should be (and by doing so, the entities should also be helping themselves to understand better their own competitive positioning). Of course, in the presence of agency costs, there may be others, e.g. managers, who have less incentive to candour (perhaps explaining why shareholder groups may be more broadly in favour of fair valuation than some company managers).

2.3.2 More generic rationales also exist:

(a) *Financial markets in effect exist to promote the 'law of one price'*, i.e. the idea that, for any financial instrument, there should be, at any particular point in time, a single price at which the instrument should trade (defined by the interaction of market participants). If the relevant instrument is freely traded, then this price will be the fair value of the instrument. To be more precise, there will actually be a range of prices, but one of the aims of a properly functioning market is to keep this *dealing spread* as narrow as possible, thereby providing as much liquidity as possible in the given instrument.

(b) *There has been, over the last 20 or 30 years, a huge amount of innovation in financial markets, particularly in relation to derivatives.* Much of the financial theory underlying these instruments is based on how they can be hedged by transactions in physical instruments. Thus, their valuation is intimately linked to the price at which we might expect to be able to carry out such hedging transactions in the relevant underlying financial market. Moreover, derivatives can be used to create an extremely wide range of potential payoff profiles, including some where each individual derivative instrument within a portfolio may be quite complicated, but where, taken as a whole, they have very simple pay-offs. Ensuring that the value of any potential combination of derivatives is sensible (not least that the value of a series of derivatives with zero payoff has zero value) becomes a real challenge if you do not adopt a fair valuation framework, closely allied, in this context, to what is known as a *no arbitrage* framework (see Section 7).

(c) For some market participants, whom we might refer to as *market makers*, the *only sensible valuation metric to use is the market price* (adjusted in some suitable way to reflect current or potential dealing spreads). These are participants who, in effect, seek their return on capital employed by providing liquidity to the market, carrying an *inventory* of financial assets (or liabilities) that they add to or subtract from on an opportunistic basis. The capital that they are employing is, in effect, this inventory (plus IT and human capital), and the market price reflects the cost of replacing their existing inventory with a new one.

(d) You might expect that the other main sort of market participants, whom we might refer to as *position takers*, could focus more on the 'intrinsic' value of a particular position, if this can be differentiated from its current market/fair value, and to be able to take idiosyncratic views as the value of particular assets and liabilities. Indeed, active investment managers, acting as agents for these position takers, are specifically paid to take such views, *but even they cannot ignore market prices*. There is a risk that their views prove erroneous. How ought an organisation to control this risk? An obvious element is to monitor how assessments of these 'intrinsic' values compare with the value assessments that others ascribe to the instruments, as represented by their current market values.

2.3.3 Of course, few market participants are exclusively market makers or exclusively position takers. Most participants have some elements of both, even if the vast majority of their activities can be categorised into one or other camp, and participants that might normally be firmly in one box can temporarily flip into the opposite box, or might need to consider what might happen if such a switch were involuntarily imposed on them.

2.3.4 This is of particular relevance to capital adequacy. Banks, insurers, pension schemes and other financial services entities maintain appropriate capital bases to protect their deposit holders, policyholders and beneficiaries against the risk that their assets might prove insufficient to meet their liabilities. An obvious question is whether, if you tried to transfer all the assets and liabilities to some other entity, you would be able to find an entity prepared to accept them without further capital injection. Thus, the underlying premise of a capital adequacy calculation is, or ought to be, that you hypothetically 'market make' the entity itself. A natural starting metric for this purpose is some estimate of the combined fair value of its assets and liabilities.

2.4 *Limitations to the Objectivity of Fair Valuations*

2.4.1 However, there are some limitations to the fair valuation concept.

An important point to realise is that they still potentially involve *computational subjectivity*. This point is expounded in detail for life insurance liabilities by Sheldon & Smith (2004).

2.4.2 Untraded assets or liabilities need to be marked to some sort of modelled value that is calibrated using instruments that are sufficiently similar, in terms of their economic characteristics, to be useful calibrators, but how similar do they need to be to be 'sufficiently similar'? Also, if several instruments fit the bill, how much weight do you give to each? Some of this sort of subjectivity can be expressed via assumed wider bid/offer spreads.

2.5 *The Impact of Discretion*

2.5.1 Fair valuations also potentially involve *inherent subjectivity*, because they can depend on the exercise of discretion, either by the firm or by the customer. Sheldon & Smith (2004) also consider this point in some detail, by reference to the discretion that a with-profits insurer has on what bonuses it declares in the future on these sorts of contracts. This particular area is one that the FSA has focused on in its recent refinements to U.K. insurance regulation (see Sections 4 and 11).

2.5.2 Discretion may also be exercised the other way round. It is then often closely linked to the knotty question of profit recognition. Take, for example, a time deposit with a retail bank. The lower the interest rate the bank provides (relative to market norms), the more profitable the contract is likely to be to the bank; or rather, the more profitable it would be until the depositor exercises his discretion to deposit his money elsewhere. Usually, the 'fair value' of such contracts for regulatory purposes would exclude the value of future profits generated by what one might describe (depending on your point of view) as customer inertia or customer goodwill. One reason for doing so is that a 'shock' to the bank sufficiently large to imperil its solvency might reasonably be expected to invalidate persistency assumptions otherwise needed to justify capitalising this profit.

2.5.3 A special form of 'discretion' available to an entity is the discretion not to honour its debts because it has gone insolvent. We might call this the *solvency put option*. To this extent, the market value of an entity's liabilities will never exceed its assets, because its liabilities will always be pro-rated down in such circumstances. This is, of course, of little help to regulators (or any party interested in the entity's accounts, see above). They would normally want any of this default optionality stripped out of fair value calculations, particularly ones linked to solvency risk capital, as such capital is there precisely to provide protection against such defaults.

2.6 *Different Types of Fair Valuation*

2.6.1 One might, therefore, distinguish between several types of fair value including:

(a) the *fair mid-value* of an asset or liability, at a particular instant in time, would be the market price (or, in the absence of a liquid market, the valuer's best estimate of the market price) at which marginal trades in either would occur between willing buyers and willing sellers if markets were frictionless;

(b) a *prudent fair value* might include a best estimate of how asset values might fall and liability values might rise because of market frictions (e.g. bid/offer spreads in the underlying instruments, lack of liquidity, etc.);

(c) a *no goodwill fair value* would be a fair value (or prudent fair value) that excluded the value of future profits arising from contract persistency that was at the discretion of other parties (principally customers); and

(d) an *entity credit spread eliminated fair value* would be what the fair value (or prudent or no goodwill fair value) would be were the default risk inherent in the entity itself to be removed from the market value of its liabilities.

3. Categorising Risk

3.1 *Classifying Risk*

3.1.1 There are many different ways of categorising risk. The FSA rules for U.K. regulated entities refer to a six-way categorisation of risk:

(a) *Market risk*. This is the risk that investments will perform adversely. For example, if I hold equities, then one aspect of the market risk which I face is that these equities might fall in value (or, if I am being assessed relative to a benchmark that these equities might fall in value relative to the equity element of the benchmark).

(b) *Credit risk*. This is the risk that the entity will suffer loss because of defaults or significant declines in the creditworthiness of its *counterparties*, including issuers of instruments in which it has invested. For example, if I hold bonds, then they may default (or, if I am being assessed relative to a benchmark, I may suffer more defaults by value than the benchmark does). The FSA includes, within its thinking on this topic, the degree to which such exposures might be *concentrated*, i.e. not *well diversified*, and the extent to which a solvency regime can be *procyclical*, and hence exacerbate the business cycle if it requires banks to strengthen their reserves when they can least afford to do so.

(c) *Liquidity risk*. This is the risk that a firm will not have sufficient liquidity to meet its liabilities as they become due, or can secure them only at excessive cost. I might have plenty of assets, but they might be impossible to sell at the time I need to (or to borrow against at a sensible rate), in order to meet actual cash flows which I have committed to paying.

(d) *Insurance risk*. This might be defined as any risk relating to insurance activities. However, this is not always a helpful classification. For example, if an insurance company provides a 'guarantee' it will typically be structured as an insurance policy, whilst if a bank provides a 'guarantee', then it will typically be structured as a banking contract, even though the two can have essentially identical economic characteristics.

(e) *Operational risk*. This, according to the FSA, is the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. A wide range of risks fall into this category, e.g. legal risk is a sub-category of operational risk.

(f) *Group risk*. This is the additional risk caused by being in a group company structure. For example, one part of the group may suffer a big loss. Resources may then be diverted from other parts of the group, causing a knock-on effect which would not have arisen had the other companies not been part of the same group.

3.1.2 For the purposes of this paper, we deem most types of group risk to be special instances of operational risk, applicable only to entities with a group structure. We differentiate between operational/group risk and the first three sorts of risk (market, credit and liquidity risk), on the grounds that:

(a) There seem to be inherent differences between the characteristics and mathematical analysis that can be applied to these two broad groupings. If I consider two entities with identical *external* relationships and characteristics, i.e. identical assets and liabilities, then their exposures to market, credit and liquidity risk are, by definition, identical, but their operational risks, being dependent on how they operate *internally*, may be quite different. To put it another way, (internal) operational risk is innately linked to compliance cultures, control procedures, human behaviour, computer systems security and a host of other topics *specific to the company in question*.

(b) A corollary is that there will be elements of operational risk that are not amenable to mathematical analysis, because the risk element is unique to the firm in question. Of course, some parts of operational risk are amenable to mathematical analysis, particularly if you can build up a statistically significant sample from other similar organisations; but no two companies are ever identical.

(c) For certain sorts of firm, most specifically an asset manager (or any other similar organisation acting as an agent), market/credit/liquidity risk are passed on to the client rather than being retained by the entity in question. In effect, they thus involve 'misfortune' rather than 'error' (as long as the asset manager did not transgress relevant portfolio constraints). Only operational risks ought logically to incur 'errors', and therefore the risk of actually paying compensation to the client. Of course, a loss is a loss, whatever the cause. In today's litigious world, clients may seek to reclassify the root cause of the loss, and/or the firm might suffer so much adverse reputational or new business impact that it might agree to such a reclassification.

3.1.3 This leaves insurance risk. It seems to me to have characteristics sometimes closer to market/credit/liquidity risk (i.e. derivable in effect from an organisation's external positioning *vis-à-vis* the rest of the world), and sometimes closer to operational risk (i.e. derivable, in effect, from a firm's or an individual's internal characteristics, although, typically, not that of the entity writing the insurance risk).

3.2 *The Blurred Boundaries between these sorts of Risk*

3.2.1 Like any categorisation, the one above can become blurred. For example, one can easily understand how a distinction between market risk and credit risk originally arose within the banking world. A bank is typically seen as having a *banking book*, lending money to others, giving rise to credit risk, and a *trading book*, that invests in market instruments, giving rise to market risk; but where do credit derivatives fit? Do they encapsulate credit risk or market risk? Entities could view the credit risk in such instruments (and even the credit risk encapsulated in physical bonds) as a form of 'market risk'; the important thing is to take account of it in some suitable fashion somewhere within the overall categorisation. See also Section 9.

3.2.2 Another example of a potential blurring is *expense risk*. Actuarial guidance seems to assume that, within an insurance company, expense risk is necessarily a form of insurance risk; but this seems possibly inappropriate to me. I, personally, would view expense risk as a form of operational risk, on the grounds that banks and other non-insurance financial services entities are also presumably exposed to expense risk (although, perhaps, not over quite such long timescales), and it is the only logical bucket into which they would place such risks. One might also view some expense risks as a form of market risk if expenses are linked to inflation and there are assets, such as index-linked gilts, whose values move in line with inflation.

3.2.3 Perhaps the most obvious potential blurring is that of *asset/liability risk*. Within an insurance company, this, too, has perhaps traditionally been viewed as an example of insurance risk, the primary control of which has often fallen to the actuarial function, but again, there is just the same sort of need to focus on asset/liability risk in other sorts of financial services entities like banks, since they, too, have both assets and liabilities. Banks, nowadays, typically have an *asset/liability committee* (*ALCO*), or some other committee with a similar function, but a different name, that monitors and manages this sort of risk.

3.3 *Asset/Liability Measurement, Modelling and Management*

To measure asset/liability risk, it is, of course, necessary to quantify one's assets and liabilities. There is an immediate link here with fair valuation, for the sorts of reasons outlined in Section 2. Fair valuation provides a methodology for measuring assets and liabilities in a consistent fashion. Quite how this fits in with how different sorts of entities manage asset/liability risk is covered in more detail in Section 8.

4. Liability Driven Investment and Capital Management

4.1 *Liability Driven Investment Management for U.K. Defined Benefit Schemes*

4.1.1 Over the last few years, some significant changes seem to have occurred in how U.K. defined benefit (DB) pension scheme trustees (and their consultants) think about their liabilities when framing their overall investment strategies. These changes are typified by the buzz-phrase *liability driven investment*, and equivalent terms, such as *liability led investing* or *asset/liability investing*, see Appendix A. The common thread seems to be a greater focus on matching of assets and liabilities, coupled with what might be described as more refined *risk budgeting* and/or *capital budgeting*. Typically for U.K. defined benefit pension funds, this is being expressed as follows:

(a) a greater emphasis on the bond-like nature of future liability cash flows;
(b) a greater emphasis on the specific incidence of these cash flows (or on characteristics such as *duration* and *convexity* linked to them), and not just those of generic bond indices;
(c) a greater use of swaps to artificially lengthen the duration of the assets closer to the duration of the liabilities (given the limited supply of physical assets with sufficiently long duration);
(d) a more refined analysis of why equities and other non-bond assets might have been appropriate in the first place; and
(e) a greater enthusiasm for *risk budgeting*, in all its various guises, throughout actively managed parts of the portfolio (and at the strategic asset/liability level).

4.1.2 Pension scheme trustees will, of course, be forgiven for thinking that they have always taken into account their scheme's liabilities within their investment strategies. So what is the logic behind this new incarnation of liability driven investment?

4.1.3 For most of the period since the 1960s, most U.K. defined benefit pension schemes exhibited a strong bias towards equity type investments. There have always been closed or hyper-mature schemes that have focused more on bonds, but received wisdom for typical open U.K. final salary pension schemes has, until recently, been that their liabilities are very long term and inflation linked in nature, and, therefore, that they should invest heavily in assets, such as equities, that have been deemed to have similar economic characteristics.

4.1.4 Furthermore, received wisdom has also been that equities would, over the long term outperform other major asset categories (including other asset categories perceived to have long-term inflation linked characteristics, such as index-linked gilts or property). So, by investing a high proportion of their assets in equities, U.K. pension schemes could, in a sense, 'have their cake and eat it'.

4.1.5 Several factors have contributed to a reassessment of what might be the most appropriate investment strategy for a final salary pension scheme to adopt. Two long-term features are:

(a) *Final salary schemes are continuing to mature*, with rising average ages and rising proportions of pensioner and deferred pension liabilities as a proportion of total liabilities.

(b) *Guaranteed benefits, as a proportion of total benefits, have been rising for some decades, via government mandated improvements to pensions in payment and early leaver benefits (e.g. limited price indexation)*. The greater the proportion of non-guaranteed benefits, the more flexibility exists over how the assets backing these liabilities might be invested; or, perhaps I should say that the greater is the range of investment strategies that can be justified if the risk is being carried by the beneficiaries, and it is unclear what is the nature, if any, of the liabilities to which the assets relate. So, reducing the discretionary element of the benefits increases the importance that needs to be placed on the precise characteristics of the liabilities.

4.1.6 However, one might expect these factors to lead only to a gradual shift in investment strategy over time. The main drivers of the current more wholesale reviews of investment strategy seem to me to be more immediate:

(a) *The recent equity bear market has highlighted the potential risks of holding equities*. It also makes a wholesale shift in strategy away from equities perhaps unpalatable right now, as it could lock in the adverse effects of previous market falls.

(b) *Many schemes have recently closed to new entrants* ('recently' here being in relation to the usual long-term timeframe within which a pension fund operates) with new members joining a defined contribution (DC) pension scheme instead. A momentous event such as this (as far as a particular DB scheme is concerned), might reasonably be expected to lead to a rather more fundamental reappraisal of what should be done with the DB scheme's assets, potentially leading to a step change in how they are invested. If enough schemes make such step changes at the same time, then the impact on the industry as a whole becomes significant.

(c) People may not have believed that equities were a perfect match for the liabilities of a typical U.K. final salary scheme. But *more openly debated of late is whether equities are even a tolerably good match for such liabilities*, see e.g. Cowling, Gordon & Speed (2004).

4.1.7 Superimposed on these pension fund specific factors is the broader trend towards fair valuation (i.e. marking-to-market) of assets and liabilities that is the focus of this paper. This is driving people to think more explicitly about the degree of mismatch between the assets and liabilities as measured by a fair value balance sheet. For example, the accounting treatment for pensions mandated by FRS 17 and similar international accounting standards draws on fair valuation concepts. These standards adopt a more fixed income orientated perspective on how to value the liabilities than the traditional more equity orientated view that has been prevalent for most of the last few decades.

4.1.8 For U.K. pension funds, an even more important driver may ultimately be the creation of the Pension Protection Fund (PPF). This will involve a compulsory quasi-insurance arrangement that provides a safety net for scheme members of insolvent pension schemes. Since June 2003, such schemes have had a charge on the sponsoring employer, so you also need the sponsoring employer to have defaulted. The fair 'cost' or premium for the insurance coverage that such a guarantee fund will provide is linked to what might be the fair value of the scheme's assets less liabilities in the event of the coverage being triggered. It is also linked to the likelihood that the coverage is triggered.

4.1.9 International experience, particularly in the U.S.A., suggests that central guarantee funds that do not charge tolerably the right sort of premium for this risk can be a potentially large drain on the public purse. Pension fund sponsors either arbitrage the premium rate computation or you end up with additional regulatory burdens that attempt to limit how easy it is to take advantage of the pricing mis-specification. So, I think that the PPF should (and it is likely that it will) attempt to price the risks that it is underwriting, either reasonably accurately or incorporating suitable margins of prudence. This would presumably involve premium rates that are set, in part, by reference to how unfavourably the scheme's assets might diverge versus its liabilities, and hence by the magnitude of the mismatch risk that the scheme is running. Of course, even if the PPF did so, it might still, itself, run into trouble if it fails to hedge its own risks appropriately. The key point is that an explicit cost to being mismatched that involves real cash outlay is likely to focus the minds of sponsors and trustees on these risks and whether they are really worth running.

4.1.10 Again, the trends in question are not solely U.K. focused. Equity market declines have been a worldwide phenomenon. As noted earlier, there are also significant changes taking place in Europe, not least in Denmark and Holland, where a more insurance orientated regulatory approach applies to pension funds, and where shifts to fair valuation methodologies have recently been mandated.

4.2 *Expressing Shifting Conventional Wisdom within a Fair Valuation Framework*

4.2.1 It is not always easy to understand or to present, in a straightforward fashion, the dynamics underlying pension scheme funding and solvency; but one way that may help is to present this sort of information in the form of a 'fair valuation' balance sheet, that shows the sensitivity of different parts of the balance sheet to different sorts of economic factors. Incidentally, such a reporting format has similarities to how one might try to present the impact of derivatives on portfolios, see e.g. LIFFE (1992a) and LIFFE (1992b). This is no accident, given fundamental links that exist between fair valuation and derivative pricing.

4.2.2. For example, several decades ago the 'fair valuation' balance sheet of a hypothetical pension scheme with 'assets' of 110 and 'liabilities' of 100 might have looked something like that set out in Table 1.

4.2.3 Making reasonably plausible assumptions, it is possible to argue that the investment strategies that schemes were then adopting would still be reasonably appropriate had fair valuation principles then been adopted, given the difficulties involved in identifying any type of asset that is a particularly good match for salaries over anything but the very long term. To put it another way, if a large proportion of a DB scheme's liabilities are actually discretionary in nature (and therefore ultimately dependent on the investment experience of the non-matched element of the total portfolio), then the trustees are largely free, in principle, to do whatever they like (or they think that their members would like) with this portion of the assets (subject to usual prudent person principles). Such a DB pension scheme is actually quite DC-like in nature.

4.2.4 Contrast this with a DB pension scheme today, as typified by a hypothetical fair valuation balance sheet, as set out in Table 2. There is relatively little discretionary element left (after taking into account legislation that has converted previously discretionary benefits into guaranteed benefits). A more significant mismatch is revealed. More of the investment risk is now, in effect, being borne by the sponsor.

4.3 *Likely Future Trends in Pension Fund Investment Strategy*

4.3.1 Bond exposures of U.K. DB pension funds have increased significantly over the last few years, see Figure 1. Whatever your views on the state of DB pension schemes in the U.K. (or how much you think will be the take-up of the 'liability driven investment' concept), it seems likely that

Table 1. A 'fair valuation' representation of the balance sheet of a typical U.K. defined benefit pension scheme a long time ago (from the perspective of the beneficiaries)

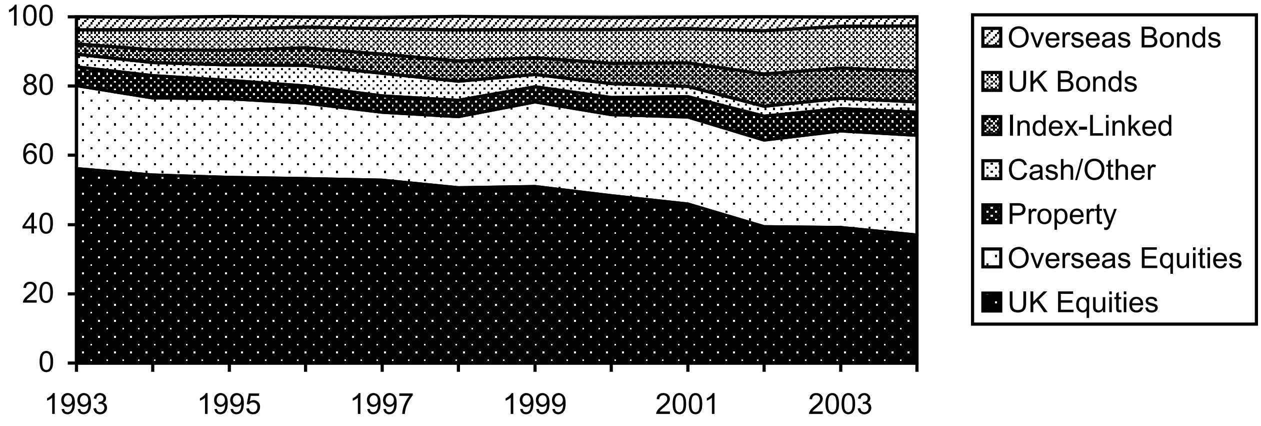
	Fixed nominal payments	Inflation linked payments	Salary linked payments	Other linkages	Total
Liabilities					
Guaranteed benefits	30[1]	–	–	–	30
Additional 'accrued' discretionary benefits	–	15[3]	15[2]	40[4]	70
Sponsor's share of future surpluses	3[8]	−15[8]	−15[8]	37[8]	10[5],[8]
Total liabilities	33	–	–	77	110
Assets	33[6]	–	–	77[7]	110

Explanation:
A long time ago guaranteed benefits were typically nominal in nature[1]. Guaranteed benefits were typically only a relatively modest part of the total accrued value of the pension benefits that beneficiaries could reasonably expect to receive. The difference, i.e. the discretionary enhancements the beneficiaries might reasonably expect to receive, would be partly be linked to future salary uplifts[2], partly to inflation linking of pensions in payment and deferred pensions prior to payment[3], but also strongly linked to how favourable future investment experience might be[4]. This last element is in a column titled 'Other linkages', here primarily relating to the performance of the unmatched element of the underlying asset base, typically the equity component, but also conceptually including mortality experience, etc. We have assumed that the pension scheme had a surplus, even after allowance for an appropriate level of discretionary benefits[5], that would ultimately return to the sponsor by way of contribution reductions or support via the pension fund for restructuring exercises. We assume that circa 30% of the asset portfolio was invested in (fixed-interest) bonds[6], the remainder in unconstrained assets[7]. The value to the scheme of the implicit guarantee of solvency from the sponsor (less an offset in relation to expected future contribution reductions) provides the balance[8].

Table 2. A more up-to-date 'fair valuation' representation of the balance sheet of a typical U.K. defined benefit pension scheme (from the perspective of the beneficiaries)

	Fixed nominal payments	Inflation linked payments	Salary linked payments	Other linkages	Total
Liabilities					
Guaranteed benefits	40[1]	50[1]	–	–	90
Additional 'accrued' discretionary benefits	–	–	5[1]	5[1]	10
Sponsor's share of future surpluses	−15[4]	−45[4]	−5[4]	55[4]	−10[2][4]
Total liabilities	25	5	–	60	90
Total assets	25[3]	5[3]	–	60[3]	90

Explanation:
The scheme is assumed now to have fewer active members, and therefore a higher proportion of guaranteed benefits, many of which are now inflation linked[1]. It is also assumed to be in deficit[2], and to have a somewhat higher bond exposure[3]. The claim on the sponsor still represents the balancing item[4].



Source: The WM Company

Figure 1. Average asset allocation of U.K. pension funds

the proportion of their investments in bonds will rise further, to unwind some of the mismatch highlighted above that has now opened up.

4.3.2 Quite how the non-bond assets might be structured is less clear. If you accept the argument that the rationale for their existence ultimately derives from the existence of discretionary elements to the liabilities, then there is no single 'right' way to invest them, as this, too, is then at the discretion of the trustees. Indeed, it may even be challenging to identify a single 'right' level of aggression to adopt within this element of the portfolio. Somewhat ill defined buzzwords, such as *unconstrained investment*, are now being used in this context. Perhaps one would fall back on economic logic, which, in a capitalist society, might be taken to imply a high weighting in equities, given the extra reward one might expect society to award to entrepreneurs and risk capital providers (although it is still then difficult to identify precisely how high the exposure should be).

4.3.3 Perhaps a helpful way of characterising unconstrained investment and liability driven investment is as two sides of a *core satellite approach*. The liability driven investment element is the core low risk element of the total portfolio, anchored by reference to the scheme's liabilities. It would focus on a strategy with relatively low risk versus the liabilities. The unconstrained element is the part focusing more on adding value. Investment consultants seem keen to promote the idea that unconstrained investment might involve paying relatively little attention to the exact construction of any specific market index and might involve a long time frame. Whether the fiduciary responsibilities imposed on pension scheme trustees will permit them to review such a manager's performance only infrequently, if the assets are reasonably liquid, is less clear to me.

4.3.4 We may, over time, also see more dynamic approaches to the allocation between the bonds (the liability driven core) and other asset types (the unconstrained satellite). This may be what is meant by another buzz-phrase which you sometimes hear in this context, namely the *'new' balanced management approach*. Option like characteristics arise in a number of contexts within a DB scheme's fair valuation balance sheet. For example, the benefit underpin provided by the sponsor might be thought of as like a put option given to the scheme by the sponsoring company. To limit the likelihood of the underpin being called upon, the sponsor could merely encourage adoption of a more matched position. However, it might, alternatively, encourage the scheme to invest either directly in an equivalent option that minimises the likelihood of the underpin being triggered, or in a dynamic hedging approach that provides some hedge against such a risk, see Appendix A. To date, sponsors who have focused on these option like exposures seem, more commonly, to have hedged such risks on their own balance sheets. This may reflect the practical complications of persuading a legally separate body, the pension scheme trustees, to adopt such a course of action.

4.4 *Liability Driven Investment for Insurers*

4.4.1 Much the same sort of change is beginning to materialise within (the non-profit and with-profits components of) life insurers, and, one might argue, has, to a considerable extent, already occurred within general insurers. The new regulatory framework introduced by the FSA has arguably made them more conscious of asset/liability mismatch risk. Over the last three to five years, U.K. with-profits funds have been major sellers of equities and buyers of bonds.

4.4.2 Insurers, typically, have shorter-dated liabilities than pension funds (except, perhaps, in their pension business books) so there is less need for them to enter into swaps purely to lengthen the duration of their assets. They have often been significant purchasers of swaptions (i.e. swaps with option elements) to hedge the *guaranteed annuity options* (GAOs).

4.4.3 It is possible that this will lead to what might be called a *capital budgeting* approach to investment management. Instead of, as at present, typically choosing some asset mix itself, and then handing out the assets to be managed in approximately these proportions, insurers might agree some way of measuring the capital being utilised by a particular investment strategy, and then give the asset manager a capital 'budget' to be used in whatever way the manager thinks fit, as long as it adds value; but this may prove to be a step too far for many insurers. Asset managers are paid to outperform a benchmark, and so they take the benchmark very seriously. A capital budgeting style approach, like the request for 'absolute returns' from a hedge fund, may merely implicitly equate the benchmark with a cash like return, which may not necessarily be what the insurer wants. Perhaps capital budgeting is more likely to be successful in the general insurance space, where insurers are, typically, more conservative in terms of the investment risk which they are prepared to take, and, anyway, more often view their liabilities as most closely matched by a cash like return. For life insurers, there are complications arising from what used to be called *Policyholders' Reasonable Expectations* (PRE), but now go under the more generic name of *Treating Customers Fairly* (TCF), see Section 4.5.

4.5 *Fair Valuation and its Interaction with Discretion and With-Profits Life Insurers' Principles and Practices of Financial Management (PPFM)*

4.5.1 The FSA has recently imposed changes to governance arrangements of U.K. with-profits funds. Insurers now have to issue statements (PPFMs) setting out how their with-profits funds are to be managed. They need to appoint a separate *with-profits actuary* to look after the interests of the with-profits policyholders. The with-profits actuary cannot have certain other roles within the insurer deemed likely to lead to potential or actual conflicts of interest.

4.5.2 With-profits liabilities can be thought of as involving an asset share element subject to some minimum sum assured (the 'put option' representation), or a guaranteed benefit plus some market upside (the 'call option' representation). The two give the same answer because of *put/call parity* (in practice, put and call options do not always exactly satisfy put/call parity, because of, say, discrepancies in tax treatment). There are, of course, complications in practice, e.g. regular rather than single premiums, mortality, lapses and the existence of 'market value adjustments' (MVAs) that might, typically, be applied to surrender values, but might not be applied on certain policy anniversaries.

4.5.3 Superimposed on these 'contractual' liabilities are those arising from TCF. One aim of a PPFM is to define more precisely how an insurer's discretion in computing the asset share algorithm is likely to be exercised, and what, in practice, treating customers fairly might mean. This has obvious attractions to the regulator in an era when a high value is now being placed on transparency.

4.5.4 Within such a framework, the shareholder provides a solvency underpin (or policy guarantee), implicitly receiving a reward for doing so (via a future profit stream). This is typically in addition to any pro-rata share of profits (e.g. via a 90:10 type subdivision). Fair valuation theory has an important implication. The fair value of this underpin, being market derived, is largely independent of what the shareholder actually does in order to hedge the risk it has taken on by providing this underpin. There is a small second order linkage via the impact which such actions might have on the credit exposure which the policyholders have to the shareholder, which we ignore for the purposes of the following analysis.

4.5.5 Consider a highly stylised example involving a five-year with-profits bond, start asset share of 100, no lapses/withdrawals, and with a guaranteed floor (i.e. underpin in five years' time) of 100. Suppose that the assets backing this contract are invested in a combination of risky assets ('equities') and risk free assets ('cash'), with a start mix of 50:50. The (fixed) return on cash over the five-year life of the contract is, say, 4% p.a., and the volatility of equity returns is 20% p.a. Most importantly, suppose that the asset share algorithm stated in the PPFM permits the insurer to reduce the proportion in equities by up to ten percentage points (but to no lower than 20%) at each year end, if the equity market has fallen over the preceding year, and to raise it by up to ten percentage points (but to no higher than 70%), if the equity market has risen over the preceding year.

4.5.6 What is the fair value of the underpin? It is relatively straightforward to model it using a *risk neutral* valuation framework as per usual derivatives pricing theory, see Section 7. As might be expected, the answer depends on how the discretion available within the asset share algorithm will be exercised, see Table 3.

4.5.7 The key point is that fair valuation theory highlights the complete lack of commonality of interest between shareholders and policyholders as far as the underpin is concerned. What, in this respect, is a liability to the shareholder is an equal and opposite asset to policyholders! In this example, the shareholder has an incentive to get the fund to use the maximum possible flexibility available to it to adjust asset shares in the light of observed returns (see the first three scenarios analysed in Table 2, which fall as the use of flexibility increases). Even better is if it can arrange for the with-profits fund to exercise the flexibility regarding equity proportion only in a downward direction (compare the first line with the last two lines). If possible, the policyholders should try to achieve exactly the opposite!

4.5.8 It is possible that new business marketing pressures could create greater commonality of purpose (at least for those with-profits funds that still remain open). More likely, in my opinion, is that PPFMs will, over time, *cease to describe* how funds might exercise discretion, and will, instead, describe how funds will *not* exercise discretion, defining in detail exactly how the asset shares will be invested.

4.5.9 If so, what then is the point of a with-profits contract? Will it not,

Table 3. Value of shareholder underpin to with-profits policyholder of different approaches to exercising discretion

Up movement actually applied if market rises (%)	Down movement actually applied if market falls (%)	Value of shareholder liability at outset, and hence value to policyholder of the shareholder underpin
0	0	2.6
5	5	2.2
10	10	1.9
0	5	1.8
0	10	1.1

in effect, mutate into a unit-linked look-alike, and are not unit-linked contracts typically more capital efficient than with-profits contracts? Perhaps the only sorts of insurers where traditional with-profits contracts really do have a long-term likelihood of thriving, are those where a commonality of interest is enforced by some other means, e.g. by having the insurer mutually owned.

4.6 *Fair Valuation Theory and its Implications for DB Pension Schemes*

4.6.1 The same general point about benefit discretion and divergence of interests is also relevant to defined benefit pension schemes (although, in the U.K., normally investment decisions are the responsibility of a separate party, i.e. the scheme trustees, and not by the effective provider of the solvency underpin, i.e. the sponsoring employer) and to current discussions within the actuarial profession about when actuaries can simultaneously advise both sponsor and trustees. Some of the issues involved are discussed in Chapman *et al.* (2001); they, too, note that the sum of the fair values of every party's interests in a pension scheme equals the fair value of the whole arrangement.

4.6.2 There is another important corollary of fair valuation theory for underfunded pension schemes. Table 2 indicates that beneficiaries in such a scheme typically have an exposure to the creditworthiness of the sponsoring employer. If this exposure were via a debt instrument issued by the sponsor to the scheme itself (or was via a loan from the pension scheme to the employer), then it would be subject to the usual *self-investment* concentration limits applicable to a pension scheme's asset portfolio. Why should a deficit be treated any differently, other than because it is perhaps less obviously a scheme 'asset'?

4.6.3 More to the point, why do prudent trustees of underfunded schemes not seek to mitigate their exposure to the credit worthiness of their sponsoring employer by using *credit derivatives* to purchase *credit protection* against their sponsor defaulting. Maybe trustees avoid this, on the grounds of actual or perceived cost; but, maybe, there is a lack of appreciation amongst some pension scheme actuaries about how rapidly the credit derivative market is developing, and therefore how practical such a strategy is, or might shortly become. Volumes in the credit derivatives market are exploding at present (in part due to CDO activity, see Section 9). A year or two back, most secondary market activity in the credit market occurred via physical bond transactions. A year or two from now, most brokers seem to be expecting the majority of secondary transactions to take place via credit derivatives, most notably *credit default swaps*. Many are integrating their cash bond and credit derivatives dealing activities to reflect this change.

4.6.4 Banks were the first substantial users of the credit derivatives markets. A bank that has lent more to a given entity than its credit officers would ideally like can now lay off the excess to other market participants via the credit derivatives market. A big advantage, as far as the bank is concerned, is that the credit exposure can be passed on anonymously, since the entity might otherwise view this as a sign of disloyalty, hindering the ongoing business relationship.

4.6.5 Why should trustees not do likewise, if they end up with more exposure to a single entity, in this case their sponsor, than they would like? You pay a premium for buying such protection, but only over time, so, at outset, the fair value funding level should be largely unaffected by taking out such protection. The greater the likelihood of default, the more the protection costs to buy, but the more likely it is then to be claimed upon. There are some practical details like the need to collateralise the derivatives positions, see Appendix A, but these are relatively unimportant in the context of the bigger picture impact that such a strategy might have.

4.6.6 A scheme purchasing such protection is, in effect, charging back to the sponsor (via increased future contributions) the credit spread which the sponsor has to pay to its other creditors, but is not paying to the scheme; or, equivalently, one can think of it as moving the status of the scheme up the credit priority ladder in the event of the sponsor defaulting. It may make it less easy for the sponsor to raise fresh loans or debt from third parties, as third party appetite for the sponsor's credit would be partly sated by its sale to them by the scheme via the derivatives market.

4.6.7 There is an interaction here, with the knotty question of to what extent it is in the trustees' interests to maintain the long-term viability of the sponsor, since making it more difficult for the sponsor to raise fresh debt may not always help the beneficiaries by as much as protecting them short term against the potential default of the company. There is an assumption here that the sponsor will ultimately try to make good any shortfalls in the scheme. If future company contributions remained absolutely unaltered, then the protection costs will ultimately result in a lower asset base to meet future liability outgo. It is not obvious to me how such a strategy melds with the AA yields mandated for the accounting treatment of liabilities in FRS 17, as it highlights some of the flexibility that the trustees have in terms of taking credit risk. Widespread adoption of such credit mitigation strategies might also influence adoption of liability driven investment. Trustees might become less worried about adverse movements in their invested assets versus their liabilities, but the sponsor might worry more (as more of this risk would economically fall to it).

4.7 *Fair Valuation Theory and Pension Scheme Buy-Outs*

4.7.1 Another topical pension fund issue that fair valuation theory sheds light on is the relevance, or otherwise, of the cost of buying-out pension liabilities from insurance companies in computing pension fund discontinuance liabilities. Some commentators seem to believe:

(a) if all schemes wanted to buy out their liabilities at the same time then it would be impractical for insurers currently active in this market to satisfy the potential demand;
(b) current buy-out quotations are typically ‘prohibitively expensive’; and
(c) so, the logic goes, buy-out quotes are inappropriate to use as the basis for calculating pension scheme discontinuance valuations.

4.7.2 It seems to me that this is missing the point. Insurance company buy-out quotes ought to form some guide to the ‘fair’ value of the discontinuance liabilities, just as market prices of bonds or equities form some guide to the ‘fair’ value of these assets. All involve prices at which market transactions occur. There may perhaps be reasons for excessive margins in buy-out prices; but making *no* use of this market available data is implicitly assuming that the scheme could guarantee to provide the same liabilities more cheaply via a closed scheme run-off. Why should an individual scheme be better at running off a closed book of liabilities than an insurer that can, presumably, gain economies of scale by running off multiple such books?

4.7.3 Is it not it also possible that these commentators may be misvaluing some of the risks involved in a run-off strategy? For example, are they understating the potential for further mortality improvements, or rather the cost needed to transfer this risk to someone else? Are they being too optimistic on administration expenses (or again the cost needed to transfer this risk to someone else)? If there is still a sponsor at the time, then, perhaps, it will be happy to shoulder these risks, but maybe not. And if there is no sponsor, will the remaining beneficiaries be keen to carry these risks themselves? How can you tell, unless you identify the sources and sizes of these risks and costs, by working out why it is that insurers seem to want to charge healthy premiums for taking on such risks?

4.7.4 It seems to me that the sort of exercise conceptually needed is somewhat like the one described in Yiasoumi *et al.* (2004), were one to be deciding what to do with a closed scheme or one in run-off. For all its imperfections, the buy-out market does provide some ‘mark to market’ data relevant to the fair valuation of the liabilities, and some clues regarding the risks which the scheme will find most difficult to pass on to others. The suggestion, in that paper, of trying to persuade beneficiaries to swap their existing benefits for others that are easier to hedge or buy out may be worth considering, see also Section 11.

4.7.5 Whether it would be reasonable to diminish the fair value of beneficiaries’ entitlements, in such a process, is less clear to me, so such a suggestion may not help to reduce closure liabilities. If you have granted someone some valuable benefits, then merely because the benefits are costly does not seem to me to be a compelling reason for not honouring them, and this did not apparently seem to be a compelling reason to the House of Lords in their judgment on the Equitable Life. Some might argue that the benefits were not voluntarily granted; instead, members got a windfall of higher guaranteed benefits thanks to pensions legislation, and sponsors are now trying to claw some of this back. However, few employees would take kindly to their employer seeking to void new employment rights (e.g. rights to ask for flexible working conditions), merely because the rights did not exist when they first joined the company.

4.8 *Regulation of Pension Funds*

The trend towards fair valuation might also lead to changes in how U.K. DB pension funds are regulated. The fair value of a scheme's liabilities is essentially independent of the form in which the benefits are delivered to scheme beneficiaries. U.K. DB pension schemes are now often closed to new entrants, so the proportion of their liabilities linked to uncertain future salary increases is less than it was before. Over time, a typical DB pension fund will look more and more like a closed insurance book. So, why should they not also be regulated in the same manner as a closed insurance book? This already happens in some E.U. countries; or, maybe, insurers should be regulated more like pension funds, if you believe that insurers are currently overregulated; or, maybe, you can develop an argument that having a sponsoring company with other activities that might support the scheme changes the picture, although this seems debatable to me.

5. Risk Measurement

5.1 *Portfolio Risk Measurement and Reporting from an Asset Manager's Perspective*

5.1.1 At the most fundamental level, asset management clients give their (active) investment managers the task of *adding value* without taking *undue risk*. 'Adding value' is the subject of *performance measurement* (and the analysis of where the added value has come from is the subject of *performance attribution*), see Appendix B. *Portfolio risk measurement* is about trying to quantify what we mean by not taking 'undue risk', see Kemp *et al.* (2000).

5.1.2 Unlike performance measurement (strictly speaking the measurement of *past* investment performance), risk measurement can involve two complementary, but different, time frames:

(a) measurement of *past* risk, which attempts to answer questions such as: "What level of risk did the manager adopt and was the reward worth the manager taking these risks?"; and

(b) estimation of likely *future* portfolio risk, which attempts to answer questions such as: "What level of risk might the portfolio experience, looking forwards, were it to remain as currently structured?"

5.1.3 Risk measurement is an inherently imprecise science. Given sufficiently accurate data, we can calculate historic portfolio returns arbitrarily accurately. The same is not true with many sorts of risk measures. Any reasonable definition of risk will take into account the likelihood or otherwise of various (adverse) outcomes. Even after the event we will only know with certainty what actually happened. We still will not know what might have happened. 'Risk' also has different meanings for different people. Even within the asset management context it can mean, for example, the risk of underperforming other similar funds, the risk of underperforming relevant market indices, or the risk of loss of capital or failure to maintain an adequate level of income.

5.1.4 All of these different sorts of risk can, in some fundamental sense, be thought of as variants of risk relative to a suitable *benchmark* (in the case of the risk of loss of capital, as with the 'absolute return' objective often applied to hedge funds, the benchmark would be a suitable cash return). Measurement of risk involves some assessment of how far away from the benchmark the portfolio is, or has been. By implication, choosing the right sort of risk to focus on, and therefore the right benchmark to use, is a key task for any client. It forms the manager's neutral position. To outperform, you need to deviate from the benchmark; but the further you deviate, the more you might underperform.

5.1.5 Similar principles apply to any other financial entity when it is measuring portfolio or 'financial' risk. The main differences in detail boil down to the benchmarks against which risk is measured and the precise metrics used to quantify 'how far away'.

5.2 *Ex-Post (i.e. Historic, Backward Looking, Retrospective) Risk Measures*

5.2.1 A simple backward looking risk measure would be to calculate the maximum underperformance in, say, any given month during the last five years (or the average size of any such underperformance, or the worst cumulative amount of underperformance during the period under analysis). Some sectors of the fund management industry do just this (particularly hedge funds, often referring to such concepts by the term 'drawdown').

5.2.2 These sorts of measures can be particularly sensitive to one or two extreme movements within the period being analysed. Two funds may have been adopting equally risky sorts of positions in the past. The first may have been particularly 'unlucky', in that its positions might have been particularly hard hit by the market circumstances that it encountered. The second may have been more 'fortunate', without necessarily running any less 'risk' in some fundamental sense of the word.

5.2.3 All practical historic risk measures suffer from these sorts of difficulties. They are only imprecise measures of the 'intrinsic' (but ultimately unobservable) risk that the portfolio has been running. Statisticians, faced with this problem, tend to prefer risk measures that are not overly sensitive to a small number of extreme movements, and have other intuitively appealing mathematical characteristics, whilst still being appropriate for the task in question.

5.2.4 For this reason, the most usual sort of historic risk measure adopted in the fund management industry is the ex-post or *historic tracking error*. Tracking errors are based on the statistical concept of *standard deviations*. If the returns relative to the benchmark are Normally distributed, then, in roughly two periods out of every three we would expect the return to be within plus or minus one standard deviation of the average. The historic (i.e. retrospective) tracking error is merely another way of describing this standard deviation, usually annualised, referring to the *actual spread of returns experienced in the past*.

5.2.5 Standard deviations give equal weight to positive and negative outcomes, whereas, in practice, the negative ones are the ones we most dislike. There are various ways of constructing *downside* risk statistics that focus more on adverse events, e.g. the *downside semi-standard* deviation, or various 'drawdown' statistics (which only focus on negative relative returns). However, blind use of, say, drawdown measures would imply that a fund that has consistently outperformed in each period has taken little or no risk. This is at odds with the concept that risk involves deviating from the benchmark. As we have noted earlier, it can be difficult to distinguish between funds that were 'fortunate' that their high risk stances did not come home to roost and funds that actually adopted a low risk stance.

5.2.6 Returns that are negatively skewed or exhibit excess kurtosis (i.e. are *fat tailed*) are typically disliked by recipients. So, both of these measures may be calculated. There are even ways of graphically analysing the entire shape of the return distribution (and all of its moments).

5.2.7 Historic risk and return can be analysed jointly through scatter plots of the sort shown in Figure 2, if there are other comparable portfolios that can be used for reference purposes. The ideal is to appear towards the top left hand corner of this chart, since this corresponds to having both performed well relative to the benchmark and having adopted a probably low risk stance in doing so. A statistic often quoted in this context is the *information ratio*. It is the ratio between the relative return and the historic tracking error, i.e. it is the slope of the line joining the origin to the point representing the fund in question. If the fund manager concerned could have doubled the sizes of all the positions (relative to the benchmark) then both the risk and the return of the portfolio (relative to the benchmark), would be doubled, leaving this ratio unchanged. If the benchmark is cash (or an absolute return) then this statistic more normally goes under the name of *Sharpe ratio*. If, instead, we are focusing on downside risk, then the equivalent statistic is the *Sortino ratio*. A glossary of such terms is given in Kemp *et al.* (2000).

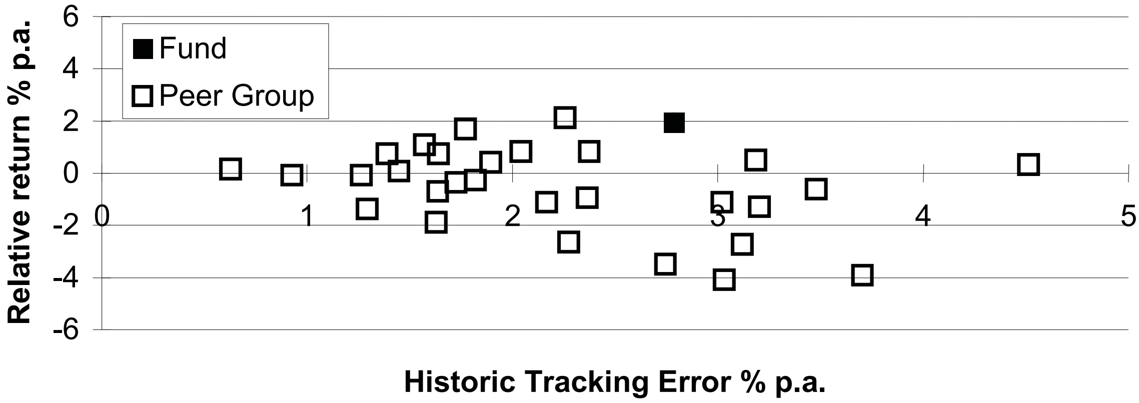


Figure 2. Illustration of a peer group scatter plot

5.3 *Forward Looking (i.e. Prospective) Risk Measures*

5.3.1 Forward looking risk measures are estimates of how much the portfolio return might deviate from benchmark return. An obvious metric, if you have measured historic risk using historic tracking errors, is to use *forward looking tracking errors*. A forward looking (i.e. prospective) tracking error is an estimate of the standard deviation of returns (relative to the benchmark) that the portfolio might experience in the future (were its current structure to remain unaltered).

5.3.2 The further that you are away from your benchmark, the more you might deviate from it, and the greater should be the risk that you are running. So, we should expect there to be an analogy between measuring 'distance' in, say, the real world (i.e. Eucleidean geometry) and 'risk' in the financial world. This analogy is particularly strong with forward looking tracking errors.

5.3.3 This *geometrical analogy* works as follows. Suppose that our positions relative to the benchmark can be described via a position vector **x** (written in bold lower case) whose terms are x_i (written in italicised indexed lower case), where x_i is the relative position in the ith security. Suppose that we describe the random variable that is the future (relative) return arising from holding **x** (over a suitably defined period) by the equivalent italicised capital letter, i.e. here X. Suppose, now, that we have two different sets of positions **a** and **b**, that create corresponding future returns A and B. The combination of the two positions $\mathbf{c} = \mathbf{a} + \mathbf{b}$, then creates a corresponding future return $C = A + B$. We note that C has a standard deviation (i.e. forward looking tracking error) of σ_C, which can be calculated as follows, where ρ_{AB} is the correlation coefficient between the random variables A and B:

$$\sigma_C^2 = \sigma_{A+B}^2 = \mathrm{var}(A+B) = \mathrm{var}(A) + 2\mathrm{cov}(A,B) + \mathrm{var}(A) = \sigma_A^2 + 2\sigma_A\sigma_B\rho_{AB} + \sigma_B^2.$$

5.3.4 If $\rho_{AB} = 0$, then the formula is very similar to Pythagoras' celebrated theorem, which tells us that the length P of the hypotenuse of a right-angled triangle can be found from the lengths Q and R of the two sides next to the right angle, using the formula $P^2 = Q^2 + R^2$. Indeed, one can derive it mathematically using similar principles. If $\rho_{AB} \neq 0$, then it is very similar to the more general formula $P^2 = Q^2 + 2QR\cos\theta + R^2$ applicable to a non-right-angled triangle, where θ is the angle between sides Q and R, if we equate $\cos(\theta)$ with ρ_{AB}, see Figure 3.

5.3.5 The one subtle difference is that, in Euclidean geometry, the 'magnitude', i.e. length, of a distance vector (between zero and a point **x** with Cartesian coordinates x_i) is calculated as $|\mathbf{x}| \equiv \sqrt{\sum x_i^2} \equiv \sqrt{\mathbf{x}^T\mathbf{I}\mathbf{x}}$ (where **I** is the identity matrix), whilst in tracking error analysis the 'magnitude', i.e. tracking error, of a set of relative positions **a** is calculated as $\|a\| \equiv \sqrt{\sum a_i V_{ij} a_j} \equiv \sqrt{\mathbf{a}^T\mathbf{V}\mathbf{a}}$ (where **V** is the covariance matrix of the joint probability distribution from which X is drawn, the elements of which are V_{ij}).

5.3.6 The key point is that the same underlying geometrical concepts that apply to 'distance' in the real world can typically be applied to 'risk' in the financial world (as long as 'risk' is equated with forward looking tracking errors), by scaling the different axes in suitable ways and by including shears to the coordinate framework to reflect non-zero correlations between different securities.

5.3.7 I find this geometrical analogy to be a very powerful way of explaining conceptually, how tracking errors work. It helps me to understand intuitively several practical features of tracking errors and also some of the more complex characteristics of the models that people have developed to estimate tracking errors. For example, we can use the analogy to conclude that:

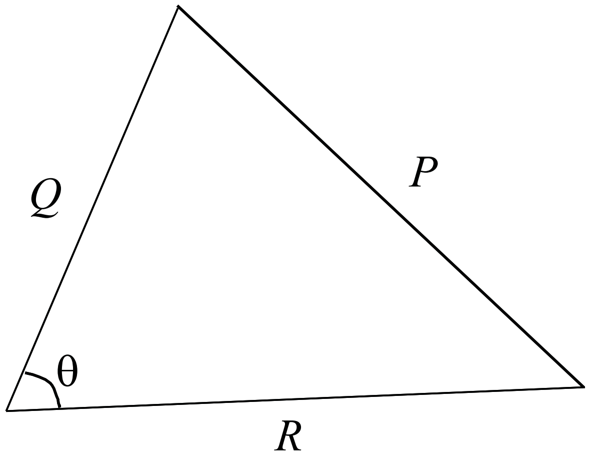


Figure 3. The geometrical meaning of tracking error:
$P^2 = Q^2 + 2QR\cos\theta + R^2$

(a) Forward looking tracking errors (and risk measures more generally) are dependent on *assumptions* both *about the likely future volatility* of individual stocks or markets relative to the benchmark (i.e. the degree of scaling required) and *about the correlations* between different stocks/markets (i.e. the extent of shearing needed).

(b) If we wish to rank portfolios by their riskiness, then, as long as the sort of risk is similar, the rankings will often be relatively insensitive to the precise risk methodology used.

(c) Ranking portfolios in this manner is less reliable if the sort of risk being measured is quite different.

5.4 *Value-at-Risk*

5.4.1 A more common forward looking risk metric in many parts of the financial community is *Value-at-Risk* (*VaR*). VaR is enticingly simple as a concept, and therefore relatively easy to explain to lay people. It requires specification of a confidence level, say 95% or 99%, and a time period, say one day, five days or one month. If we talk about a fund having a five-day 95% confidence VaR of X (X could be a monetary amount or a percentage of the fund), we mean that there is a only a 5% chance of losing more than X over the next five days, if the same positions are held for this five-day time frame. VaR originally referred merely to losses on some absolute numerical basis, but it is more helpful to use a generalised definition, in which VaR can also refer to percentage losses versus a suitably chosen benchmark.

5.4.2 There are several reasons why VaR is potentially more attractive as a risk measure to lay people than (forward looking) tracking errors, including:

(a) Tracking error requires an understanding of statistical concepts like standard deviations.

(b) Tracking error can also be applied both in a fully backward looking and in a fully forward looking manner, which again may give rise to misunderstandings. VaR can, in principle, do too, but, in practice, is much more commonly forward looking. For example, even the term 'historic VaR' is most usually taken to mean taking current positions and working out what would have happened based on, say, the last five or ten year's worth of past daily market movements, and so is still an estimate of what might happen going forwards if the current positions are retained (to the extent that past market movements are a guide to how markets might move in the future). Historic tracking error, in contrast, refers to historic positions. A fund that is perfectly indexed should, therefore, have a nil VaR (versus the index), whether 'historic' or otherwise (using the above terminology), but could still have an appreciable historic tracking error if it had only recently been converted into an index fund.

(c) If returns are Normally distributed, then an annualised forward looking tracking error is the same as a 15.9% one-year VaR; but which client is interested in a 15.9% break point?

(d) Flexibility in choice of confidence interval means that it may be easier to take practical account of the non-Normal distributions typically perceived to apply to returns in the real world.

5.4.3 Of course, life is not this simple. Estimating VaR, like estimating forward looking tracking error, involves subjective inputs that may not be immediately apparent to the recipient of the statistic. It is helpful to realise that VaR is frequently calculated using the same Normal distribution assumption as might be used to derive a forward looking tracking error, by reading off the appropriate confidence interval from a tabulation of the Normal distribution (or by using corresponding spreadsheet functions).

5.4.4 In such circumstances, VaR is, in effect, just an alternative way of presenting tracking error, albeit in a format which may be more intuitively appealing and perhaps focusing on a different (and often shorter) time horizon. *If* returns are Normally distributed, then the return distribution is completely characterised by two parameters, its mean and standard deviation, which means that, mathematically, the 'best' way of estimating any VaR statistic will typically involve reference to the sample standard deviation (the square of which is the 'minimum variance' unbiased estimator of the population variance).

5.4.5 The difference becomes more intrinsic if, as most commentators consider to be the case, future return distributions are not Normal; but I would then pose the question: "How practical is it to differentiate between different distributional forms?"

5.4.6 Sometimes the situation directly prescribes a non-Normal distribution. For example, the performance of a poorly diversified corporate bond fund is naturally likely to be significantly skewed, because of the risk that one or more of the bonds held might default, causing a significant relative loss versus the benchmark. Similar skews can naturally be expected to arise with portfolios that contain significant exposure to options (or other financial instruments with option like characteristics).

5.4.7 However, in most other circumstances it is more difficult to identify exactly how skewed or fat-tailed might be the underlying distribution of future returns. The distributional form underlying the VaR statistic might be taken directly from the historical data distribution (or from Monte Carlo or bootstrap simulations that, themselves, ultimately derive from this distribution), but this may be placing too much reliance on the particular sample of the past that was observed. Sampling errors may be particularly acute if we are focusing just on the extreme tail of the distribution. It may be better to adopt more robust methodologies that are less sensitive to the actual sample used, see e.g. Appendix C.

5.4.8 Alternatively, we might dispense with any distributional assumptions and merely use scenario tests (or 'stress tests') to identify the impact that some particular set of events, occurring simultaneously, might have. However, someone still needs to decide what stress tests to carry out (and how severe they should be). Without distributional assumptions, it is very difficult to identify objective criteria that can be used to aid this selection.

5.5 *Other Forward Looking Risk Measures*

5.5.1 As Leippold (2004) points out, the VaR of a portfolio is actually the *minimum loss* that a portfolio can suffer in *x* days in the *y*% *worst* cases when the portfolio weights are not changed during these *x* days. His reasons for clarifying the definition of VaR in this manner are to highlight that VaR fails to take any account of the shape of the distribution beyond the VaR cut-off point and to highlight that it is not a *coherent* risk measure in terms of how risks might add together.

5.5.2 For example, suppose that I have two return distributions, both with the same 95% VaR of £1m, but in one the average loss, in the event that the VaR cut-off point is breached, is £1.5m, and in the other it is £15m. The latter would, in most people's eyes, be riskier than the former, even though both have the same 95% VaR statistic. Or consider an insurer only underwriting a single catastrophe insurance risk, without any reinsurance. If the probability of it occurring is one in 300 years, then its yearly 99.5% VaR will be zero (or better, if it takes credit for the premium it receives); but, every once in a while, it will suffer a massive loss way beyond its VaR. If the same insurer underwrites 1/30th of 30 such risks, each independently with the same one in 300 year chance of occurring, then its yearly VaR will be much higher, even though it has a more diversified book of business. Or take a bond or CDO. Such an instrument might have an expected default rate, if held to maturity, of, say, 0.5%, which would imply that its 95% VaR over this period is 0, but, of course, it is not thereby riskless.

5.5.3 Possible risk metrics that are more useful in these circumstances are *expected loss* (or *expected shortfall*, sometimes called *TVAR*, that is, tail value at risk) which is the *average* loss that a portfolio can suffer in *x* days in the *y*% worst cases (rather than the *minimum* loss, as is used in the basic VaR computation), or other similar metrics that take better account of the shape of the tail.

5.6 *Risk Attribution*

Whatever risk metric is used, there will be a natural desire, just as there is with the performance measurement of returns, to understand what are the sources of the risk. This process is known as *risk attribution*, see Appendix D. As with performance attribution, there is no unique way to decompose risk into its various parts in such an analysis.

6. Time Series Based Risk Models

6.1 *Risk Models*

6.1.1 To calculate a forward looking risk statistic, you need a *risk model*, i.e. a mathematical framework for estimating the future spread of returns which a portfolio might generate, were its positions versus the benchmark to remain unaltered in the future.

6.1.2 A risk model can be differentiated from a *risk system*, which is a practical software tool that can be used to derive these sorts of statistics (or to carry out other related tasks, e.g. risk/return optimisation,, see Section 8). It is the underlying risk model that defines what answers you will get out of a risk system, even if ease of use, cost and run times are also key elements in deciding which system to buy. There are now quite a few specialist third party providers of risk systems. Quantitative research departments within investment banks also provide such services to their broking clients. Some asset managers have their own internally developed systems (which they sometimes then try to commercialise for third party use). The larger investment consultants provide similar services to their pension fund clients (usually in conjunction with a commercial risk system provider). Risk modelling capabilities are increasingly being added to asset/liability software supplied by insurance consultants and actuaries.

6.2 *Characterising Risk Models*

6.2.1 We can characterise the main sorts of commercially available risk models in several different ways:

(a) *How factors driving the behaviour of multiple securities are developed.* The main sub-classifications here are between fundamental, econometric and statistical models.

(b) *The shape of the underlying joint probability distribution that the risk model assumes will govern the behaviour of different securities*. There are some differences here between equities and bonds (and between securities/portfolios that do, or do not, contain optionality).

(c) *The mathematical algorithm used to calculate the risk metric*. The main distinction here is use of *analytical* versus *simulation* techniques. The latter can typified by Monte Carlo simulations, although other, more sophisticated, approaches, such as antithetic random variables, variance reduction techniques and low discrepancy quasi-random variables, may be used in an attempt to reduce the number of simulations needed to achieve a suitably accurate answer.

6.2.2 However, there are fewer underlying distinctions than appear at first sight. For example, categorisation by mathematical algorithm does not really define different underlying risk 'models' *per se*. Given unlimited computing power, simulation techniques will give the same answer as any corresponding exact analytical result; it is just that, in most cases, you cannot find an exact analytical answer without making some approximations that you may not feel are appropriate for the task in hand. For a mathematical/computational discussion of simulation techniques more generally, see e.g. Press *et al.* (1992). Such methods can normally be thought of as special cases of numerical integration techniques.

6.3 *Fundamental, Econometric and Statistical Risk Models*

6.3.1 There are three main types of time series based risk models:

(a) A *fundamental* risk model ascribes certain fundamental factors (such as price to book) to individual securities. These factor exposures are exogenously derived, e.g. by reference to a company's report and accounts. The factor exposures for a portfolio as a whole (or for a benchmark, and hence for a portfolio's active positions versus a benchmark) are the weighted averages of the individual position exposures. Different factors are assumed to behave in the future in a manner described by some joint probability distribution. The overall portfolio risk (versus its benchmark) can then be derived from its active factor exposures, this joint probability distribution and any additional variability in future returns deemed to arise from security specific idiosyncratic exposures held within the portfolio.

(b) An *econometric* risk model is similar to a fundamental model, except that the factor exposures are individual security specific sensitivities to certain pre-chosen exogenous economic variables, e.g. interest rate, currency or oil price movements. The sensitivities are typically found by regressing past returns on the security against past movements in the relevant economic variables (typically using multiple regression analyses, such as described in Appendix E).

(c) A *statistical* risk model eliminates the need to define any exogenous factors, whether fundamental or econometric. Instead, we identify some otherwise arbitrary time series that, in aggregate, explain well the past return histories of a high proportion of the relevant security universe, ascribing to these time series the status of 'factors'. Simultaneously, we also derive the exposures that each security has to these factors. This involves *principal components analysis* (or techniques that are mathematically equivalent, but might go under other names), see also Appendix E.

6.3.2 These different types of model are less different than might appear at first sight. It would be nice to believe that factors included within a fundamental or econometric model are chosen purely from inherent a priori criteria. In reality, however, the factors will normally be chosen, in part, because they seem to have exhibited some explanatory power in the past. They are, therefore, almost certain to have some broad correspondence to what you would have chosen had you merely analysed past returns in some detail as per method (c). How can we ever expect to decouple entirely what we consider to be a 'reasonable' way of describing market dynamics from past experience as to how markets have actually operated?

6.3.3 This blurring is particularly noticeable with bond risk models. A key driver of bond behaviour is *duration*. Is this a 'fundamental' factor, since we can calculate it exogenously by reference merely to the timing of the cash flows underlying the bond; or is it an 'econometric' factor, because a bond's modified duration is also its sensitivity to small parallel shifts in the yield curve; or is it a 'statistical' factor, because, if we carry out a principal components analysis of well-rated bonds, we typically find that the most important driver for a bond is its duration?

6.3.4 All three types of risk model have the same underlying mathematical framework, which we can derive from the geometrical representation of risk developed in Section 5. We model the ith security's return as coming from 'exposures' f_{ij} to the jth 'factor loading', one unit of each factor generating a prospective return (in the relevant future period) of r_j, where the r_j are random variables with, say, a joint covariance matrix $\mathbf{V}$. So, a portfolio described by an active weights vector $\mathbf{w}$ has an overall risk (equating for this purpose tracking error with risk) of σ, where $\sigma^2 = \mathbf{w}^T\left(\mathbf{F}^T\mathbf{V}\mathbf{F}\right)\mathbf{w} = (\mathbf{Fw})^T\mathbf{V}(\mathbf{Fw})$, where the matrix $\mathbf{F}$ contains the terms f_{ij}.

6.3.5 It is worth noting here that there are two different ways, in practice, of handling 'residual' risk within such a framework, i.e. the idiosyncratic risk that is relevant only to specific individual securities:

(a) the matrix $\mathbf{F}$ might be deemed to *include* all such idiosyncratic risks, i.e. the set of 'factors' which we consider includes idiosyncratic factors that predominantly affect only individual securities, and in this paper we concentrate on this approach, unless otherwise stated; or

(b) the matrix $\mathbf{F}$ *excludes* these idiosyncratic risks. In such a formalisation, the idiosyncratic risk of the ith security might be, say, s_i, and we might have the total risk of the portfolio now defined as $\sigma^2 = \mathbf{w}^T\mathbf{V}\mathbf{w} + \sum w_i^2\sigma_i^2$, where $\mathbf{w}$ is now a vector of active *factor* weights (not *security* weights), and $\mathbf{V}$ is a much smaller sized matrix that only refers to the factor covariance matrix. Some refinements occur, in practice, where two different securities are exposed to the same idiosyncratic risk. For example, some companies have a dual holding company structure, with one holding company being domiciled in one country and the other in a different country. The equities of the two holding companies may not trade at identical prices, but clearly do exhibit a strong linage. Bond issuers often have multiple bonds in issue.

6.3.6 Carrying out a principal components' analysis, in effect, involves identifying an orthogonal matrix $\mathbf{L}$ for which the matrix $\mathbf{M} = \mathbf{L}^T\mathbf{V}\mathbf{L}$ contains non-zero elements only along its leading diagonal (with the elements of the leading diagonal also typically sorted into, say, descending order). The sizes of these terms and the structure of **L** are intimately related to the eigenvalues and eigenvectors of **V**, see Appendix E. Usually, when people talk about 'principal' components analysis, they mean truncating this matrix, so that all bar a few of the leading diagonal terms are set to zero (equivalent to applying a further transform **P**, which is unity for the first few leading diagonal terms and zero everywhere else) and then backing out an adjusted covariance matrix $\bar{\mathbf{V}} = (\mathbf{L}^{-1})^T\left(\mathbf{P}^T\mathbf{L}^T\mathbf{V}\mathbf{L}\mathbf{P}\right)\mathbf{L}^{-1}$ from an original covariance matrix **V** derived from historic data. As L is orthogonal, $\mathbf{L}^{-1} = (\mathbf{L}^{-1})^T$ and $\mathbf{L}^T = \mathbf{L}$ (and likewise **P**), so this expression simplifies to $\bar{\mathbf{V}} = \mathbf{L}^{-1}\mathbf{PLVLPL}^{-1}$.

6.3.7 For example, if we carry out a principal components analysis on the entire (conventional) gilt market, then, typically, we would find that nearly all of the behaviour of nearly of all the gilts is well explained by a very small number of factors. By 'nearly all of the behaviour' we mean that only the first few of leading diagonal terms in **M** (i.e. only the largest few eigenvalues) are much different to zero. Bond risk models often focus on just the first three principal components, equating them with, say, *shift*, i.e. parallel shifts in the yield curve; *twist*, i.e. uniform steepening or flattening of the curve; and *butterfly*, i.e. uniform curving up or down of the curve, with the first being typically significantly more important that the other two.

6.3.8 Readers familiar with this subject will recognise that we are repeatedly applying *transformations* characterised by a matrix **A** that translates $\mathbf{X} \rightarrow \mathbf{A}^T\mathbf{X}\mathbf{A}$. Geometrically, these transformations can be equated with the same sort of rotation and/or stretching/shearing introduced in ¶5.3.6, or with the special case of such a matrix in which we shear away an entire dimension, i.e. we project the geometrical representation of the matrix onto some lower dimensional space.

6.3.9 Multivariate regression can be expressed using similar matrix algebra, see Appendix E. The process of creating econometric risk models is thus mathematically equivalent to deriving a covariance matrix covering all securities, using the historic returns on each security, and then projecting this matrix onto a lower dimensional space (in a manner that equates to regressing these return series versus whatever are the base econometric time series being used in the analysis).

6.3.10 So *mathematically*, econometric risk models essentially only differ from statistical risk models in the way that they rank and discard eigenvectors and corresponding eigenvalues. Of course, how they are created is different. With statistical models, the explanatory variables — the principal components — emerge endogenously from the variance/covariance matrix, whilst with econometric risk models, they are selected on a priori grounds, but the point is that the econometric time series most likely to be incorporated in an econometric risk model are ones that correspond (in aggregate) with a significant fraction of the leading eigenvectors, so the two modelling approaches should actually produce relatively similar results, to the extent that they are being based on the same underlying return series. What econometric risk models really bring to the party is a more intuitive description of the covariance matrix, i.e. primarily presentation rather than underlying mathematical content. Presentation should, however, not be dismissed as unimportant, not least because it makes explanation of the results much easier to non-experts, or, perhaps, it can be argued that, if there is some underlying economic logic to the ranking of eigenvectors, then their use for prospective risk may become more reliable.

6.3.11 A complication is that econometric risk models may not only include factor exposures, but also security specific idiosyncratic elements. In our geometrical representation, the inclusion of these sorts of idiosyncratic elements involves reinserting additional dimensions into the covariance matrix by reinserting non-zero eigenvalues whose eigenvectors largely align with individual securities (i.e. correspond to active weights that are one for a given security and zero for all others). One can, in principle, also do this within a statistical model framework. We discuss below some of the challenges that arise, in practice, in choosing a suitable structure for these 'residual' terms.

6.3.12 Similar mathematics also underlies fundamental risk models. In these models, we exogenously assign factor exposures to individual securities. We then back out the returns on individual factors from these factor exposures by a suitable matrix inversion and projection into a suitably dimensioned space (assuming that there are fewer factors than there are securities, as otherwise the problem becomes ill defined). Once again, therefore, the essential difference (from a mathematical perspective) is in how we rank and discard eigenvalues and eigenvectors (and then, in effect, reinsert back other eigenvectors defining the 'residual' terms). Once again, the factors deemed useful in this process are likely to be ones that have exhibited predictive power in the past, i.e. ones that, in aggregate, span the main eigenvectors that a principal components analysis might generate.

6.4 *Choice of Underlying Distributional Form*

6.4.1 The other main way (mathematically speaking) in which risk models can be differentiated is in terms of choice of distributional form. Here there are potentially larger inherent differences.

6.4.2 Risk models for equity securities often, but not always, assume that returns on individual securities are jointly Normally distributed over suitable time intervals, with the same mean for all securities, and with some suitable *covariance matrix* that defines the joint second moment of the distribution. The use of a common mean involves taking an a priori stance that risk measurement ought not to assume any expected added value from investment 'skill' in an analysis that is attempting to assess the downside implications if that skill fails to materialise. This assumption would, however, be suspect if, say, known charge differentials between the portfolio and the benchmark justified a non-zero differential expected return between them (or if there were liquidity arguments that justified the same conclusion, see Section 10). From a formal mathematical perspective, such risk models are, therefore, completely characterised by their underlying covariance matrix.

6.4.3 Of course, equities do not exhibit perfectly Normal return distributions. The methodology is therefore relying on the active positions within the portfolio being sufficiently diversified for the Central Limit Theorem to bite. This mathematical law states that the probability distribution of the sum of a large number of independent identically distributed random variables tends to a Normal distribution as the number of underlying random variables tends to infinity, subject to certain regularity conditions, such as each random variable having a (known) finite standard deviation. So, you need to treat with caution tracking error and VaR computations for highly concentrated portfolios, as any deviations from Normality may then have less scope to be smoothed away by the Central Limit Theorem.

6.4.4 For many types of bond portfolios, an assumption of Normality is more suspect. One can conceptually split the return behaviour of bonds into several parts:

(a) a part driven by general levels of interest rates (*curve* risk), by which is generally meant prevailing interest rates as derived from yields on relevant well-rated government bonds of different durations (also known as the *government* or *gilt* yield curve);

(b) a part driven by the currency of the bond (*currency* risk);

(c) a part driven by changes in general levels of spreads versus government bond yields, for issues with a similar credit rating as the issue/issuer in question (*spread* risk);

(d) a part driven by spread changes not being uniform across industries/sectors (*industry/sector* risk); and

(e) a residual element arising from issuer-specific idiosyncratic features (*idiosyncratic* risk), mainly the possibility that a particular issuer might default (*default* risk), but also covering other issuer idiosyncratic characteristics, e.g. idiosyncratic yield differentials between different issues from the same underlying issuer (perhaps driven by liquidity considerations).

6.4.5 In essence, exposures to (a) to (d) are similar, in a mathematical sense, to the sorts of equity style factor exposures described above, just translated into bond-speak. For certain types of bond portfolio, e.g. single currency/country government debt, there may be so few factors (or one so dominating, in this instance duration), that the point noted in ¶6.4.3 becomes particularly pertinent. We may then want to spend more time attempting to estimate more accurately the distributional form likely to be relevant to just this one factor.

6.4.6 Portfolios containing significant amounts of credit will, typically, have less to worry about from this perspective. Instead, they will, typically, be exposed to default risk. This sort of risk (for any given issuer) is also highly non-Normal, because of the highly skewed returns that such bonds can deliver, depending on whether or not they default.

6.4.7 The geometrical analogy developed in Section 5 is arguably less effective for highly skewed returns, but it still provides hints as to how we might develop risk models that cater for such skewed behaviour. We might, for example, develop a *granularity-based approach* to risk modelling, by noting that we can decompose the relative return (and the risk) that a credit portfolio exhibits (versus its benchmark) into parts that derive from:

(a) its active 'factor' exposures, as per ¶6.4.4(a) to (d) (e.g. its relative duration, industry positions and exposures by rating bucket), as if the portfolio had *infinitely diversified active credit exposures* within each such dimension; and

(b) its issuer specific credit exposures (relative to the benchmark), as per ¶6.4.4(e), that arise because the portfolio (and benchmark) exhibits *credit granularity*, i.e. is not infinitely diversified as per (a).

6.4.8 For example, suppose that we wish to identify suitable issuer specific limits to apply to a bond portfolio. These, in effect, seek to limit the risks arising from ¶6.4.7(b), not ¶6.4.7(a). So, our focus, when setting them merely needs to revolve around the impact of granularity. We can, in turn, equate this with default risk (or, more generally, rating migration risk, although default risk ought, in some sense, to encapsulate all other ratings migration possibilities, since by maturity either a bond has defaulted or it has not). We might then proceed as follows:

(a) Suppose that the active position (versus benchmark) in bond i with rating R is w_i. Hence $\sum w_i = 0$. Suppose that the annualised probability of default of a bond rated R is $p(R)$ and the recovery rate in the event of default is $y(R)$. For simplicity, we assume that all bond defaults and recoveries are independent of each other, and hence uncorrelated (an assumption that is not, in practice, accurate, see Section 9).

(b) We note that the sum of independent random variables that take the value of $w_i(1-y)$ (i.e. the actual loss that we would suffer if the ith bond defaulted), with probability p and 0 otherwise has a variance of:

$$\sigma^2 = \sum w_i^2 (1-y)^2 p(1-p).$$

This simplifies to $\sigma^2 = nw^2(1-y)^2 p(1-p)$ if we have n such bonds, each with the same weight w.

(c) Consider now what happens if $L(R)$ is the maximum allowable holding of an issuer (as a percentage of the portfolio) carrying a rating of R.

Suppose that we are investing a proportion z of the portfolio in differently rated debt, and that, when doing so, we will use a fraction k of the relevant rating dependent limit (where k will depend on the strength of our conviction and we assume is independent of R). Then n, the number of holdings in which we invest this z will depend on the rating, namely $z/(kL)$. We have $w = kL$, and the contribution to variance from the granularity from these holding is $C = zk(1-y)^2Lp(1-p)$. For investment grade debt, p should be reasonably small, so $(1-p)$ should be close to 1, and C then simplifies to $C = zk(1-y)^2Lp$.

(d) The most appropriate choice of limit structure is one that is indifferent between ratings for the same level of conviction regarding return outcomes, i.e. has C the same for each rating category. If the recovery rate is the same, then this implies that L should be inversely proportional to p.

(e) This approach can, for example, be used to justify limits that scale approximately 1:2:3 for BBB:A:AA rated corporate debt, since the recovery rates for these sorts of debt instruments seem to be reasonably similar, and default rates for most terms seem to scale in approximately the inverse of these ratios for these sorts of bonds (see Table 4). There is some dependency on term, which we might, perhaps, also reflect in the limits applied to individual issuers.

6.4.9 This sort of analysis does not, by itself, identify how large, in absolute terms, the limits should be, only their ratios. Overall limit sizes will depend, in part, on the overall level of outperformance, and hence risk which we might want the portfolio to target, on how much of this we want to come from issuer selection (presumably reflecting what the manager is perceived to be good at), and on how much of a given limit a manager might typically expect to use (this will depend on the typical conviction levels the manager exhibits and how he or she expresses them, versus pre-set limits within the portfolio). However, it does provide a way of estimating the contribution to tracking error from granularity, using the formula for Q

Table 4. Historic corporate bond recovery rates and annualised cumulative default rates

Credit rating[1]/[2]	Recovery rates (for differing years before default)[1]		Annualised default probabilities (for differing bond terms)[2]		
	3 years	5 years	3 years	5 years	10 years
Aa/AA	30.8%	41.1%	0.13%	0.15%	0.20%
A/A	42.0%	45.7%	0.18%	0.22%	0.31%
Baa/BBB	43.7%	38.1%	0.40%	0.51%	0.63%

Source: Threadneedle and[1] Moody's (1982-2003), [2]S&P (1970-2003) described above. It also provides a means of estimating the distributional form of this contribution, since Q is distributed according to a multivariate binomial distribution (which tends to a Normal distribution as the granularity tends to zero, because of the Central Limit Theorem).

6.4.10 Of course, we made certain simplifications in the above analysis. We focused on a single (assumed constant) annualised default rate for each rating bucket (and term). This may not fully reflect potential sources of idiosyncratic return dispersion (including views on potential ratings migrations). We also assumed no correlation of default between different issuers. A more sophisticated approach might be to assume some correlation, akin to the *diversity scores* which ratings agencies use to rate *collateralised debt obligations*, see Section 9.

6.4.11 Care regarding distributional form may also be needed with options (or, more generally, any instruments that contain optionality). These instruments, like bonds, also generally have significantly skewed return profiles; indeed, this is normally their underlying attraction, but, the extent of 'optionality' that a portfolio (or a benchmark) exhibits in this context is not always easy to identify. If a portfolio were, say, to consist of lots of different options, each one small in isolation and each one uncorrelated with each other, then the Central Limit Theorem is still likely to apply. In contrast, a portfolio consisting of a single, but large, relatively short-dated at-the-money index call option would almost certainly exhibit significant optionality in this context. So also would a portfolio consisting merely of short-dated at-the-money call options on each individual security within a typical market index. Usually, much of the movement of individual securities is explained by how the index, as a whole, moves, so such options will typically move in tandem.

6.4.12 A special case of choice of distributional form is that underlying a *historical simulation*. Here, in effect, the distributional form is exactly the distribution observed in the past (subject typically to the adjustment that all return series have the same mean). See ¶5.4.6 regarding the robustness of this sort of approach.

6.5 *Refinements to Time Series Risk Models*

6.5.1 Most commercially available risk systems can be categorised into one of the above forms. In essence, one can view time series risk modelling as an example of the more general problem of forecasting the characteristics of return series, see Appendix E, but applying the constraint that all assets (and liabilities) must have same mean return, see ¶6.4.2. Most refinements to such models are essentially ad-hoc in nature, although, again, such refinements can normally be considered as special cases of tools that also have more general application within return forecasting activities.

6.5.2 A common complaint levied at risk systems is that they typically understate (or overstate, depending on the time period) the overall risk characteristics of a portfolio. This may partly reflect cognitive bias. People, typically, remember those times when the estimated tracking errors significantly mis-state actually observed tracking errors more than they remember the times when they are closer (even though tracking errors are statistical tools and therefore necessarily subject to error) but it also reflects the *heteroscedastic* nature of most financial time series, i.e. that they seem to exhibit time varying volatilities. Various time series analysis tools can be used in an attempt to make risk models more responsive to such time varying characteristics, e.g. use of GARCH (i.e. *generalised autoregressive conditional heteroscedastic*) processes that seek to predict the current level of, say, index volatility (or perhaps even sector or security volatility) from its recent past, and then to adjust the covariance matrix in a manner consistent with this forecast. Correlations also appear to exhibit time varying characteristics (i.e. do not appear to be stable over time). GARCH style modelling may also be used in an attempt to capture their dynamics. More sophisticated approaches, akin to those used for return forecasting, could also be used, see Appendix E.

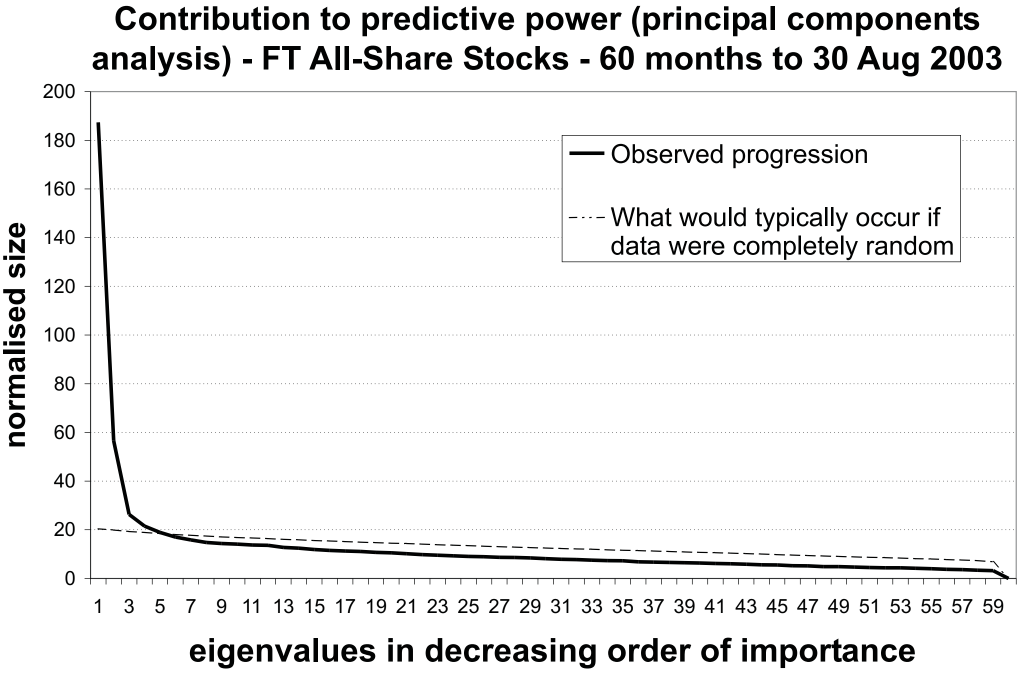
6.6 *Inherent Data Limitations Applicable to Time Series Risk Models*

6.6.1 A major issue that afflicts all time series risk models, and indeed time series return forecasting more generally, is the *sparsity of the data available for the task*. We are used to thinking that there is a veritable cornucopia of data relating to financial markets available from brokers or via third party data vendors. How then can there be insufficient data for such purposes?

6.6.2 High dimensional vector spaces are incredibly large. If, for example, one could conceive of a 100-dimensional cube, each side of which is ten units long, then its volume would be 10^{100}. As an aside, we note that the distance between two randomly selected points within such a cube would then be roughly Normally distributed, with mean 41 and standard deviation 2.5, which means that the likelihood of coming across two randomly chosen points that are substantially closer together than any other two randomly chosen points is very low. In our geometrical representation of risk, each instrument, in principle, creates a new dimension, and there are hugely more than 100 securities traded in the global market place (let alone all the liabilities that might also be considered if, say, we treated each individual insurance policy as a separate 'instrument'). The vector space describing all such instruments is truly vast!

6.6.3 For time series risk modelling purposes, an even more important constraint is the limited history available to us. Suppose, for example, that I wish to develop a risk model for the S&P 500 index or the FTSE All-Share index using monthly returns over the last five years. Ignoring, for the moment, that not all of the 500 companies in the S&P index will have a complete five-year history, the covariance matrix for the S&P 500 would appear to have $500(500+1)/2 = 125{,}250$ separate terms, and for the FTSE All-Share quite a few more (as it has more constituents). However, both covariance matrices only have, at most, 59 non-zero eigenvalues. This is because we can replicate every single return series using linear combinations of 60 different base series (the jth return series having a one in month j and a zero in every other month), and one further degree of freedom drops away, because, for risk purposes, we a priori assume a common underlying mean for all the security returns. This means that the covariance matrix is embedded in, at most, a 59 dimension vector space, however many securities we are analysing. Even fewer will exist if some of these 59 degrees of freedom are 'consumed' by incorporating time varying behaviour within the model. The true underlying probability distribution describing the joint behaviour of different securities almost certainly contains many more factors, *but there is no possible way of identifying any of the remaining factors from the historic return data alone*.

6.6.4 Even this understates the magnitude of the estimation problem. If we actually analyse the observed eigenvalues, we discover that nearly all of them appear to be little different to what might arise purely by chance, see Figure 4. This plots the magnitude of the observed eigenvalues of the covariance matrix described above for the FTSE All-Share index (giving



Source: Threadneedle, Thompson Financial Datastream

Figure 4. Contribution to predictive power (principal components analysis) for FTSE All-Share index; 60 month's data to 30 August 2003

equal weight to each month's return in the computation of the covariance matrix), ordered by size, against those that would typically occur by chance, even if all the securities were independent Normally distributed returns (each scaled to have the same volatility). Some of the observed eigenvalues from the random data will be randomly larger than others, so they, too, will show a declining pattern when ordered by size. Only perhaps five to ten of the eigenvalues appear to be obviously different to what you might observe by chance. Using weekly data does not increase the number of obviously statistically significant eigenvalues by very much.

6.7 *Other Practical Challenges*

6.7.1 The inherent mathematical limitations described above make choice of idiosyncratic risk elements in time series risk models highly subjective. In essence, there simply is not enough data to estimate these contributions reliably. One could, in principle, attempt to overcome this problem by increasing the time period over which we carried out our analysis, but long ago individual securities may have had quite different dynamics (if they existed at all!). There comes a point where what you might gain in this respect will be lost, because the information is more out-of-date. Moreover, the challenge of working out what to do for securities that do not have a complete history becomes greater. Often risk systems allow users to choose how to 'fill in' such missing data, or they merely model aggregates, such as industry/sector/rating bucket/duration bucket portfolios, which do have complete histories, and ignore or otherwise guess at the idiosyncratic risk characteristics of securities within these buckets.

6.7.2 It is important to bear these fundamental limitations in mind when using optimisers based on time series risk models, see Section 8, as it means that some of their answers are more subjective than appears at first sight. It also highlights some of the challenges which arise if we want to develop a risk model that can simultaneously estimate risk well, both for a broad global or regional portfolio and for a narrower market segment, e.g. just securities in some individual sector within a single country. The problem is that the eigenvalues (or to be more precise the corresponding eigenvectors) that work well at the big picture level are unlikely to be the same as the ones that work well at every single micro level at which the model might be used.

6.7.3 We glossed over a subtle point earlier in this respect. We presented principal components analysis as if there were only one way of extracting the 'most important' eigenvectors, and hence drivers to observed return series. However, suppose that, instead of using the raw return series, we scaled one by a factor of 1,000, and only after doing so calculated the covariance matrix on which we then carried out our principal components analysis. It has the same number of non-zero eigenvalues (and hence eigenvectors) as before, but greater weight is now given to the security return series which we scaled up in magnitude. In this instance, the largest eigenvalue would be almost identical to this security's return series (adjusted to have the same mean as the other return series). Time series risk modelling actually includes a weighting schema (which, for simplicity, we assumed involved equal weights for each security being analysed). When attempting to create risk models that simultaneously cater for widely different types of portfolios, we, in effect, ideally want to use different weighting schemas for each discrete type of portfolio. We will, therefore, be pretty lucky if we get consistent risk models across all possible such weighting arrangements.

7. Derivative Pricing (or Fair Valuation or Market Consistent) Based Risk Models

7.1 *An Alternative Approach to Risk Modelling*

7.1.1 The framework that I described in the previous section could probably be categorised as *conventional wisdom* as far as risk modelling is concerned (albeit you do not often hear much about its fundamental limitations, and different risk system providers have the unsurprising tendency to trumpet their own particular variant over all others). So, readers may be surprised to discover that its whole theoretical framework is potentially shaky.

7.1.2 We will demonstrate this from three different angles. First, we consider some ad-hoc refinements which we could make to the granularity-based risk model described in Section 6.4. These hint at a conceptually quite different framework that might be adopted. Next, we discuss what might otherwise appear to be an aside regarding the most appropriate time horizon to adopt for risk measurement purposes, discovering that this, too, points to a different sort of framework. Finally, we analyse more explicitly the interaction of risk measurement, fair valuation theory and derivative, pricing to put these hints onto a firmer theoretical foundation. What we discover is that risk theory, itself, needs to be re-evaluated in the light of fair valuation principles. In theory, this allows us to circumvent the inherent limitations on time series risk models, noted in Section 6.6, although, in practice, there still are rather less data than we would like to formulate risk models.

7.2 *Refinements to the Granularity-Based Approach to Risk Modelling*

7.2.1 The approach to risk modelling which we developed in Section 6.4 seems to offer some means of circumventing the inherent data limitations that otherwise plague time series risk models. Via it, we created a structure to apply to the 'residual', or idiosyncratic, risk affecting individual securities/issuers, even though elsewhere in Section 6 we noted that there are not enough historic time series data to allow us to estimate reliably these residual contributions.

7.2.2 The granularity-based approach did so by referring to an exogenous characteristic of the instrument in question, i.e. its credit rating, which had a natural link to the credit's idiosyncratic risk (via default risk). To do so, we needed to make a number of assumptions. Specifically, we glossed over ratings migrations, assuming that they were all encapsulated in some way in default and recovery rates. We also needed to rely on some exogenous mapping of rating to likelihood of default, as provided by a rating agency's historic data.

7.2.3 How accurate are the assignments of ratings to credits by rating agencies, and how relevant are default histories to the task of assessing how likely a bond is to default in the future? If we dislike *historic default rates*, then we could, instead, use *market implied default rates*. Credit rating agencies can take some time to reflect in their ratings what is happening to a particular credit, so individual bonds can trade as if they have a different rating to the one assigned to them by the ratings agencies. Market implied default rates are easily derived from market data. As Schönbucher (2003) notes, they are essentially the same as the bond's *credit spread*, i.e. its yield spread versus an equivalent risk free bond of similar currency and duration. We leave to Section 10 the question of what we mean by *risk free* in this context. Market implied default rates are 'risk neutral' (using derivatives pricing terminology) and, of course, 'market-consistent'.

7.2.4 Default rates are a rather bond orientated concept. Are there any corresponding market implied data that are relevant for equities? Yes. The financial theory of *firm valuation* demonstrates that bond default rates and equity volatility, in some sense, form two sides of the same coin, see Schönbucher (2003). This was recognised even within pioneering papers on derivative pricing, such as the celebrated Black & Scholes (1973) paper (which, interestingly, is actually titled 'The pricing of options and corporate liabilities') and Merton (1974). So, to create a market implied granularity-based risk model encompassing equities, we should derive an individual equity's idiosyncratic risk from its implied volatility (assuming that there are options trading on it), or, failing that, from its credit spread (if available) and some analysis based on firm value theory.

7.2.5 Such an approach is conceptually a huge step forward compared with a pure time series model. It enables us, in principle, to populate all of the idiosyncratic risk components that were otherwise out of reach using purely time series data.

7.2.6 And why stop at idiosyncratic risk? In principle, we ought also to be able to identify market implied idiosyncratic cross-correlations, and, if we have all of these too, then we have an entire *market implied covariance matrix*. In practice, relative performance options between two different individual securities rarely, if ever, trade, although some of the more complicated structures used to back some retail products are sensitive to market implied average correlations between baskets of securities (*correlation*, in the parlance of the derivatives markets), which might help here.

7.2.7 Hold on, you might say, we cannot, in practice, observe this level of detail, so surely we are back to using historic observed correlations, etc, but there is a flaw with such reasoning. Financial markets have shown tremendous innovation over the last few decades. In certain markets it is now possible to identify from market observables some of these data. For example, in equity land, one can infer an approximate average level of correlation from the relationship between index implied volatilities and single stock implied volatilities (and, indeed, there seem to be hedge funds that are attempting to arbitrage between the two). In bond land, prices of different CDO tranches (see Section 9) are sensitive to average implied correlation between defaults on different bonds, and so one can infer information about average implied correlations from this market place. Indeed, close analysis of these data indicates that there is currently a correlation 'smile' linked to the subordination level of the tranche; and who knows how much more will become possible over time.

7.2.8 The point is that the philosophical basis of derivatives pricing based risk modelling is quite different to that of time series based risk modelling. With time series risk modelling, we extrapolate the past to identify how the future might behave; but, with derivative pricing based risk modelling, we infer, where possible, how risky the future might be from current market observables. Only if we cannot find current market observables do we fill out the missing data via general reasoning (which, in this instance, might typically involve use of historic data).

7.2.9 Understanding the difference becomes particularly important in situations which are sensitive to how the two methods differ, or for instruments which, in effect, arbitrage between the two methodologies. It is, as we shall see in Section 9, highly relevant to CDOs.

7.3 *Return Horizons in Risk Management Tools*

7.3.1 Forward looking risk measures can be *horizon dependent* or *horizon independent*. VaR is naturally a horizon dependent measure, e.g. we might be interested in a five-day (rather than a one-day, 15-day, One-month, ...) 95% confidence level VaR, in which case its *return horizon* is this five-day period.

7.3.2 Tracking errors are less often horizon dependent. They are usually annualised, but this is actually a quotation convention. The underlying logic behind the convention is to assume that the return distribution through time has *stationary* second moments. This means that the standard deviation of the log of the return between t_1 and t_2 is dependent only on the length of the time period. In continuous time (subject to suitable regularity conditions), it would be proportional to $\sqrt{t_2 - t_1}$. The standard quotation convention therefore involves annualising tracking errors derived from, say, monthly returns, by adjusting the square root of time. Tracking error engines that quote a single annualised tracking error are implicitly assuming that the tracking error applicable to any other return horizon can be derived by the same square root of time convention but in reverse. However, some risk systems do quote horizon dependent tracking errors, e.g. 4% p.a. for yearly returns, but 5% p.a. for monthly returns.

7.3.3 Conventional risk modelling wisdom holds that we ought, in principle, to be interested in a *time horizon*. The argument goes that we are interested in risks that we might experience between now and when the portfolio might change, since, if the portfolio is completely realigned, then its risks become completely different. Such wisdom thus holds that the most appropriate time horizon to focus on for risk measurement and management purposes ought, in principle, to depend on your investment analysis and/or decision making timescale/time horizon. A short time horizon (e.g. days, or even hours or minutes) might be appropriate for a hedge fund, a somewhat longer one (e.g. weeks or months) for a more traditional ('long only') asset manager, and a longer one still (quarters or years) for a longer-term investing institution, such as a pension fund, when considering asset/liability management.

7.3.4 Conventional wisdom also holds that return distributions typically do not have stationary second moments, and so the time horizon which you choose actually makes a difference to the answer. It usually holds that return distributions typically exhibit some *autocorrelation*. In such an assertion, 'autocorrelation' is typically used loosely to refer to any *intertemporal dependency* (even though such dependency does not always relate to the second moment and hence to correlation). There is, for example, some apparent evidence of intertemporal dependency, even for major market indices (and more for some individual securities), see e.g. Table 5. It is not easy to identify how the observations set out in this table might plausibly arise without there being at least some sort of intertemporal dependency, albeit that it might, perhaps, merely be an artefact of how market values for the underlying instruments are being observed and recorded (see Appendix D.4 for a further discussion on this point).

Table 5. Annualised observed volatilities, skewness and kurtosis of (log) returns % p.a. (1 January 1990 to 30 September 2004)

	Day	Week	Month	Quarter	Year	2 years
FTSE All-Share (U.K. equities)						
Volatility	15.1	14.7	14.8	16.2	16.3	16.5
Skewness	−0.15	−0.08	−0.49	−0.77	−0.86	−0.40
Kurtosis	3.52	1.73	0.51	0.77	−0.76	−1.26
FT-Actuaries All-Stocks (gilts)						
Volatility	5.0	5.3	5.4	6.0	7.4	6.6
Skewness	0.04	−0.11	0.00	−0.58	−0.59	1.13
Kurtosis	3.80	1.99	0.34	0.62	−0.40	0.98

Source: Threadneedle, Thompson Financial Datastream

7.3.5 However, there are some puzzling features with conventional wisdom. If such autocorrelation really did exist, then why has it not been arbitraged away more effectively over time? Perhaps its existence, or non-existence, is less important than conventional wisdom might imply is the case; autocorrelation does not necessarily have to be inconsistent with efficient markets, if you posit a time varying 'price of risk'.

7.4 *An Apparent Aside: Mileage Options*

7.4.1 In my opinion, one of the most powerful conceptual tools relevant to understanding option pricing theory is a hypothetical (total return) option called a *mileage option*, explored by Neuberger (1990), and referred to in Kemp (1997). An analysis of it succinctly encapsulates essentially all of the ways in which the fair value of a derivative instrument can diverge from the celebrated Black-Scholes option pricing formulae and their extensions. We concentrate on 'total return' derivatives. An example would be a European-style put option which gives the holder the right to sell at time T an index, with gross income reinvested, represented by S_t, for a price set by reference to an initial reference index level F_0, rolled up (as F_t) in line with the total return on, say, cash. The pay-off of such an option at time T is, in effect, $\max(E.F_T/F_0 - S, 0)$, where E is the strike price of the option.

7.4.2 The unusual feature of a mileage option is that it expires, not at some fixed time T, but when the *cumulative quadratic variation* of the option $CQV(t)$ reaches a certain predetermined value $CQVT$. The cumulative quadratic variation $CQV(t)$ is defined as the limit, as the partitioning into separate time steps tends to infinity, of:

$$\sum_{i=0}^{t}\left(\log\left(\frac{S_{i+1}}{S_i}\Big/\frac{F_{i+1}}{F_i}\right)\right)^2.$$

7.4.3 The no arbitrage pricing formula for this option is particularly simple, if we assume markets are arbitrage free and are frictionless (i.e. no transaction costs, no limits on short-selling, borrowing, etc.). If the evolution of $S(t)$ is constrained, so that $CQV(t)$ is always continuous, then the value of this option is, see Neuberger (1990) or Kemp (1997):

$$P(S,t) = E.N(-d_2) - S.N(-d_1)$$

where:

$$d_1 = \frac{\log(S/E) + CQVT/2}{\sqrt{CQVT}}$$

$$d_2 = d_1 - \sqrt{CQVT}$$

and

$$N(x) = \text{cumulative normal distribution function.}$$

7.4.4 This result can, perhaps, be most easily understood by remembering that the Black-Scholes formula can be derived as the limit of a binomial tree pricing approach, as per Figure 5, in the limit where h tends to zero, if $\log(u) = \sigma\sqrt{h}$ and $\log(d) = -\sigma\sqrt{h}$, see e.g. Kemp (1997). With a mileage option, we redefine the rate at which 'time' passes, i.e. the step size, to match the rate at which the cumulative quadratic variation changes.

7.4.5 Thus, as Kemp (1997) points out, in a no-arbitrage world option prices only fundamentally diverge from a generalised sort of Black-Scholes framework because of:

(a) *market frictions* (e.g. transaction costs);
(b) the *stochastic nature of volatility* (i.e. uncertainty in when 'real time' $CQV(t)$ will reach $CQVT$); and
(c) the *possibility of market jumps* (i.e. that $CQV(t)$ might not be continuous).

7.4.6 It seems to me that this analysis has a number of important corollaries for risk management. What is the most important purpose behind calculating VaR or the like? If it is merely comparison versus others, then we have already highlighted that orderings are relatively insensitive, as long as we are all focusing on the same sort of risk. If it is to do with our own assessment of the likelihood of being hit by an adverse event, then using 'real world' probabilities perhaps makes sense, although we need to remember that they do not reflect the different utility (i.e. risk aversion) that we might assign to upside versus downside outcomes. However, I would argue that the main purpose of VaR is ultimately to help with identification of capital requirements, or in the allocation of capital. It is, in some sense, attempting to quantify the capital charge incurred for a given risk. We ought, therefore,

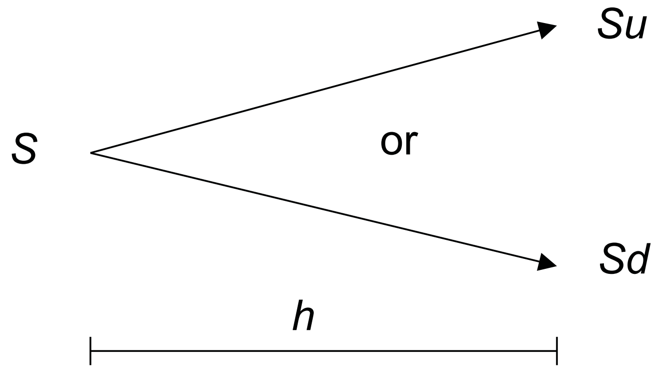


Figure 5. Diagram illustrating binomial option pricing

to be primarily interested in the *fair market price for the relevant risk*. If we want to buy an option to, say, protect ourselves against losses exceeding our VaR (or to cover your TVaR, see Section 5.5), then its *fair price* will depend on:

(a) what positions we hold (and how they might change); and
(b) the *risk neutral* probability of the loss occurring (and *not* its real world probability).

7.4.7 The reason why mileage options are not, therefore, an aside is that they highlight, very effectively, that the relevant risk neutral probability depends on the behaviour of the cumulative quadratic variations of the underlying securities, and *not* on how autocorrelated their *real world* behaviour is. So, the actual (real world) autocorrelation exhibited by return series is largely irrelevant for this sort of use of VaR.

7.5 *Fair Valuation, Derivative Pricing and Risk Measurement*

7.5.1 The above two lines of argument both include reference to derivative pricing. This is no coincidence. Two obvious axioms that we might impose on how we might value things are *additivity* and *scalability*, i.e. that if A is worth V_A and B is worth V_B, then $k(A+B)$ is worth $k(V_A+V_B)$. Nearly any valuation framework that we might wish to adopt from a risk management perspective is likely to satisfy these axioms. It is possible to envisage situations where the sum of the parts may be greater than the whole (e.g. the usual justification for one company bidding for another is that the bid will create synergies), but whether it is ever appropriate to reflect such possibilities in risk management, until after they have been realised in terms of market price movements, is unclear to me.

7.5.2 These axioms are also extremely powerful. If $k=0$, we conclude that the value of nothing is nothing. This should not be contentious; but suppose that we now have a zero coupon bond Z, paying one in T years' time, and we have a range of non-overlapping, but mutually exhaustive *digital call spread* options $Q(E, E+dE)$, $Q(E+dE, E+2dE), \ldots$ on some underlying $S(t)$, with the same maturity date. By $Q(E, F)$ we mean an instrument that pays out in T years time a sum of one if $ES(T) < F$ and zero if $S(T)$ is outside this range. Suppose that we assign values to these instruments that satisfy the above axioms, i.e. a function $V(.)$, that, at any given time, is well defined both for Z and for all of these digital call spread options. We note that, if we go long one unit of each of these digital call spreads, and go short one zero coupon bond, then the pay-off is zero, and hence so is its value. This implies (in the continuous limit) that:

$$\int_0^\infty V(Q(E, E+dE))dE = V(Z).$$

We can therefore derive a function (defined on the range of possible outcomes for $S(T)$) $p(x) = \lim_{dx \to 0} V(Q(x, x + dx)) \big/ (V(Z)dx)$ that satisfies all the requirements of a probability measure. This is the *risk neutral probability measure* for $S(T)$ for this valuation framework.

7.5.3 The underlying principle of derivative pricing is that of *no arbitrage*. In frictionless markets, the principle of no arbitrage is essentially equivalent to the combination of the above two axioms (additivity and scalability) and the use of valuations that equate the value of the instruments underlying the derivatives with their *market prices*. (Also, one needs to assume that if the payoff of an instrument A is ≥ 0 with probability one, and the pay-off is not identically zero, then $V(A) > 0$.)

7.5.4 A special case of derivative pricing is the pricing of *delta-one* derivatives, i.e. ones that move one-for-one in line with the underlying instrument, of which the simplest is the instrument itself. So if no arbitrage applies, then fair valuation theory is merely a special case of derivative pricing theory, calibrated to match observed market prices where they exist, but using the valuer's best estimate of what such market prices would be where market prices do not exist.

7.5.5 This explains why you end up in all sorts of knots if you do not fair value derivatives. If derivatives are *not* fair valued, then the valuation framework which you adopt will *not* satisfy the above axioms. For example, hedge accounting, which associates derivatives with any instruments that they might be hedging (and links the value of the derivative with say the book cost of the instrument hedged), in general, does not satisfy these axioms. If I hold the same derivative twice, once to hedge one instrument and once for some other purpose, the value which I ascribe to the former in a hedge accounting framework will, in general, be inconsistent with the value which I ascribe to the latter, even though the two instruments are identical!

7.5.6 And the link with risk measurement? In ¶5.3.3, astute readers will have noticed that we assumed that if we had two different sets of positions **a** and **b**, with corresponding future returns A and B, then we claimed that the combination of the two $\mathbf{c} = \mathbf{a} + \mathbf{b}$ created a corresponding future return $C = A + B$. This necessarily requires us to assume that the values that we ascribe to different instruments are additive (and elsewhere in the same section is embedded an assumption that values are scalable).

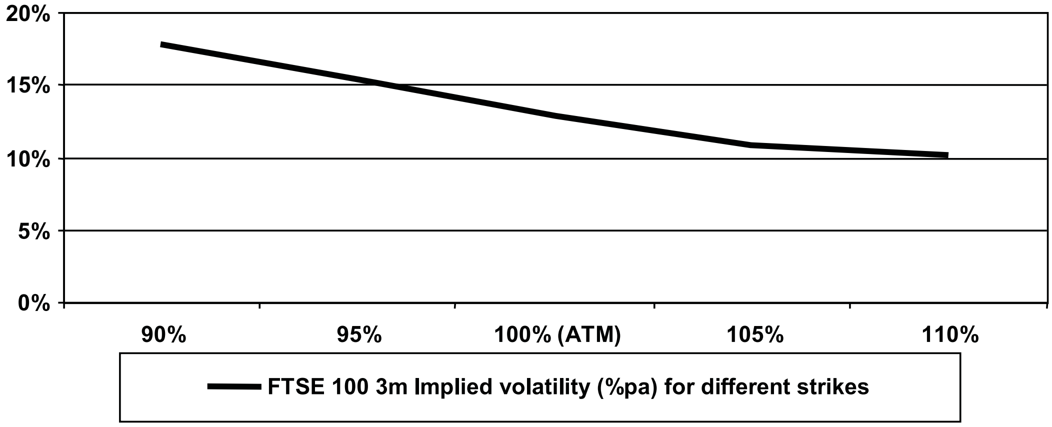
7.5.7 Thus, the whole of the risk measurement framework which we have described to date in this paper theoretically *requires us* to adopt these axioms, and hence a derivative pricing based approach to risk modelling!

7.5.8 Of course, creating a pure derivative price based risk model is, in practice, impossible, due to inadequate market data. Fair valuation methodology (like its more general analogue, derivative pricing) is heavily concerned with extrapolating from the observed to the unobserved. In it we develop models that we think are reasonable, and then calibrate them, so that they give back observed market prices for those instruments where such prices are observable. Approximations are acceptable as long as they are appropriate in the context of the answer or the purpose to which the methodology is being put; and so it is with risk measurement. However, the above analysis does suggest that relatively straightforward refinements (such as to make some allowance for market implied index volatility, idiosyncratic volatility and general levels of correlation) might usefully be more prevalent than at present in commercial risk modelling tools.

7.6 *Other Observations*

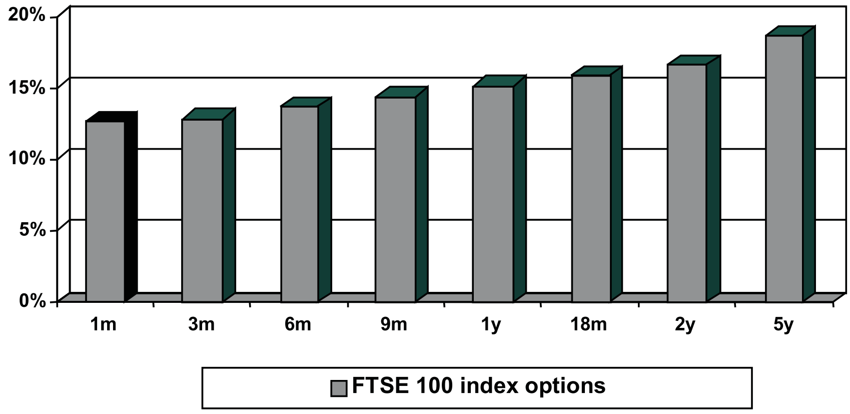
7.6.1 Derivative pricing based risk modelling also sheds further light on the incoherence of VaR, referred to in Section 5.5. The equivalent option to a traditional VaR is a digital option that pays out one if the loss is greater than the VaR and zero otherwise, but life is not often quite this black and white. Is it really plausible to assume that moving from just above to just below the VaR confidence level would really create such a massive realignment in how the company was being viewed in the marketplace? More appropriate, perhaps, is a risk measure with an equivalent option that has a somewhat smoother payoff function, such as TVaR.

7.6.2 It also sheds light on the distributional form to assume for the future return distribution, see Section 6.4, and the most appropriate time horizon to use, see Section 7.3. Option volatilities exhibit a skew structure (i.e. different volatilities for different strikes), see Figure 6, which can, in principle, be used to derive the market implied return distribution. They also exhibit a term structure, see Figure 7, i.e. they vary by maturity date. Option investors are not only taking views on *volatility*, but also on *volatility skew* and *volatility term structure*.



Source: Threadneedle, Morgan Stanley

Figure 6. Three-month implied volatilities (% p.a.) as at 26 October 2004



Source: Threadneedle, Morgan Stanley

Figure 7: At-the-money implied volatilities (% p.a.) for different terms as at 26 October 2004

8. Managing Risk in the Light of Real World Uncertainty

8.1 *Risk versus Reward*

8.1.1 Managing risk involves more than just measuring it. The traditional quantitative workhouse used to help to decide how much risk to take and of what form is *risk-return optimisation*, also, more colloquially, called *efficient frontier* analysis (strictly speaking the ‘efficient frontier’ is merely the collection of portfolios that achieve an optimal level of return for a given level of risk, or vice versa). With suitable time series data, the basic principles are beguilingly straightforward even if the mathematics can get quite detailed in places.

8.1.2 For the sake of variety, we here illustrate the concept by reference to how it might be used in a non-life insurance context. Similar approaches can also be used in life insurance, pension fund and private wealth management, or in almost any other financial services area.

8.1.3 Consider, for example, a non-life insurer with relatively short tail liabilities, that wishes to identify some suitable neutral asset mix (i.e. benchmark) to give to its investment manager. We assume that the insurer wishes to limit itself to (Sterling) investment grade bonds with term $<$ ten years, if A rated (or above), and $<$ seven years, if BBB rated, with duration and credit rating mix chosen so as to best trade off risk against return, expressed by reference to a suitable neutral weighting in a mixture of Merrill Lynch bond indices. We note that, even though we appear to have defined the problem in a relatively general manner, we have still made some implicit choices via the specification, e.g. we are ignoring other characteristics that bonds might have, such as industry category, and we have excluded from our analysis non-U.K. bond assets. We also assume that the insurer imposes an upper limit on the duration of the portfolio of three years. We assume that 'duration' means a bond's *option adjusted modified duration* (i.e. sensitivity to small parallel shifts in the yield curve), rather than *Macaulay duration* (i.e. weighted average time to payment). The modified duration for a straightforward bond equals the Macaulay duration divided by $1 + i$, where i is the annualised gross redemption yield. Option adjusted here means taking into account any callable or putable features within the bond.

8.1.4 Such analyses often use whatever is the maximum time period available for which complete return series are available for all asset categories, which, at the time of writing (if we use one to three, three to five, five to seven and seven to ten year indices, except for a merged BBB one to five year index), is from 31 March 1998 to 30 November 2004.

8.1.5 Risk/return optimisation proceeds as follows:

(a) Any optimiser requires some definition of *return* (also referred to as *reward*). Typically we might assume that we have a vector $\mathbf{r}$ of (assumed) returns on the different asset categories, and that the return to use in the optimisation exercise is the weighted average of these, weighted in line with $\mathbf{x}$, the desired benchmark asset mix.

(b) The optimiser also requires some definition of *risk*. Typically, we might assume that this can be equated with a forward looking tracking error (versus some suitable minimum risk position $\mathbf{b}$, based on a covariance matrix, as described previously.

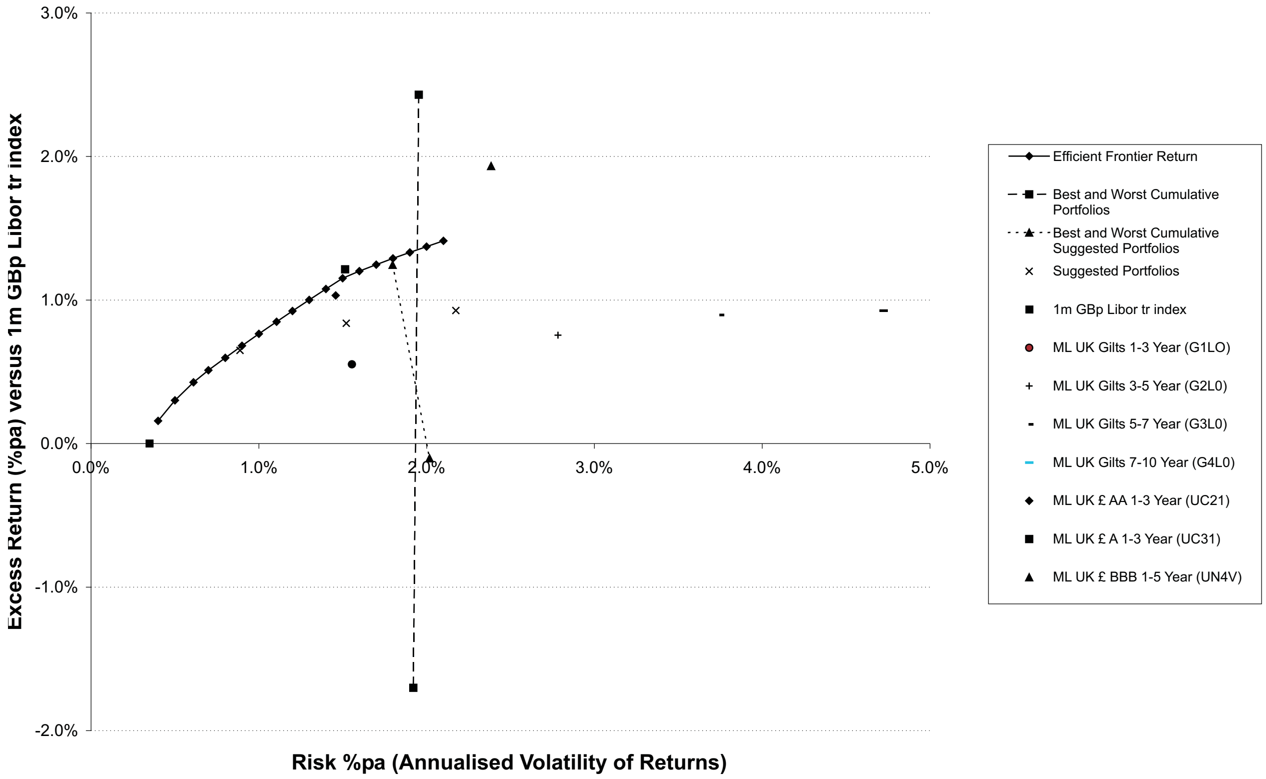
(c) The optimisation exercise then mathematically involves maximising, for some range of risk aversion λ, some risk/reward trade-off (or *utility*) function, subject to some *constraints* on the portfolio weights.

(d) The same optimal portfolios arise for any other utility function that monotonically increases as the risk metric in (b) increases. So, for example, we get the same efficient portfolios whether we use the forward looking tracking error, any VaR statistic that has been determined using a Normal distribution approximation from the same underlying covariance matrix, the variance (i.e. square of tracking error) or variance plus a constant (e.g. to reflect sources of risk independent of any assets deemed available to the investment manager). Most commonly a quadratic utility function is used, such as $U(\mathbf{x}) = \lambda \mathbf{r}.\mathbf{x} - (1 - \lambda) \times (\mathbf{x} - \mathbf{b})^T \mathbf{V}(\mathbf{x} - \mathbf{b})$, together with linear constraints $\mathbf{A}.\mathbf{x} \leq \mathbf{P}$ (typically including the two constraints $\sum x_i \leq 1$ and $-\sum x_i - 1$ to force the asset weights to sum to unity and, for long only portfolios, $x_i \geq 0 \Rightarrow -x_i \leq 0$). The advantage of this utility function is that if $\mathbf{V}$ is a positive definite symmetric matrix (which it will be if it is derived directly from historic data), then the exact solution can be found relatively easily, using a variant of the Simplex algorithm or other standard algorithms for solving *constrained quadratic optimisation* problems, see e.g. references quoted by Press *et al.* (1992). We can likewise use any other return function that monotonically increases as the return metric in (a) increases, e.g. the excess return over a base value, such as the return on cash.

8.1.6 Suppose that this particular client has specified that risk is to be measured in nominal terms (i.e. $\mathbf{b} = 0$), and that the maximum exposure to BBB paper is 10%, to A and BBB combined is 30%, and to AA, A and BBB is 70%. Suppose, also, that the assumed returns to be used in the optimisation problem are the observed annualised returns over the period 31 March 1998 to 30 November 2004, and that the covariance matrix used to measure risk is based on the observed covariance of monthly returns over this period. Then, the optimal portfolios (i.e. the efficient frontier), their risk/return characteristics and the risk/return characteristics for portfolios invested 100% in individual asset categories are as per Figures 8 and 9, quoting return in terms of excess over cash and risk in terms of nominal volatility.

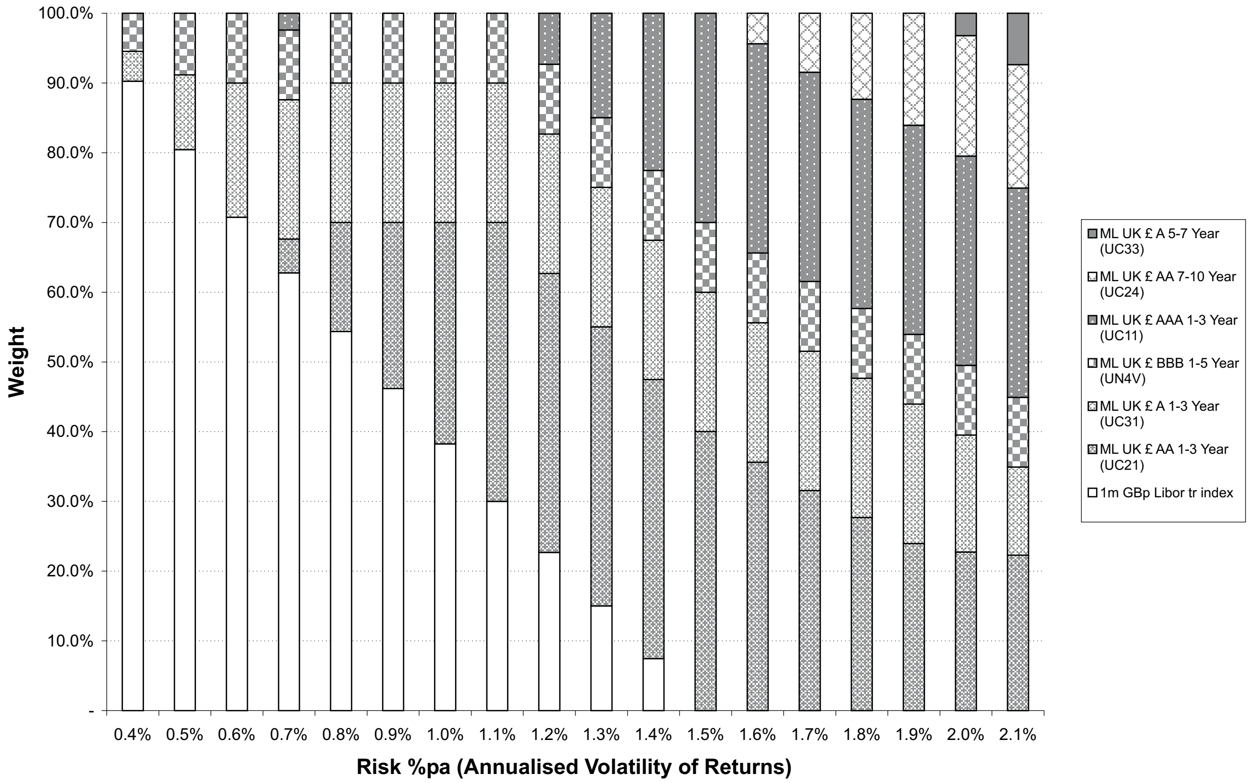
8.1.7 This example illustrates one of the problems with using historic returns to estimate future returns. The time period used happens to have coincided with credit spreads narrowing. So, it is not surprising that the efficient portfolios focus on credit rather than on gilts.

8.1.8 A practical complication that almost inevitably arises is how to translate risk, as measured by a relatively mathematical concept such as tracking error, into something that is easier for clients to understand. There are almost endless ways in which this can be done, e.g. we might calculate



Source: Threadneedle, Bloomberg, Merrill Lynch

Figure 8. Efficient and other portfolios in illustrative example



Source: Threadneedle, Bloomberg, Merrill Lynch

Figure 9. Weights in efficient portfolios in illustrative example

more intuitive VaR type statistics, e.g. the worst 95%ile outcome over a given year (or even simply the worst outcome that the efficient frontier had generated for any given calendar year within the historic time period); or we might carry out simulations (also known as *stochastic projections*) showing the range of future outcomes of different metrics of interest (e.g. solvency, funding level, contribution rate, ...) for different efficient portfolios, this being a common approach for long-term investing institutions.

8.1.9 We can also use the observed return series for various other purposes. For example, we might illustrate the best and worst outcomes that might have arisen within the client's constraints (see also Figure 8, which shows the best and worst annualised returns in excess of cash, and their volatility, had the mix been rebalanced at the start of each calendar year and had we had perfectly accurate or inaccurate foresight for the following year). This can conveniently be determined by reuse of a quadratic optimiser each year in isolation, by giving little weight to the risk element; or we might highlight the risk/reward characteristics of specimen benchmark asset mixes that we might have decided, from more general reasoning, were worth focusing on. Figure 8 includes an example of three such portfolios designed to reflect a core credit exposure and flexibility to take different duration positions using a mixture of cash and gilts.

8.2 *Practical Challenges*

8.2.1 One obvious challenge is how to decide how much risk to adopt. There is no right answer here; it depends on the client's risk appetite.

8.2.2 An even bigger practical challenge is the extreme sensitivity of the answers to the input assumptions, i.e. the returns, covariance matrix, constraints and minimum risk portfolio. Dependency on the constraints is relatively obvious and easy to explain to lay people. If the unconstrained result implied a 50% weighting to a particular asset category, and we limit it to 10%, then, not surprisingly, the answers differ by 40%. Dependency on the minimum risk portfolio is also fairly obvious, at least for less risky portfolios. The dependency on the other parameters may be less obvious, but is no less significant. Mathematically, the optimum (in the absence of constraints) lies at a point where all the partial derivatives of the utility function are zero, i.e. if one were to visualise the utility function in multi-dimensional space, then the optimum lies at a point where the utility function is absolutely flat. The situation reminds me of 'Labyrinth', a children's game, in which you attempt to navigate marbles around a flat maze board with holes in it by tipping the edges of the board slightly up or down. It requires significant manual dexterity!

8.2.3 This problem is, perhaps, particularly acute for the return assumption. In the above example, we adopted as return assumptions the observed historic returns. There is the philosophical issue of whether it is right to estimate future returns merely from past data. But even if we are happy to do so, we omitted to mention that the mean historic return is based merely on some sample from some underlying 'true' historic return distribution for the asset category in question, and so is subject to sampling error. The historic returns in our example have been derived from 80 monthly returns, so each (if they were independent of each other) has a standard error approximately equal to $\sigma/9$, where σ is the volatility of the relevant asset category. Such errors are enough to move the precise composition of the efficient frontier by a significant amount. For example, if the AA one to three year index had an assumed return lower by this amount (i.e. 6.1% p.a. rather than 6.3% p.a.), all other assumed returns being left unchanged (admittedly a somewhat unrealistic example), then its weight in the efficient frontier drops to zero, having previously been, at times, 40%!

8.2.4 One way of trying to mitigate these sorts of problems is to introduce some *anchor* that constrains the optimisation problem, not just at one point (the minimum risk portfolio), but at another point some way along the efficient frontier chosen from general reasoning. The most obvious is the Black-Litterman approach, which, in effect, assumes that the *market portfolio* is optimal for some level of risk. In mean variance space, the efficient frontier is piecewise linear, each line segment finishing when some new constraint starts to bite. Portfolio weights are piecewise linear in the risk-tracking error space, again with each line segment finishing when a new constraint bites. So, normally, a Black-Litterman approach results in the optimal portfolios being some mix of the minimum risk portfolio and the market portfolio.

8.2.5 But what does the 'market' portfolio consist of? Does it include cash, bonds and property, as well as the equities with which it is more normally associated, and, if we include asset classes like cash (or derivatives), where the two parties involved, strictly speaking, have equal and opposite positions, how much weight should we give to each side?

8.2.6 Alternatively, we might adopt Bayesian approaches or other similar tools designed to give only partial weight to historical data. We might, for example, rely wholly on some externally derived estimates of future returns, volatilities and correlations. This does overcome the problem in a certain manner of speaking, but merely introduces another, namely the potential inaccuracies that might exist in our Bayesian 'prior' forecasts.

8.2.7 A final approach might be to avoid optimisation altogether, and instead to use reverse optimisation, i.e. to work out what (typically return) assumptions are needed to justify a given portfolio structure. This is more robust than optimisation, but, of course, we still need to decide what stances to adopt in the first place. In a qualitative investment process, one might develop individual position limits using methodologies akin to those described in Section 6.4, and leave it to the human investment manager to position the portfolio accordingly. With a highly quantitative investment process, there may be no way of avoiding using optimisers.

8.3 *Risk Budgeting*

8.3.1 The sensitivity of optimisers to their input assumptions also has important implications for *risk budgeting*. Risk budgeting involves:

(a) identifying the total risk that we are prepared to run;
(b) identifying its decomposition between different parts of the investment process; and
(c) altering this decomposition to maximise expected value added.

8.3.2 The concept has wide applicability. It can be applied to asset/liability management, manager selection, stock selection, etc. It is a concept that is also difficult to fault. If the risks arising in each part of the investment process are fixed, then it implies focusing the assets on those areas where there is the highest expected level of skill. If the asset split is fixed, then it implies focusing the risk being taken on those areas with the highest level of skill. Skill, here, might be associated with an expected future information ratio, by re-expressing the definition of the information ratio as follows. An advantage of this re-expression is that tracking error can, in essence, be thought of as deriving from portfolio construction disciplines and the information ratio from the skill that the manager exhibits, and so, to first order, it is reasonable to assume that the two are largely independent of each other.

$$\text{Information ratio}(IR) = \frac{\text{Outperformance}(\alpha)}{\text{Tracking error}(\sigma)} \Rightarrow \alpha = IR \times \sigma.$$

8.3.3 The problem that we find in practice is that the results are hugely sensitive to our assumptions about future information ratio. In particular, if we adopt the usual risk manager's starting assumption that the *IR* is zero (see ¶6.4.2), then the answers become ill defined. So, to make any use of risk budgeting, you need to believe that you have some skill at choosing 'good' managers or asset categories (i.e. ones with *IR* versus the underlying benchmark > 0), as opposed to 'poor' ones (i.e. ones with $IR < 0$).

8.3.4 One can also use risk budgeting theory to help define appropriate portfolio construction discipline rules. For example, you might a priori assume, as a fund management house, that you have an upper quartile level of skill (e.g. because of the way in which you select staff, develop research capabilities, etc.). If you then adopt the working assumption that all *other* managers will behave randomly (which, to first order, does not seem very unreasonable if you analyse many different peer groups), then to target an upper quartile level of skill you should be aiming to deliver approximately a 0.7 information ratio over one year (if both return and risk are annualised), a 0.4 information ratio over three years and a 0.3 information ratio over five years. These are close to the rule of thumb of an information ratio of 0.5 that is often used by consultants to define a 'good' manager.

8.3.5 Then, once you have defined an appropriate information ratio target, you can identify what level of risk needs to be taken to stand a reasonable chance of achieving the client's desired level of relative return, and, once you have defined an appropriate target risk level, you can use simulation techniques (or other approaches) to identify the sorts of portfolio construction parameters that might typically result in you running this level of risk over time.

8.3.6 This logic does, however, imply that, for the same fixed target outperformance level, a fund manager should alter his average position sizes as general levels of riskiness of securities change, even if he has not changed his intrinsic views on any of the securities in question. This is arguably an appropriate approach if a short-term change really is a harbinger of a longer-term structural shift in the market; but what if it just reflects a temporary market phenomenon? During the recent dot com boom and bust, average position sizes in many equity peer groups did not change much, which meant that their ex ante tracking errors typically rose (and then fell again). This suggests that, in practice, fund managers often use more pragmatic portfolio construction disciplines, e.g. applying maximum exposure limits to a single name (changing these limits only infrequently), viewing with some scepticism what might arise were risk budgeting theory to be rigorously applied.

8.4 *Fair Valuation Implications*

8.4.1 What does fair valuation have to add to this analysis? There seems to be an interesting dichotomy between *asset/liability management* in a banking context (and within some parts of the investment management community) versus the sorts of approaches typified by the above, that might more commonly be used by longer-term investing institutions.

8.4.2 Typically (within, say, an actuarial or pension fund context), asset/liability modelling involves *stochastic* modelling, i.e. the projection of assets and liabilities some way *into the future* under lots of different scenarios. Implicit in any such analysis is some probability distribution describing how the future might evolve. In contrast, in the banking/investment world, the focus is generally much more on the *here and now* using tools like VaR, even though banks and investment managers also hold long-term instruments.

8.4.3 Some of the dichotomy is more apparent than real. For example, to estimate the credit risk inherent in a complex derivative instrument, banks typically calculate metrics such as 'expected positive exposure', by projecting forward how much credit exposure the position might involve at future points in time, taking appropriate account of credit mitigation techniques such as collateralisation arrangements, and then averaging these exposures merely over the ones where the credit exposure is positive. So, stochastic simulations do exist in the banking world, particularly when analysing long-term instruments; it is just that they are less emphasised. And, insurers have to worry about resilience tests, stress tests, realistic reserving, ICAs and the like, several of which can be thought of as variants of VaR type approaches.

8.4.4 However, it seems to me that there may still be a philosophical difference here, maybe linked, in part, to differences in time horizons and/or to agency versus principal stances (see Section 1.3), which influences how one mentally recognises gains or losses over time.

8.4.5 The efficient frontier approach, described above (and indeed any forward looking stochastic modelling approach, if it involves differential return assumptions) in effect places some positive 'value' on the expected future outperformance of some asset categories over others. The very act of expressing the results in a risk/reward space involves assigning some positive benefit to strategies shown as having higher returns/rewards.

8.4.6 Fair valuation theory (and derivative pricing theory more generally) squares up the values of different instruments by reference to a 'risk neutral' probability distribution. For the hypothetical marginal market participant, the future returns on a security should balance the risks. Otherwise, the participant would not buy or sell at that price. In such a risk neutral world, the future returns on different asset categories are equal, i.e. the efficient frontier is no longer upwardly sloping, but flat. Using observed (i.e. *real world*) return assumptions gives insufficient weight to the greater disutility that investors typically place on downside outcomes, or, to be more precise, market prices imply greater weight on downside outcomes than is implied by most efficient frontier approaches.

8.4.7 What is probably needed is a greater emphasis on how the client's risk appetite differs from that of the relevant marginal market participant, who is, in effect, setting the fair price of the relevant asset or liability. Some of these differences can be accommodated via the use of different minimum risk portfolios for different clients, and so can still be fitted into the above framework. This is, perhaps, more applicable to bond land, e.g. some investors will prefer fixed interest and some will prefer inflation linked assets, because of the nature of their liabilities.

8.4.8 However, some of the differences in risk appetite are more complicated to handle, because they are linked to different appetites for downside risk for the same asset category. This is, perhaps, more applicable to equity land, given the lesser matching characteristics that are nowadays typically ascribed to equities. It is not enough to say: "I expect equities to outperform bonds by *x*% p.a. and for the risks of combinations to be described by some particular covariance matrix." Instead, I probably need a more sophisticated analysis that also includes: "... and of the *x*% p.a. expected outperformance, I can expect risk adjusted to benefit from *y*% of it, because of differences between my risk tolerance to equity downside risk and that of the marginal investor in this type of asset."

9. Credit Risk and Collateralised Debt Obligations

9.1 *Is a Distinction between Market and Credit Sustainable in the Context of the Trend towards Fair Valuation?*

9.1.1 The framework set out in earlier sections of this paper might be classified as a 'market risk' orientated approach. Market risk is often, in this context, differentiated from *credit risk*, see Section 3, even though we noted in Section 3.2 that the boundary is blurred.

9.1.2 It is easy to see how such a distinction grew up, particularly during an era when, in the U.S.A., the Glass-Stegal Act largely prohibited the same company from undertaking both types of activity. Whether it remains applicable nowadays is less clear, given the growth of large multinational integrated banking houses. It also seems to me that such a distinction will struggle in the light of fair valuation.

9.1.3 Of course, this is not to say that, practically speaking, one cannot differentiate between the two sorts of risk, as defined by the FSA. In particular, individual credit risk exposures usually involve highly asymmetric, i.e. skewed, pay-offs, whereas many sorts of market risk are often more symmetric (although not all, e.g. those encapsulated in many sorts of options).

9.1.4 Another apparent differentiator would seem to be the existence of *credit ratings*. Credit rating agencies, such as S&P, Moody's and Fitch, assign ratings ranging from AAA (least risk) to BBB to C to D, etc. to individual instruments. These provide an external guide to investors as to how likely the instrument is to default. If bonds are downgraded, then, typically, their *spread* versus, say, government debt (i.e. the difference between the redemption yields payable on otherwise similar payment streams) widens.

9.1.5 The existence of external credit ratings allows one to think of a bond as potentially migrating between rating buckets over time. Based on actual histories of defaults, it then becomes possible to model what losses ought to arise (and when they might occur) on bonds of any given rating, see e.g. Schönbucher (2003). The theory is much like that used by actuaries to model state dependent behaviour in general insurance, and so is sometimes referred to as the *actuarial approach to credit ratings*. As in the general insurance applications, there are some implicit assumptions being made. For example, it is typically assumed that, say, an A rating in 1993 means the same as an A rating in 2003, in terms of underlying company strength.

9.1.6 And within the investment management world, 'market risk' and 'credit risk' are often considered separately, because asset managers are often split into equity and bond teams, and 'credit' is then seen as part of bond-land. Risk management tools (because they are being sold to different teams) often differentiate between the two types of risk, with risk models designed for credit portfolios incorporating credit spread widening factors and credit rating bucket factors that are not considered relevant for other sorts of portfolios.

9.1.7 But, as elsewhere, fair valuation is a great leveller. A bank's trading book will, in general, have credit exposures to the issuers of any securities which it holds. The bank is just as exposed to the risk of default via these securities as it is from any loans which it has made within its banking book to the same entity (if the loan and the security rank *pari passu* in the event of default), and, in a fair valuation world, the value of any loan which the bank holds in its banking book should (like the securities which it holds in its trading book) be marked-to-market. So, its rise and fall in value (including its fall in the event of default) is a 'market risk'.

9.1.8 The distinction also, in my opinion, becomes untenable from a theoretical perspective, given the development of *collateralised debt obligations* (CDOs) and analogues.

9.2 *Collateralised Debt Obligations*

9.2.1 Traditionally, a CDO involved the establishment of a Special Purpose Vehicle (SPV) that held one set of debt instruments and funded these positions by itself, issuing several different tranches of debt, see Figure 10. The different tranches would have different priority levels and therefore command different credit ratings and credit spreads.

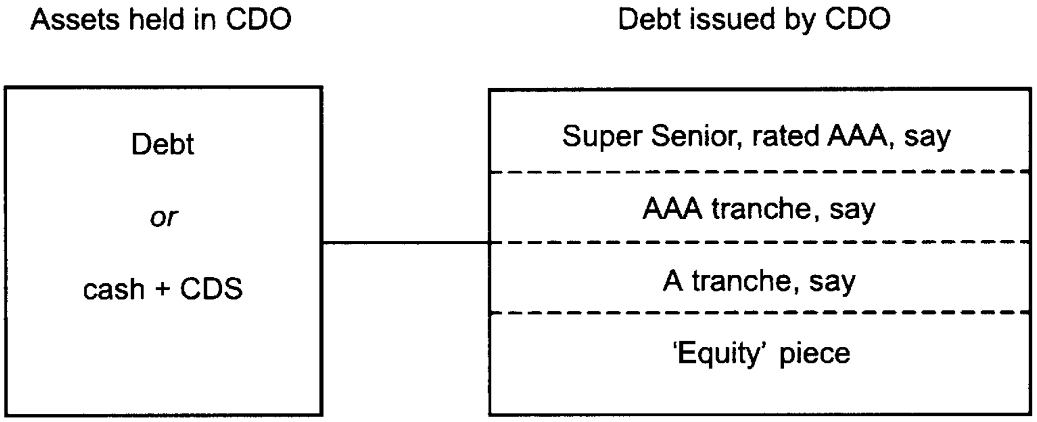


Figure 10. Schematic diagram of a traditional CDO

9.2.2 A bank could, for example, have a portfolio *P* of debt or loans that it wanted to remove from its balance sheet. It could do so by creating an SPV and selling *P* to this SPV. The SPV needs to raise sufficient funds to be able to purchase *P* from the bank. So, the SPV would have its own capital structure, issuing various tranches of debt (and at the bottom of the priority ladder an '*equity*' element). Different entities would subscribe to the different tranches of the SPV's debt, the spreads being demanded being dependent on where in the priority ladder the relevant paper lay. The 'technology' underlying CDOs is also known as *tranching*, as it involves a rearrangement of who suffers what if there are credit losses within a portfolio.

9.2.3 To understand better the impact of tranching, consider the following example. Suppose that the underlying portfolio contained ten debt securities (equally weighted). If one of them defaulted with zero recovery value, then the portfolio value would fall from, say, 100 to 90. This loss would be borne first by holders of the CDO's 'equity', i.e. the lowest priority tranche of the SPV's own balance sheet structure. If the equity tranche was not sufficiently large to absorb the loss, then other tranches sequentially higher up the priority ladder would suffer a loss. Holders of super-senior debt, i.e. the tranche at the top of the priority ladder, would typically only suffer a loss in the highly unlikely situation of there being multiple defaults in the asset portfolio. Actually, what is relevant is not the default frequency *per se*, but the degree to which the observed recovery default frequency exceeds that implied by the credit spreads ruling on the bonds held within the CDO. All other things being equal, this spread accrues to the CDO in the absence of defaults.

9.2.4 Each tranche is defined by its attachment and detachment points. The *attachment point* is the level of loss which, if not reached, results in that tranche being repaid in full at maturity. The *detachment point* is the level of loss which, if exceeded, means that the tranche holders receive nothing at maturity. Figure 11 shows how, in broad terms, the maturity proceeds provided by a specific tranche might differ from those arising from the portfolio as a whole, depending on the default experience of the CDO.

9.2.5 The underlying economic rationale for CDOs (and tranching more generally) is that different market participants may find different parts of the credit risk spectrum particularly relevant to their own needs. For example, different investors will have different risk profiles, perhaps because of regulatory considerations. By repackaging risks, so that each tranche can be sold to the sort of investor to which it is most suited, the theory is that the sum of the parts can, in some sense, become worth more than the whole. This is also the ultimate economic rationale behind the development of other risk transference or risk sharing mechanisms, such as the derivatives market (or, one might argue, the insurance market).

9.2.6 Originally, the debt instruments held within CDOs were typically passively managed or subject to very limited substitution rights, i.e. defined rules for replacing, say, a bond that had defaulted with another non-defaulted bond, to avoid the CDO having defaulted paper on its books. More recently, it has become more common for CDOs to be actively managed. Good active management benefits the investors in the CDO (just as it benefits investors in any other sort of actively managed investment product). The primary beneficiaries are the equity tranche holders, because they are then more likely to be repaid in full, or even to receive repayment above par; but good security selection can also result in the more highly rated tranches being upgraded, and hence revalued upwards.

9.2.7 Traditional tranched CDO structures suffer from the significant disadvantage that the SPV needs to sell all of its tranches to raise the funds it

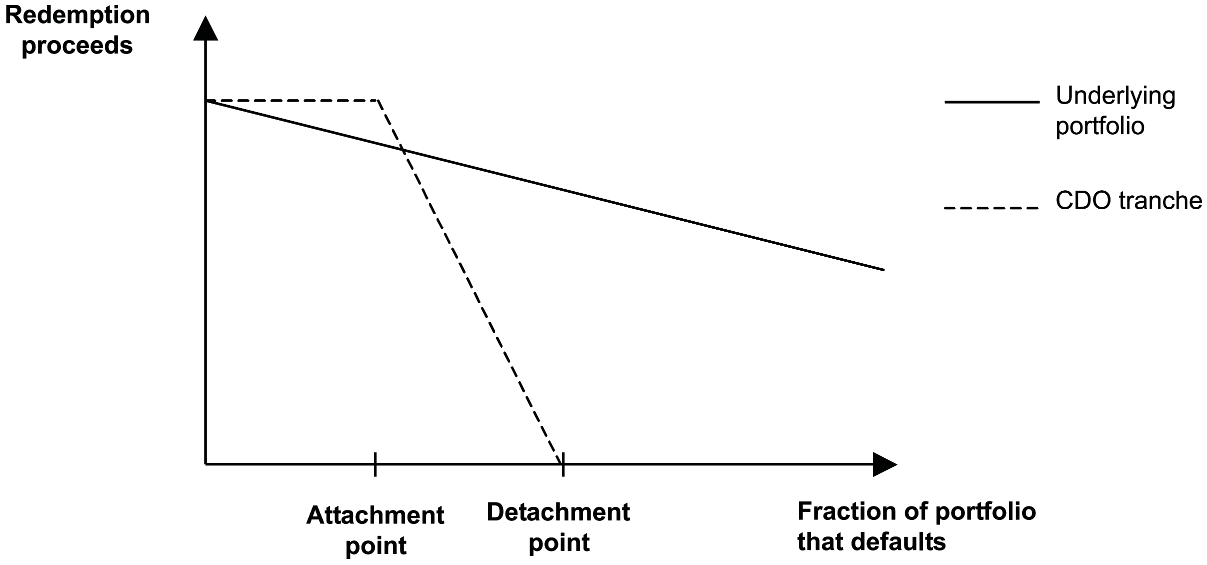


Figure 11. Redemption proceeds for a particular CDO tranche

needs to buy its debt portfolio. A particular issue is the equity tranche. It is typically unrated, making it difficult to sell to third parties, but it is precisely the piece of the capital structure that the creator of the CDO usually most wants to pass to someone else.

9.2.8 To circumvent this difficulty, investment banks have developed the concept of the *single tranche* CDO. In this structure, an investment bank synthetically acquires all bar a given tranche (say what would have been the A rated tranche), by selling to (and/or buying from) the SPV some credit protection that replicates what would have happened had there been the remaining tranches and these had been sold to third parties, see Figure 12. These transactions can be thought of as specific examples of *basket credit default swaps* (i.e. credit derivatives dependent on a whole basket of credit names), rather than the more standardised *single-name credit default swaps* (that depend merely on the behaviour of a single credit).

9.2.9 The investment bank will want to hedge the risks which it incurs by entering into these *tranche CDSs*. A good way for it to hedge at least some of these risks is for it to buy single name CDS protection on each of the individual credit risk exposures contained within the underlying portfolio. Typically, these sorts of hedges would reside in some notional hedge portfolio that the investment bank owns. Single tranche CDOs are nowadays, typically, actively managed, so that the investment bank will, ideally, want to be able efficiently to modify its hedge portfolio whenever the investment manager makes a change to the underlying portfolio.

9.2.10 A single tranche CDO is, therefore, typically structured so that its credit exposures are implemented using credit default swaps rather than physical bonds, and so is often called a *synthetic CDO* (but see below for an alternative meaning that might be ascribed to this term). This makes it easier for the investment bank to alter its hedges whenever the investment manager wants to alter the underlying exposures. The fund manager adds an extra element on behalf of the investment bank to any transaction that it

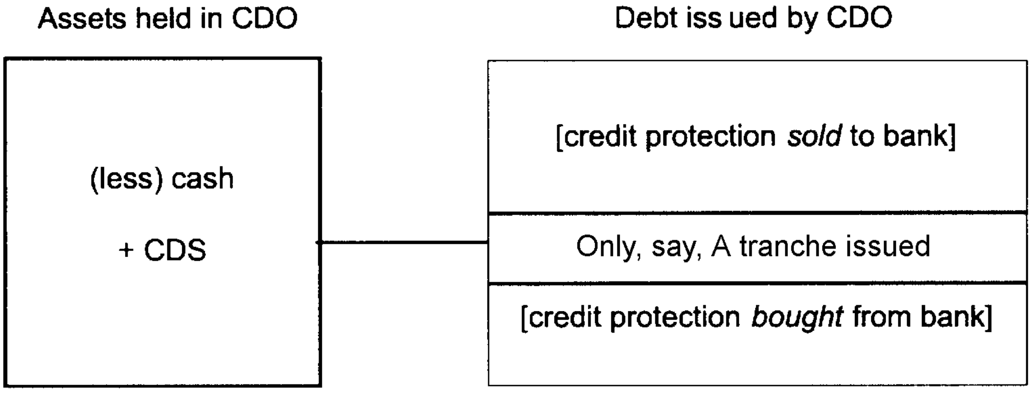

Assets held in CDO
Debt iss ued by CDO
(less) cash
+ CDS
[credit protection *sold* to bank]
Only, say, A tranche issued
[credit protection *bought* from bank]

Figure 12. Schematic diagram of a single tranche CDO

wants to undertake, which implements a corresponding change to the investment bank's hedge portfolio. The CDO is not required to carry out its trades through the relevant investment bank, but it does need to alter the part of the hedge portfolio that it controls according to some suitable pre-agreed mathematical algorithm whenever it changes its own underlying portfolio.

9.2.11 There is a further subtlety, at least with more modern single tranche CDOs. Suppose that a manger wishes to alter the underlying reference portfolio in some way. Then, all other things being equal, this will alter the value of the tranche CDSs. Originally, there were restrictions on what the manager could do to stop the manager changing the value too much to the detriment of the issuing bank (i.e. in this case, the other side of the tranche CDS transactions). Nowadays, what happens instead is that the attachment and detachment points are altered whenever a trade occurs, in a manner that leaves the replacement tranche CDSs worth the same as the value of the tranche CDSs immediately prior to the trade.

9.2.12 Investors, typically, rely heavily on the rating assigned to each tranche by a ratings agency when assessing its attractiveness. The ratings agencies use Monte Carlo simulation and other techniques to identify how likely they think a given tranche is to suffer a default (and its likely recovery rate). Usually, attachment and detachment points are set so that the tranche in question achieves a certain rating when issued, which the manager will typically wish to protect quite vigorously. Paper actually issued often does not use the minimum possible subordination level, to provide some protection against a downgrade in somewhat adverse circumstances.

9.2.13 The detailed methodologies used by different credit rating agencies to rate CDO paper varies. For example, one agency apparently concentrates just on the subordination level (i.e. where the attachment point is placed), whilst another one apparently takes into account the expected loss if the attachment point is reached (which also then depends on where the detachment point is placed). A rather more important point is that the rating that a given agency assigns to a traditional bond may not necessarily mean the same as an apparently identical rating that it awards to a CDO tranche; or rather, it may do as far as the formal meaning assigned to a rating by the rating agency is concerned, but it is not necessarily then correct to assume that, for the purpose the investor then uses the rating, it has the same meaning, see ¶9.5.9.

9.2.14 The use of CDSs to 'manufacture' single tranche CDOs highlights a close linkage between these two relatively recent financial innovations. Indeed, it is possible to avoid having a SPV entirely, i.e. to have a *totally synthetic CDO*. This would involve the owner retaining the debt portfolio on its balance sheet, rather than transferring it into a SPV, and for the owner then to purchase suitable basket CDS that provide the same risk transference as a SPV structure would have done. Again, the debt portfolio could be actively managed, and again it may be beneficial to devise mechanisms that allow the provider of the basket CDS to hedge the basket CDS efficiently. This would again imply a preference (if practical) for exposures to be traded in a hedge friendly way, such as via single name CDSs.

9.3 *The Leverage often Present in CDO Tranches*

9.3.1 An important element in understanding CDOs is to appreciate their typically *leveraged* nature. There are several angles to this. One can tell that there must be some possibility of leverage by noting that if, say, a tranche is £100m in size and has an attachment point at, say, 6% and a detachment point at, say, 10%, then this £100m must, in some sense, have some underlying 'reference portfolio' of £2.5bn, i.e. £100/(0.10 – 0.06)m. However, it is difficult, using just this sort of basic analysis, to work out the actual characteristics of any tranche, given the subordination protection that a tranche typically benefits from.

9.3.2 Somewhat more helpful in this context is to refer to Figure 11. We see that, if spreads widen, i.e. market implied default rates rise, then the market implied likelihood of total or substantial loss from holding at least some CDO tranches is likely to increase by significantly more than the spread increase on the CDO's underlying holdings. Thus, all other things being equal, CDO paper may experience a magnified mark-to-market spread movement. The magnification ratio, called the *tranche delta*, is sensitive to a number of parameters, but JP Morgan estimates that the weighted average delta across all CDOs issued in 2004 was roughly as per Table 6. We note that, with a traditional CDO structure as per ¶9.2.1, the combination of all of the individual tranches has the same redemption characteristics as the overall underlying portfolio. This explains why there are some tranches which have delta of less than one, although Table 6 shows that these are not the ones most typically purchased by investors in single tranche CDOs.

9.3.3 Of course, all other things are not always equal. In particular, CDO paper is itself subject to supply and demand considerations. The observed volatility in the price of such paper may not, therefore, exhibit the same degree of magnification (or dampening in the case of super senior) as implied above, at least if volumes traded are limited.

Table 6. Approximate average delta of different CDO tranches

	% of issuance	Average delta
Junior (subordination of 0% to 3%)	11.6	14.3
Mezzanine (subordination of 3% to 7%)	33.5	9.6
Senior (subordination of 7% to 10%)	24.2	4.3
Super senior (subordination of >10%)	30.7	0.4
Total/average	100.0	6.0

Source: JP Morgan

9.4 *Collateralised Obligations involving Debt, Loan, CDO, Equity, ...*

9.4.1 One reason why I think that tranching ultimately renders the distinction between 'credit' and 'market' risk, as defined by the FSA, somewhat spurious is that CDOs are not the only type of collateralised structures in existence. SPVs exist in which other sorts of credit assets substitute for the debt held within a traditional CDO, e.g. loans (in which case the SPV is called a collateralised loan obligation vehicle) or even other CDOs (in which case the SPV is called a CDO squared vehicle).

9.4.2 However, more interesting still, in this context, is that SPVs exist where the substitute assets are not what anyone would typically associate with 'credit', e.g. they can involve equities or hedge funds. How do the resulting structures then differ economically from equity investment trusts and other similar closed end vehicles, such as Real Estate Investment Trusts (REITs), and where then is the boundary between 'market' and 'credit' risk?

9.4.3 Can we fall back on a practical definition? External investors in CDOs paper typically rely heavily on the ratings assigned to the different tranche by an external ratings agency, such as Moodys or S&P. So, maybe we can define 'credit risk' as the sort of thing that credit rating agencies look at. The problem with this heuristic definition is that the ratings agencies are commercial organisations that will pursue revenue opportunities. It has recently been reported in the press that Standard & Poors are going to rate U.K. pension funds. Given the mismatch versus the liabilities typically exhibited by these funds, much of the driver to these ratings will presumably be the (equity) market risks that they are exposed to. Also, the credit rating of a CDO tranche ought to depend on any duration (i.e. interest rate) mismatch risks present between the CDO's assets and its debt. Once again, this is more traditionally thought of as 'market risk'.

9.5 *Risk Capital Computations in the Presence of Tranching (and CDSs)*

9.5.1 Another reason why I think that tranching messes up any previously well defined boundaries between market and credit risk is that tranching technology also potentially fundamentally alters the way that you need to think about credit risk from a risk capital perspective. This is most easily seen by considering the impact that tranching can have on solvency risk capital computations.

9.5.2 Take, for example, the regulatory capital computation that became applicable to U.K. life insurers as at 1 January 2005. In general, there are three regulatory computations applicable to U.K. life insurers. Two derive from Pillar I requirements, namely the *realistic peak* and the *regulatory peak*, both of which are ultimately mandated by FSA rules (although it is possible to obtain waivers from certain of these requirements from the FSA in certain circumstances). The third computation derives from Pillar II requirements. It involves each insurer preparing its own *Individual Capital Assessment* (potentially supplemented, if the FSA thinks that it is too optimistic, by the

FSA issuing further *Individual Capital Guidance*). For convenience, we use ICA as shorthand for the combination of the Individual Capital Assessment and Individual Capital Guidance.

9.5.3 The regulatory peak has been around for some time and ultimately derives from E.U. insurance directives, that even the E.U. Commission accepts are ripe for revision. For it, the life insurer might typically:

(a) Calculate the spread of each bond over the gilt yield curve.

(b) Calculate, using the rating assigned to the bond in question, the spread corresponding to its expected default loss, based on historic default experience.

(c) Deem the spread differential between (a) and (b) as the illiquidity premium ascribable to the bond.

(d) Capitalise this illiquidity premium (if the liabilities are also illiquid, and therefore illiquidity in the assets is acceptable), and take credit for it as a reduction to its required capital base by increasing the yield at which the liabilities are discounted.

9.5.4 The regulatory peak computation has the perverse characteristic that moving away from gilts into less creditworthy debt typically reduces, rather than increases, capital requirements. Also, the computation seems to be typically carried out on a security-by-security basis, and so probably fails to take full account of the diversifying effects of holding portfolios of bonds rather than isolated ones. One has to be a little careful with this logic, as there is plenty of flexibility, in practice, afforded in exactly how the illiquidity premium is calculated. Even if it is exactly as described above, the computation in effect often assumes that each position is actually a diversified basket of that particular rating category, but, typically, one would expect a suitably retranched portfolio to have what is, in effect, an 'overall' rating that is better than the average of its parts because of these diversification effects. So, use of a CDO structure should logically permit some further release of capital for companies to which this computation basis applies.

9.5.5 The realistic peak computation was only introduced a small number of years ago, and is rather closer to what we might describe as 'underlying reality' (as one might hope given the use of the term 'realistic' in its name); but even it comes unstuck with CDOs.

9.5.6 Strictly speaking, the realistic peak *Risk Capital Margin* (RCM) only applies to larger with-profits companies (or smaller ones that have opted to adopt it), and then only for their with-profits business. The RCM for credit risk is, broadly speaking, calculated as follows (for each bond and then summed), where $D =$ duration, $s =$ spread (yield) over gilts and F is a factor depending on credit rating:

$$RCM = MV \times \begin{cases} F \times D \times \sqrt{s} & \text{if rated B — or better} \\ \max(F \times D \times \sqrt{s}, 5\%) & \text{if rated below B — or unrated.} \end{cases}$$

Credit rating	F
AAA	3.00
AA	5.25
A	6.75
BBB	9.25
BB	15.00
B or below	24.00

9.5.7 There are two ways in which this computation fails to reflect underlying reality. Firstly, relative to the differential historic default experience on differently rated paper and paper of different durations, this computation seems to favour better rated shorter duration paper at the expense of worse rated longer duration paper. Secondly, it, too, is done on a security-by-security basis, and therefore presumably does not fully reflect portfolio diversification. So, if one could hypothetically retranche a typical corporate bond portfolio in a manner that both assigned it a rating more in line with its inherent underlying expected default experience and that also better reflected any additional diversification characteristics, then it should again be possible to reduce the realistic peak RCM to closer to underlying reality.

9.5.8 Only the third computation basis (the ICA computation) is really likely to come close to 'underlying reality'. The ICA is designed to reflect what the insurer believes is the 'true' amount of risk capital that it needs as a business, based on some standardised 'ruin probability' (the FSA has asked to see a figure based on a 99.5% one-year confidence limit, designed broadly to equate to a BBB rating).

9.5.9 What is happening here? You actually have several somewhat disjoint viewpoints, all converging on the same question. The different viewpoints are:

(a) *The market*: it is using derivative pricing based techniques to work out what are by definition, fair values for the risk transference involved with CDOs.

(b) *The rating agencies*: they are adopting other techniques to come up with their view of the intrinsic creditworthiness of the relevant issuer/issue. The key point is that these do not necessarily map one-to-one onto what we might call the *market implied rating*, as derived from the observed credit spread, i.e. credit ratings provided by ratings agencies are not market consistent.

(c) *The regulator*: appears to favour a fair valuation, i.e. market consistent, framework, which is only consistent with (a) and not (b), but it has still introduced an RCM framework (and inherited a regulatory capital framework) that relies, in part, on market inconsistent data, such as the ratings assigned to instruments by ratings agencies.

(d) *Individual insurance companies*: they, not surprisingly, have been receptive to strategies that efficiently minimise their Pillar 1 capital, including CDO structures that, in effect, arbitrage between the other three points of view.

9.5.10 There are shades here of our earlier discussion of time series versus derivative pricing based risk modelling. It seems to me that, only if the regulator moves to a fully market consistent based RCM (i.e. one where observed credit spreads override those derived from credit ratings and historic default/recovery experience) will you eliminate the potential for inconsistencies.

9.6 *Wider Ramifications of Tranching*

9.6.1 In fact, though, even such a shift is not radical enough, as it still differentiates too much between market risk and credit risk. Tranching seems to me to have potential ramifications for how the financial services industry and its regulation might develop over the longer term, which I think will ultimately render superfluous this distinction; for, it seems to me that there is little obviously fundamentally different in an economic sense between a bank, insurance company or pension fund and a suitably defined CDO. Take, for example, an insurance company. We traditionally think of such a company as having policyholders who give it money to invest; but we could equally think of it as a structure that holds assets and has funded their purchase by issuing a (policyholder) debt tranche, in the form of life policies, and a residual equity tranche (held by its shareholders), maybe supplemented by other tranches relating to other debts issued by the insurer.

9.6.2 The 99.5% one-year ICA requirement, recently introduced for life insurers by the FSA, can, in such a representation, be thought of as identifying an attachment point deemed appropriate by the regulator for the policyholder tranche (and therefore a required minimum equity base). So, a *fully market consistent approach* to setting capital requirements would, in effect, seek to answer the following question (for some suitable value of x):

> "What capital does the company need (and in what form) to ensure that, if the company restructured itself into something akin to a CDO, the tranche relating to policyholder liabilities (or the equivalent for a non-insurer) would command a market spread (over the appropriate risk free rate) of less than 0.5% p.a.?"

Such an approach does not differentiate between market and credit risk (or, indeed, any other type of risk).

9.6.3 Obviously we are not there yet, but perhaps, in time, market discipline (as per Pillar III of Basel II) will become the primary capital discipline imposed on financial services entities via such a computation. There may also be pressure on companies to release information about their ICA, so that others can attempt to answer the same sort of question.

9.6.4 Barriers between different types of financial services may also become increasingly blurred or untenable if the different sorts of risks become more easily tradable, and therefore held by different sorts of entities. Why should a conglomerate need an insurance company if it wishes to market one sort of unitised fund (a unit-linked life product), an asset manager if it wants to market another (an OEIC), and a bank if it wants to market a third (a market index-linked term deposit)? Why should industrial companies be able to provide entitlements with annuity like characteristics (pension benefits) through a trust subject to one sort of regulatory framework (with one set of concomitant credit risk exposures to the company in question), when insurance companies providing similar sorts of entitlements are subject to a different sort of regulatory framework (generating a different set of credit exposures)?

10. Liquidity Risk

10.1 *Managing Market Exposures versus Managing Short-Term Liquidity*

10.1.1 Active investment management is about achieving good returns at an acceptable level of risk. Up to now, in this paper we have been thinking of this in terms of taking market positions (and/or credit positions, to the extent that it is appropriate to distinguish between the two). This is not the whole picture. Managing market exposures involves deciding when to trade and in what. You also have to settle the trades into which you enter. This is known as managing *funding* or *short-term liquidity*.

10.1.2 The main challenge with managing funding, for a traditional long only asset manager, is to ensure that you have enough cash available to settle purchases as and when settlement of them falls due. Different instruments have different settlement cycles in this respect, e.g. some instruments are same day settlement ('T + 0'), some are settled the day after the transaction ('T + 1′), etc. You also need to be careful about what 'settling on a particular day' means. Is it by 5 pm local time (and, if so, what time zone is 'local'), 12 noon U.K. time, or some other time? Sometimes, you need to settle in advance ('T − 1' or before), e.g. if you are buying a fund, and the fund provider requires you to provide cleared funds before the actual pricing point of the fund.

10.1.3 Typically, you might assume that any sales that come due for settlement before purchases will generate cash, but this assumes that your counterparties will settle, on time, their purchases of securities from you. This does not always happen, although you may agree a 'contractual settlement' with your brokers/custodians, which, broadly speaking, puts you in the same position as if they had settled on time.

10.1.4 Settlement processes can seem relatively arcane to those not intimately involved with day-to-day market activities; but, given the sizes of flows through the marketplace, it is vitally important that settlement processes are orderly and involve minimum risk. Much effort has been expended to automate settlement activities and to develop approaches that minimise settlement failure risk. For example, typically security transactions nowadays occur via a process called true 'delivery versus payment', which means that, if you are buying securities you only release cash to buy the securities contemporaneously with the transfer of securities to you. It is worth noting that this does not completely eliminate market risk. Suppose that you agree to buy some securities for £1m with a T + 3 settlement, and two days after you trade (i.e. before you were due to settle the trade) your counterparty defaults. In a true DVP market, you should not ever part with your £1m, but, if the value of the security has risen to £1.1m, then it will now cost you £0.1m more than you had previously expected to enter into your desired transaction, i.e. you had some contingent credit exposure to your counterparty, contingent on the market moving against the position which you were attempting to enter into with that counterparty.

10.1.5 Management of funding in this context requires a cash 'ladder' that indicates when cash will become available or will be needed to settle transactions. One might think that the easiest way of minimising the risk of having insufficient liquid funds to settle transactions would be always to hold a large cash buffer, but then you are exposed to credit risk in terms of where you place the cash, and you may not be able to create such a buffer, if, for example, the fund is required to be almost fully invested. You could, if your client agreement permits, create liquidity by borrowing against the assets in the portfolio (typically, nowadays, in a collateralised fashion via *repo* or *stocklending* transactions). If the portfolio can go 'short', you want 'liquidity' of the opposite sort, i.e. to deliver to you the right sort of stock when you want to close your position (accessing the stock to enter the initial short transaction can then be achieved by *reverse repo* or *stockborrowing*).

10.1.6 Credit risk measurement and management techniques are relevant to these activities. For example, when you lend securities to a counterparty, they typically post *collateral* to you. If your counterparty defaults, then you will suffer a loss if the collateral (also referred to as *margin*, in line with derivatives nomenclature) which you hold is insufficient to purchase back the stock you no longer have. Under Basel II rules, banks typically need to assume that it might take them ten working days to do this, during which time the value of the collateral and the stock lent out may have diverged. One might, for example, compute a *probability of loss on default*, i.e. the likelihood that collateral will prove inadequate on default (assessed using, say, some suitable VaR style risk model), the *expected loss on default* (assessed, say, using option pricing theory) and hence an *annualised risk premium* (being the product of the previous two numbers, expressed in some suitable units). Dealing costs might be incorporated in the computation, taking into account the sizes of the positions involved, if these were expected to have a material impact on the option adjusted expected loss on default. In theory, a full derivative pricing based approach would jointly price both the expected loss on default and the probability of loss on default using a risk neutral probability of default, derived as if the combination were a single derivative instrument.

10.2 *Management of Longer-Term Liquidity*

10.2.1 Ensuring that you have access to longer-term liquidity may also be an issue. Corporates need to ensure that they have sufficient working capital. They may arrange *credit lines*, i.e. *credit facilities*, with their banks, that enable them to borrow from the bank on prearranged terms, even when the bank might otherwise be unenthusiastic about making them such a loan. Undrawn credit lines create contingent liabilities for the bank that can, again, be priced (and valued) using the above sorts of techniques. The spread (versus, say, LIBOR) that the bank will receive on such a loan, if the line is drawn, can be compared with the expected default rate that the loan will exhibit, and the probability of the line being drawn can also be derived, probably assuming that the corporate acts rationally when deciding whether to draw down the line.

10.2.2 Longer-term investors also, in principle, have similar issues, which can, in principle, be priced and valued in a similar manner. Take, for example, a pension fund. It needs to have sufficient cash available to pay its liabilities as they fall due. However, it might have a high proportion of its assets invested in equities or other assets with a low running yield, low enough not to provide the level of benefit outgo projected for the coming year or two.

10.2.3 Some pension schemes, in these circumstances, have been known to set up a cash flow matching portfolio (using gilts) that generates sufficient cash flow to ensure that the pension fund is almost sure of being able to meet its expected benefit outgo over some suitable number of years into the future. The idea would be to replenish this portfolio from time to time, regularly extending out the period over which the extra income requirement was needed.

10.2.4 Whilst such a strategy may incidentally have been a sensible investment call (if it involved selling equities and buying bonds at opportune times) and relatively straightforward to explain to clients, it is less clear to me that it is theoretically sound, purely from a liquidity management perspective. It involves regular sale of non-gilt assets to park them in a gilt portfolio to provide guaranteed ‘liquidity’ (in this case maturity proceeds from the gilts) sufficient to meet the required liability outgo shortfall. Schemes could, instead, ensure that they had sufficient liquidity to meet such shortfalls by selling the assets at the time when the liability needed to be paid (rather than in advance), or by negotiating a credit facility that enabled them to borrow against their assets in such circumstances, if they were worried that the assets might be temporarily depressed in value.

10.2.5 Of course, there are all sorts of possible risks that might stymie such alternatives, e.g. the provider of the credit line might have defaulted by the time that the fund wanted to draw on it, but, leaving aside the asset allocation element of the decision, it could be argued that using gilts in this sort of fashion is a rather belt and braces approach, which reduces the risk of there being a liquidity problem to a disproportionately low level. If everyone followed such a low risk strategy, then there would be a lot of funds holding a lot of gilts, thereby bidding up the price of gilts and making the approach relatively expensive.

10.2.6 One can argue that there are similar such systemic features within the financial system, more generally, that do, indeed, tend to bid up the price of gilts in this sort of fashion. Many OTC derivative transactions are now collateralised, because it keeps down both sides' credit risk. Acceptable sorts of collateral are most normally specified as cash or high quality government debt (e.g. U.S. treasuries, Euro government debt or U.K. gilts). If cash is used, some suitable interest rate will be payable by the holder of the collateral back to the provider, which can sometimes prove onerous to achieve, so a surer way of not being out of pocket via the collateral is to stick to suitably secure government debt. There is probably a virtuous circle here, with highly liquid government debt being particularly attractive for such uses, making it even more in demand and therefore usually even more liquid (ultimately to the benefit of tax payers, as it reduces the cost to the Government of funding its debt).

10.3 *The Risk Free Rate*

10.3.1 This has implications, in a fair valuation world, for certain types of assets and liabilities. For example, suppose that we have an annuity book and suppose, also, that we know the right (risk adjusted) mortality rate and expense costs to use in the valuation. The fair value of these liabilities that we ought to use in capital adequacy computations depends on the 'risk free' rate(s) at which we discount the liabilities. Even small differences in how we define 'risk free' will mount up if the liabilities are long term, as many annuity books are.

10.3.2 Sheldon & Smith (2004) assert that swap rates are the wrong rates to use to define 'risk free', because of the risk of default on the cash held to generate the floating rate payments, and go on to conclude that the right rate is the gilt rate. Section 5.1.3 of the current actuarial guidance note on 'Determining the With-Profits Insurance Capital Component' (GN45, 2004), is less categorical, suggesting that some rate between the gilt rate and the swap rate might be applicable.

10.3.3 The following analysis may, perhaps, shine some further light on this issue. Suppose that I buy a zero coupon bond paying 100 in one-year's time, issued by entity A. Suppose that I simultaneously buy protection on entity A, via, say, a credit default swap (CDS), which I enter into with entity B, with a likelihood and incidence of potential default that is not well correlated with that for entity A. To keep life simple, suppose that the CDS involves entity B, paying me 100 in one-year's time (less any recovery, then from a non-zero value to the zero coupon bond), if, and only if, entity A defaults within one year. In return, of course, I will need to pay a premium, which, again to keep life simple, we will assume is paid up-front in one lump sum payment. To receive anything other than 100 in one-year's time, both A and B need to default during the coming year.

10.3.4 There is, of course, some possibility that both A and B will default, but suppose that I enter into a further 'two name last-to-default' CDS on A and B both defaulting, this time with another independent counterparty C; and so on with counterparty D, E, ... Eventually, I should be able to make the risk of not receiving 100 in one year vanishingly small.

10.3.5 Although this logic might seem to be rather contrived, in theory, such a structure is becoming more practical to access as the credit derivatives market develops. It is widely accepted within the credit derivatives market that the theoretical pricing of CDSs is driven off asset swap rates (or, to be more precise, general collateral repo rates to the extent that these are observable), see e.g. Schönbucher (2003). This seems to me to favour using these rates as the starting point for working out the 'risk free' rate rather than gilts.

10.3.6 What is the right liquidity risk component to incorporate in the 'risk free' rate used to value liabilities? It seems to me that this depends on the liability. For some life insurance policies, the 'correct' amount of liquidity to assume is probably one consistent with the hypothetical credit hedged matched portfolio, as described above. For example, suppose that I had a pure unit-linked contract linked to a portfolio invested exactly in line with the matched portfolio, with policyholders suffering a suitable surrender penalty if they wished to break the contract early. Theory would suggest that the correct value to place on the liabilities is the same as the value placed on the assets (with gilt rates having no relevance here), as long as the potential liquidity implications of the underlying portfolio were somehow appropriately explained to policyholders.

10.3.7 However, perhaps the caveat in the preceding sentence is important. How many with-profits policyholders expect to bear the risks arising from the illiquid nature of any assets being held to back their liabilities, or would even understand such a question if posed? How many life insurers alter their surrender terms to reflect fluctuations in the costs of notionally realising the assets backing these contracts (and/or the fluctuations in value of these assets due to factors driving liquidity), in a manner that might equate to the cost of the insurer arranging a credit facility that provides them with the required liquidity? I suspect not all. As the pricing of credit facilities indicates, the 'cost' of arranging liquidity is linked to the credit worthiness of the insurer itself. If it were perfectly credit worthy, then, presumably, it could always borrow whenever it needed funding, but few insurers are in such a fortunate position.

10.4 *Incorporating Liquidity Risk in Capital Computations*

10.4.1 There seems, at present, to be some debate as to whether liquidity risk should contribute to ICAs and other capital calculations, alongside market, credit, insurance and operational risks.

10.4.2 I would answer this question by going back to the proposed fully market consistent approach to setting capital requirements, set out in ¶9.6.2. If we restructured the company into a CDO like structure, would the presence of liquidity risk alter the attachment/detachment points and/or market spreads relating to the policyholder liabilities' 'tranche'. It seems to me that, in general, the tranche pricing would be sensitive to liquidity premia (and hence liquidity risk should be included in ICA and other capital computations). However, it may be that, if there were nearly perfect cash flow matching between the assets and the liabilities, and that if (in the case of an insurer), when policyholders lapsed early (assuming that they can do, which may not be the case with an annuity book), then their policy proceeds contained adjustments reflecting the potential illiquidity of the assets backing their policies, then the level of liquidity risk to take into account may be immaterial.

11. Insurance Risk

11.1 *The Relevance of Fair Values to Valuing Different Sorts of Insurance Liabilities*

11.1.1 Sheldon & Smith (2004) note that it is difficult to estimate a fair value for many sorts of insurance liabilities. Is fair valuation, therefore, unimportant for insurers? By no means! In Section 2, we noted that capital adequacy is intrinsically about working out what value the market would place on the assets and liabilities were they to be 'put up for sale'. So, the fact that the computation involves subjective elements does not make it unimportant, merely difficult.

11.1.2 Insurance liabilities are not unique in this respect, and, even if they were, derivatives markets have a history of innovation that may make fair valuation more practical, going forwards. Corporate loans are not normally freely tradable on any organised market (the direct lender may, for example, acquire inside information on the company in the process of making such a loan); but it is now possible to hedge the credit risk involved in such loans using CDSs (and their interest rate risk using interest rate swaps). Identifying objective valuations for corporate debt is also trickier than it looks, since there is no *organised* exchange on which these instruments trade. Liabilities may be securitised or entire books sold on to third parties (which, of course, does not necessarily mean that the prices at which such books trade are necessarily easy to reconcile with each other, in part because there may be other business characteristics or relationships being transferred as well in such transactions). Retail bank deposits are equally individualistic in nature. They also present some theoretical challenges from a fair valuation basis, see Section 2.3.

11.1.3 Indeed, any 'marking to model' involves some subjectivity. By definition, any liability (or asset) that you cannot hedge perfectly is in the same boat; if it was possible, then you would be able to mark it 'to market' instead of 'to model'. Marking 'to model' involves, in some sense, an identification of some suitable hedge portfolio that does consist of market traded assets or liabilities and can be used to replicate, with some reasonable degree of accuracy, the relevant liability in question. The more the liability deviates from the hedge portfolio (in some risk sense), the greater the potential inaccuracy involved, i.e. the greater is the element of subjectivity in deriving a fair value. Perhaps, therefore, the key perceived issues for insurance liabilities are:

(a) how far away from the liabilities in a hedging sense are any available market observables; and
(b) how concentrated is this mismatch risk to a small number of factors, limiting the diversification principles that might otherwise limit the practical impact of such divergences to the computation of the total fair value of the entire liability book?

11.1.4 For insurance liabilities containing options, how might I derive such a hedge portfolio? Again, the theory of derivative pricing is relevant, bearing in mind that practical hedging strategies may involve more risk than otherwise strictly necessary, so as to mitigate dealing costs, see Kemp (1997), or to maintain liquidity, see Section 10. The 'distance', in risk terms, between the liability and the (potentially dynamically adjusted) hedge portfolio can again be measured using VaR or tracking error type approaches, as described earlier in this paper.

11.1.5 So, it seems to me that there is no fundamental difference between insurance liabilities and any other sorts of liabilities in a fair valuation world. Even the time horizon is theoretically irrelevant, as we discovered in Section 7.

11.2 *The Long-Term-Ness of Certain Types of Life Insurance Contracts*

11.2.1 Where I think that, qualitatively, there may be differences is in some of the characteristics typically exhibited by insurance company contracts. For example, many life insurance contracts are small and quite long term in nature. This has some important implications for expense reserving. Typically, the company needs to reserve for potential expense overruns for the entire life of the contract. It must also reserve for such overruns for new policies which it might write over the coming year. However, if the company can unilaterally terminate the contract (or, perhaps, if it can unilaterally increase charges without limit and it would be treating policyholders fairly were it to do so) then, presumably, it only needs to reserve for a much shorter period (i.e. only until such time as it might be reasonable to assume that it has exercised these powers).

11.2.2 This contrasts with other types of savings contracts, e.g. OEICs or unit trusts, which can be thought of as short term in nature (albeit typically renewed), as these sorts of vehicles can normally, *in extremis*, be closed down or merged by the relevant provider, without the consent of the investor (albeit, probably, at some cost to the reputation of the provider).

11.2.3 Do customers actually value this long-term-ness? Take, for example, defined contribution pension scheme provision. In the U.K., this is typically provided via a life insurance route, in contrast to the U.S.A., where 401(k) schemes typically involve direct investment in mutual funds. Previously, many U.K. sponsors set up *occupational* DC schemes involving a trust arrangement, in which beneficiaries looked to the trust to provide their benefits, and the trust might itself enter into contracts with a suitable DC provider life insurer. However, sponsors are becoming keener to have nothing to do with any legal structure linked to pensions for which they themselves are responsible. They are, therefore, becoming keener on *contract* based DC schemes (whether in 'stakeholder' form or otherwise), in which the employees enter into a contract directly with the life insurer. From the insurer's perspective, these involve slow accumulation of funds. They carry a genuine risk that, if things do not work out, then the insurer may be saddled with a sub-scale book of business that it cannot offload.

11.2.4 However, do policyholders actually value a forced obligation on the insurer to administer the contract, come what may, for many years, even if the business is sub-scale? Would not a better structure allow the insurer (and equally the policyholder) to walk away from the arrangement in suitable circumstances, returning the investment to the policyholder and leaving it to him to find a suitable home for it? This would reduce the expense reserve that the insurer needs to carry, which should ultimately mean better value for money for the policyholder. Being left with a sub-scale business line also sounds like an invitation to deliver a less than market leading service proposition, which is ultimately not in the customers' best interests either.

11.2.5 In a similar vein, do typical annuity structures best meet the needs of customers, in the light of uncertainties concerning future mortality improvements? We noted, in Section 4.7, how annuity buy-out prices seem expensive to many pension funds. One issue is how unpalatable is the risk of potential further improvements in longevity to insurers, who may already have more of this sort of risk than they can easily cope with. Perhaps there is a way in which beneficiaries could be given the opportunity to buy annuitisation year by year, rather than in one fell swoop, for their entire remaining life, if it is felt that market implied annuitant longevity is too pessimistic. Most commentators who argue in favour of annuitisation of people's wealth when they are older do so because it reduces the risk of the individual running out of funds by living too long. Less focus is placed on the entire cohort living too long. Perhaps the only way practically of providing annuitisation to all who will eventually need it is to 'experience rate' the cohort, as a whole, in some way.

12. The Future?

12.1 *The Future for Risk Management*

12.1.1 It seems likely to me that, over time, even greater focus will be placed on 'portfolio' or financial risk measurement and management within the financial community. Fair valuation methodologies are an inherent underpin to this trend, since without them much of the mathematics behind risk measurement becomes unsound. As we saw in Section 9.6, fair valuation principles also have important messages to the risk management community, itself and so are likely, over time, to have a big influence on how the relevant calculations are carried out. We might, in this context, prefer to describe 'portfolio' risk as 'market' risk (since it involves exposures to things external to the company, i.e. to 'the market'), except that the term 'market risk' has already typically been applied merely to a sub-set of these risks.

12.1.2 From a regulatory perspective, there is, not surprisingly, an enthusiasm to attempt to apply the same sorts of mathematical disciplines to 'operational' risk management. I think that this may prove more difficult, given the fundamentally different nature of the risks involved. Of course, there still needs to be a close dialogue between the two, as portfolio/market risk and operational risk can, from time to time, transmute into each other.

12.2 *The Future for Actuaries and the Actuarial Profession*

12.2.1 The future should also be bright for those with market/portfolio risk management expertise (coupled, of course, with good communication skills). It would also be churlish of me not to promote a good combination of financial, mathematical and economic expertise as part of this skill-set, tempered with a healthy dose of pragmatism.

12.2.2 One might logically expect many actuaries to have (or to be able to acquire) the sorts of skills that a portfolio/market risk manager should ideally exhibit, but little of what I have covered in this paper is exclusively actuarial in nature (as astute readers will have noticed from the relatively few references to 'actuaries' or 'actuarial' elsewhere in the paper). Other professional groupings can develop, and are already developing, similar expertise outside the current actuarial professional framework. This presents some threats and opportunities for the actuarial profession — threats that others might encroach on what actuaries might have previously seen as their own preserve, but opportunities to expand into new fields and/or to expand the coverage of the profession to embrace these newer risk management professional groupings.

12.2.3 Astute readers will also have noted that there is relatively little that is specifically U.K.-centric in relation to the fundamental impact that fair valuation trends will have on risk management disciplines. The U.K. Actuarial Profession may therefore also need to consider exactly what role a specifically U.K. orientated professional body should have in a world where national boundaries may have increasingly little relevance to the answers to actuarial problems.

Acknowledgements

The author would like to thank two anonymous referees and Gareth Taylor and Paul Bourdon for their helpful comments on earlier drafts of this paper. He would also like to thank members of the U.K. Actuarial Profession's Liability Driven Benchmarks Working Party for commenting on Appendix A.

References

Abarbanel, H.D.I. (1993). The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, **65**, 4.

Abramowitz, M. & Stegun, I.A. (1970). *Handbook of mathematical functions*. Dover Publications Inc.

Billah, M.B., Hyndman, R.J. & Koehler, A.B. (2003). Empirical information criteria for time series forecasting model selection. Monash University, Australia, Department of Econometrics and Business Statistics, Working Paper 2/2003, ISSN 1440-771X.

Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, **81**, 637-654.

Booth, P.M. & Marcato, G. (2004). The measurement and modelling of commercial real estate performance. *British Actuarial Journal*, **10**, 5-61.

Chapman, R.J., Gordon, T.J. & Speed, C.A. (2002). Pensions, funding and risk. *British Actuarial Journal*, **7**, 605-662..

Clark, P.K., Hinton, P.H., Nicholson, E.J., Storey, L., Wells, G.G. & White, M.G. (2003). The implication for fair value accounting for general insurance companies. *British Actuarial Journal*, **8**, 1007-1044.

Cowling, C.A., Gordon, T.J. & Speed, C.A. (2004). Funding defined benefit pension schemes (to appear).

Financial News (2004). New focus on risk and return brings derivatives boom. *Financial News Briefing Notes*, March 2004.

GN45 (2004). Actuarial Guidance Note GN45: Determining the with-profits insurance capital component (version 1.1). Institute of Actuaries, 31 December 2004.

Heywood, G.C., Marsland, J.R. & Morrison, G.M. (2003). Practical risk management for equity portfolio managers. *British Actuarial Journal*, **8**, 1061-1123.

Hurst, M. (2004). Fair value's big time implications, *Investment & Pensions Europe Netherlands Supplement*, May 2004.

Investment Management Association (2004). Market timing: guidelines for managers of investment funds, Paper published by the Investment Management Association, October 2004, see www.investmentuk.org.

IMA & Data (2004). Pricing guidance for investment funds: fair value pricing, Paper published by the Investment Management Association and the Depositary and Trustee Association, September 2004, see www.investmentuk.org.

Kemp, M.H.D. (1997). Actuaries and derivatives. *British Actuarial Journal*, **3**, 51-162.

Kemp, M.H.D., Cumberworth, M., Gardner, D., Johnson, J. & Sandford, C. (2000). Portfolio risk measurement and reporting: an overview for pension funds. Institute of Actuaries, 2000.

Leippold, M. (2004). 'Do not rely on VaR' *Euromoney*, November 2004.

LIFFE (1992a). The reporting and performance measurement of financial futures and options in investment portfolios. *The London International Financial Futures and Options Exchange*.

LIFFE (1992b). Futures and options: standards for measuring their impact on investment portfolios. *The London International Financial Futures and Options Exchange*.

Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, **29**, 449-470.

Neuberger, A.J. (1990). Option pricing: a non-stochastic approach. London Business School Institute of Finance and Accounting, IFA Working Paper 183.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in C: The Art of Scientific Computing* (2nd edition). Cambridge University Press.

Schönbucher, P.J. (2003). *Credit derivatives pricing models*, Wiley Finance.

Sheldon, T.J. & Smith, A.D. (2004). Market consistent valuation of life assurance business. *British Actuarial Journal* (to appear).

Weigend, A.S. & Gershenfeld, N.A. (1993). Time series prediction: forecasting the future and understanding the past, *SFI Studies in the Sciences of Complexity, Proc Vol. XV*, Addison-Wesley.

Yiasoumi, C., Canham, D., Miller, J., Wharmby, N. (2004). The management of the discontinuance of large defined benefit schemes. Paper presented to the Staple Inn Actuarial Society, 16 November 2004.

APPENDIX A

LIABILITY DRIVEN INVESTMENT FOR DEFINED BENEFIT PENSION SCHEMES

A.1 *A Typical Structure (for a U.K. Defined Benefit Pension Scheme)*

A.1.1 There seems to be growing interest in the concept of *liability driven investment* for U.K. defined benefit pension schemes. Large mature schemes, with a greater bond focus, typically seem to be more interested in this type of investing than less mature, more equity, focused clients.

A.1.2 There are several different ways in which a liability driven investment portfolio might be structured. Perhaps the simplest involves two parts:

(a) *An underlying physical component*, typically consisting of an actively managed bond portfolio chosen, in broad terms, to look like the relevant liabilities. For example, if the liabilities are partly fixed in monetary terms and partly linked to movements in the Retail Price Index (RPI) (in other countries, the Consumer Price Index (CPI)), then it might incorporate some fixed-interest and some index-linked bonds.

(b) *A swaps overlay component*. This would typically consist of one or more *swap contracts* (or other similar derivatives), that involve the pension fund giving up one set of future cash flows (e.g. ones like those arising from the portfolio in (a)), and receiving, in return, another set of future cash flows (e.g. ones more closely matching the relevant liabilities). Precisely how these swaps might be structured can vary. For example, there might be one swap that pays away to the bank cash flow akin to that arising from the portfolio in (a), in return for interest payments on some notional principal linked to prevailing LIBOR cash rates. There might then be a second swap that paid away this LIBOR cash flow in return for a cash flow that more closely matched the pension fund's expected liability outgo; or there might be several swaps on each side that handled different parts of the cash flow (e.g. differentiating by term or by liability type); or all of the cash flows might be wrapped up in a single overarching swap.

A.1.3 The concept is similar to the actuarial theory of matching. Indeed, if the liabilities are short enough and the trustees want a passively managed low risk approach, then (b) might become superfluous and (a) might be merely involve a more traditional cash flow matched portfolio using, say, gilts.

A.1.4 The core 'new' idea is the use of swaps or other similar derivatives. They are used because the liabilities are, typically, of too long duration to be matched merely using physical bonds. So, you need a 'synthetic' method of artificially lengthening the duration of the assets if you do not want to be exposed to the risk that very-long-dated yields will fall more than you expect. Using swaps also gives you a wider range of underlying bonds in which you can invest.

A.1.5 If the liabilities are RPI linked (or contain inflation linked characteristics such as Limited Price Indexation (LPI)), then the same overall concept is still applicable. The only difference is that the cash flows that the swaps pay to the pension fund need to include these features, i.e. they need to involve the investment banks selling *inflation* to the pension fund. Of course, banks typically want to hedge their exposures. So, they will be on the lookout for other market participants (e.g. utility companies or PFI projects) prepared to sell them inflation. The two sides do not need to be in identical form (e.g. one might be strictly increase in line with the RPI, the other might be more LPI in nature). The 'art' of good derivatives intermediation is to be able to access both sides of the flow, to make a good return between the two and to keep the inevitable residual mismatches well controlled and hedged (and to charge an appropriate spread for carrying this risk).

A.2 *The (Typically Bond Based) Core Element of such a Structure*

A.2.1 An important advantage of the above structure is that it divorces the managing of the 'core' asset base from the 'bespoke-ness' needed to achieve a close match to the liabilities. The core can then be managed in a practical manner, e.g. along the lines of a manager's standardised investment process against some relatively standard benchmark, offering potential economies of scale.

A.2.2 The precise structure of the core element can still express trustee preferences, but these preferences can now primarily refer to the assets in isolation, rather having simultaneously also to cater for the precise shape of the liabilities. For example, the core element might eschew gilts in favour of a greater proportion of less well rated credits. This might be because the yield spread of such bonds over gilts is believed by the trustees to over-compensate the holder for the likely future default loss experience on such bonds, on the grounds of liquidity criteria, see Section 10. It can also incorporate a wider range of assets. There are relatively few long duration bonds in either the government debt or corporate bond markets.

A.2.3 It is not necessary for the core component to be exclusively bond orientated. It could involve *portable alpha*. Nowadays, swaps come in a very wide variety of forms. It is now possible to swap almost any sort of return stream, property-like, equity-like, bond-like, cash-like or inflation-like, into any other sort of return stream, embedding into the swap, if you so wished, caps, floors and other option-like characteristics. So, if you have confidence in a given active manager's skill at adding value, it can be in any asset class that you like, and you can still 'port' this added value onto a liability orientated benchmark merely by swapping the return on the relevant active manager's benchmark into the return on the benchmark which you set by reference to your liabilities.

A.2.4 However, whether such refinements are likely to be appreciated by most sets of trustees is less clear to me. A few asset managers do offer portable alpha products, but take-up to date has been relatively limited, perhaps because of the difficulties involved in educating trustees in the concepts involved (or in being sure that there is no leakage of value by the porting process). Also, one can argue that the swap contracts might be more keenly priced if they are swapping similar sorts of return streams. So, all other things being equal, if your desired cash flows are akin to fixed or inflation linked bonds (just rather longer than is easily available in the physical market place), then starting with similar sorts of cash flows may be preferable.

A.3 *The Swaps Element of such a Structure*

A.3.1 Divorcing the core physical portfolio from the derivatives overlay helps to clarify who is responsible for what decisions. The following parties are involved, and would typically have the following responsibilities:

(a) *Trustees* carry ultimate legal responsibility for the fund. They would be responsible for choosing who manages the core element and the swaps overlay. In the above structure, they would also be responsible for instructing the investment manager when to execute exactly what swap transaction (although, in practice, there would have been prior liaision with the investment manager in choosing how best to frame these instructions).

(b) *Scheme Actuary* would normally prepare any required liability cash flow projections, and update them, as necessary, at regular intervals. See below for what such projections might contain.

(c) *Investment consultant* would normally advise the trustees on overall investment strategy, on fund manager selection and on how to monitor the fund manager and measure the manager's performance. Together with the actuary, he would advise on exactly what liabilities to match (e.g. should it include pensions in payment, deferred pensions and/or actives' liabilities?).

(d) *Fund manager* is likely to be responsible for managing the underlying bond portfolio and for actual implementation of the swap transactions. The role in relation to the swaps overlay could, perhaps, best be classified as 'execution only' in the sense that the fund manager would probably help draft up any instructions formally given to it by the trustees and/or investment consultant, but otherwise the swap portfolio would be 'non-discretionary'. This would be in contrast to the core physical portfolio (which would, most typically, involve discretionary active management). The fund manager would most likely provide education to the trustees, views on transaction timing and valuations of the individual swaps. The fund manager would also most likely arrange for the collateralisation of the swap portfolios.

(e) *Investment bank* would be the trustees' actual swap counterparty, i.e. the entity whose balance sheet would honour the contractual obligations in any given swap transaction. In principle, trustees (or their consultants) could deal directly with such banks (subject to any overriding requirement on the trustees to avoid 'day-to-day' investment activity if they are not FSA regulated); but, in practice, banks' derivatives desks are remunerated on a transaction orientated basis. This is not obviously conducive to acting in the best interests of the trustees. It is most likely that the trustees would delegate choice of swap counterparty to their fund manager, who would make the choice by reference to the usual sorts of 'best execution' criteria that apply to fund manager dealing activity (subject to any overriding criteria set by the trustees, such as a credit rating requirement). There could be several such banks, as the fund manager, in principle, needs to apply best execution criteria each time new swap transactions take place.

A.3.2 In practice, there is likely to be close liaison between the actuary/ investment consultant and the fund manager when preparing suitable liability projections, and hence a proposed structure. The fund manager might also typically work with a few well-chosen investment banks, who can help to identify what derivatives are most likely to meet the client's requirements.

A.3.3 There needs to be such interaction, because overly exact cash flow matching might result in an overly complex (and therefore expensive) structure, bearing in mind the inherent approximations involved in liability projections (and the inherent approximations involved in modelling how the actively managed core portfolio might behave). There are also minimum amounts, below which it is impractical to effect swap contracts, which depend in part on how non-standard the swap is. An exact hedge of all of the risks embedded in the liabilities may be prohibitive or even impossible (e.g. liability driven 'investment' has rarely to date attempted to include scheme specific longevity protection). Experience suggests that complicated overlay structures may initially be discussed with trustees and their consultants, but, typically, only relatively simple structures seem to be used in practice.

A.3.4 At regular intervals (say yearly), the client (in conjunction with its actuary/investment consultant) would probably revise its cash flow projections and, after discussion with the fund manager, would instruct the fund manager to alter the structure of the swaps within the swap portfolio. Again, this would be done subject to the usual best execution rules, perhaps, if necessary, novating or cancelling previous swap transactions with new ones (to avoid building up large numbers of swap transactions that largely cancel each other out, and which might be burdensome to administer).

A.3.5 This flurry of activity contrasts with what happens the rest of the time. The fund manager does incur some ongoing costs, most notably the costs of sorting out the collateralisation of the swaps, as well as ongoing reporting/valuation. These costs are typically smaller than the costs of actively managing a portfolio, and might be absorbed within an all-in fee covering both arrangements. It would be possible for the fund manager of the swaps overlay to be different to the fund manager of the underlying physical bonds (just as a scheme's tactical asset allocation manager does not need to manage any of the underlying assets). However, this may make collateralisation procedures more complicated.

A.4 *Mitigating Credit Risk within Swap Contracts using Collateralisation*

A.4.1 Normally, the pension scheme would want the swap counterparty to *collateralise* the swap contract. The aim is to reduce the exposure that the pension fund has to the risk of default of the bank involved. The aim is to have moved some suitable form of collateral from the bank to the pension fund, whenever such a default might be costly to the pension fund. This involves marking to market the swap (by definition, this is the estimated cost of effecting a similar sort of swap with another counterparty), and whenever this builds up to be materially positive as far as the pension fund is concerned, for additional collateral to be 'posted' by the bank to the fund. If the mark to market then declines, some of the collateral would be released and returned back to the counterparty.

A.4.2 The counterparty might, of course, also require the swap to be collateralised for the same reason, but in reverse. Over the last few years, many life insurers entering into over-the-counter derivative transactions have discovered that they may be deemed less credit worthy than their counterparties. Underfunded pension funds may face the same learning curve!

A.4.3 For most transactions of any size, it is now common for collateral flows to occur quite frequently, even daily (although there will typically be minimum thresholds and a minimum build-up of exposure, typically dependent on credit rating, before any flow occurs). It may be possible to pledge securities held within the underlying portfolio, or, it may be necessary to hold some cash buffer within the swap portfolio itself to meet such calls. If, instead, the bank is posting collateral to the scheme then it too needs looking after, since it may need to be returned at some stage.

A.4.4 Typically, the asset manager would negotiate collateralisation arrangements on behalf of its client via a *Credit Support Annexe* within its wider negotiation of the master International Swap Dealers Association (ISDA) legal documentation governing the overall relationship between the client and its bank counterparty. Normally, the client would legally be one of the two parties to swap, with the asset manager merely acting as its agent. The pension fund might, therefore, want its own lawyers to review or negotiate these contracts, but, in practice, the investment manager is likely to have greater negotiating clout with the bank, given other relationships that it may have. The investment manager may, therefore, adopt umbrella documentation relating to all of its clients that wish to transact with the relevant counterparty. Where the client has multiple swap transactions with the same counterparty, it is normal to have them all netted off within the relevant ISDA and Credit Support Annexe. Otherwise, one party can find that, in the event of the other party defaulting, it owes money to the defaulted party on one transaction, but cannot recover what it is owed on another.

A.5 *Monitoring such a Structure*

A.5.1 There are three key elements to the above structure that might need monitoring:

(a) *The (actively managed) underlying bond portfolio*. This would be assessed as usual for the asset management product in question. For example, if it involved management of a credit portfolio against a market index, then performance and risk measurement and attribution analyses versus the benchmark in question might be reported as per the asset manager's/ pension fund's usual reporting cycle.

(b) *The (passive) swaps overlay*. This might, for simplicity, also be reported upon to a similar frequency, although most attention would be focused on those occasions when the swap positions needed to be altered.

(c) *The effectiveness of the choice of swaps overlay structure in relation to the scheme's liabilities*. Various approximations will have been interposed between the precise liability model available from the actuary and the precise structure of the swap portfolio. The swap portfolio being 'execution only' in nature, this element of the decision making is actually one that lies with the trustees, albeit only after taking advice from other parties.

A.5.2 The key additional requirement is to construct some sort of *liability benchmark* (or *index*) that reflects, in a market orientated way, the nature of the liabilities. Constructing such a benchmark may also directly guide the choice of swaps to hold within the overlay portfolio.

A.5.3 The most obvious way to proceed is first to develop some cash flow projections, differentiating between ones with different sorts of economic sensitivities (particularly those where the sensitivities have option-like characteristics, such as LPI). For example, the liability flows might be differentiated by year of projected payment into those that involve:

(a) *fixed monetary sums*, e.g. those arising from benefits not subject to any increases;

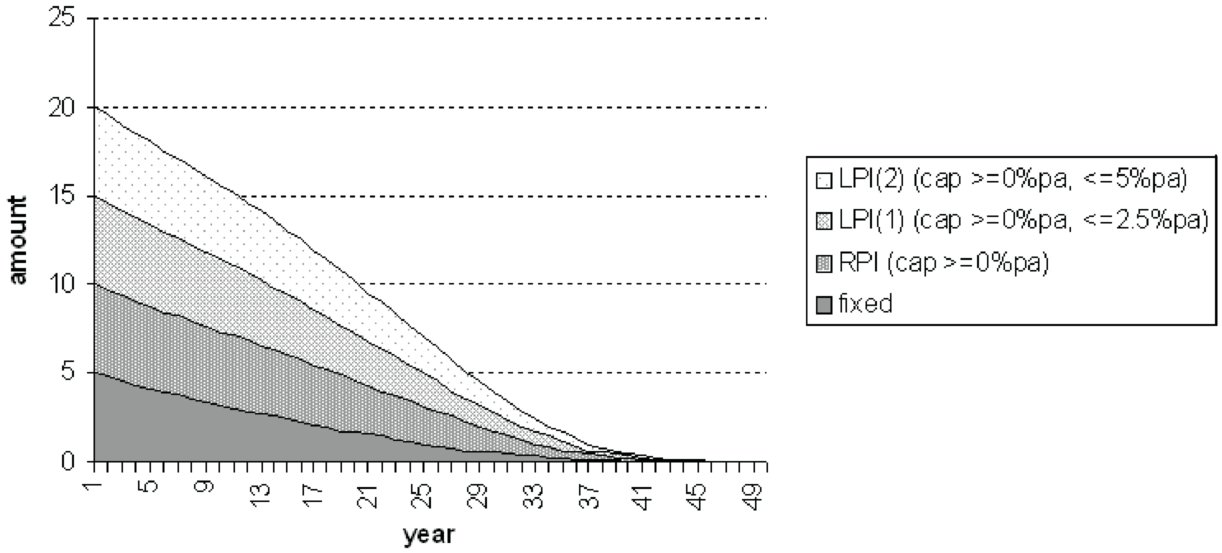
(b) *fully RPI inflation linked sums*, e.g. benefits subject to full RPI linked increases;

(c) *sums that increase on a year by year basis on some more complicated measure, driven by inflation at that time*, e.g. LPI type increases in payment. For these sorts of liabilities, the expected outgo during a given future year can still be derived from a single expected amount at outset, together with the history of RPI increases since then; if different ceilings, say 2.5% and 5% p.a. caps, apply, then these flows should, in principle, be differentiated, as swaps to match them exactly would also differ; and

(d) *cash flows governed by more complex increase formulae dependent on multi-year investment or economic conditions*. At least in principle, benefits linked to LPI in deferment fit into this category. The big difference between these sorts of cash flows and the sorts referred to in (b) or (c) are that they, in principle, require multi-dimensional matrices to specify as they depend jointly on date of withdrawal, assumed date of retirement, assumed date of payment and (for those already deferred pensions at outset) on how large RPI increases were prior to the start of the projection relative to the caps and floors present in individual member's benefits. As with (b) and (c), they also depend on RPI increases post the start date of the projection.

A.5.4 The choice of numeraire (e.g. whether the cash flows are in nominal or real terms, or if they are expressed using some present value metric) is not particularly important, as long as the cash flow analysis ultimately precisely specifies the assumed cash flows. For example, suppose that we have some nominal liabilities, some RPI linked liabilities (with a floor of 0% p.a. annual increase) and some LPI in payment liabilities, some with an annual cap of 2.5% and floor of 0%, and some with an annual cap of 5% and floor of 0%. The projected liabilities might then be expressed in present value terms (discounting, say, using a constant 4% p.a. discount factor) and using an assumed future inflation rate, say 3%, as per Figure 13. It is possible to work backwards from these projections to derive what the cash flows would be had any other future inflation assumption been used (and any other term dependent discount factor used, including one calibrated to match actual prevailing yield curves). In this illustrative example, we have assumed equal proportions at outset of each type of pension increase, with all scheme members assumed to be aged 60 and to have just retired (and with the somewhat unreasonable assumption that pensions are payable yearly in advance). The mortality assumed in this example is that underlying the PMA92 tables (with 28 years, of further mortality improvement incorporated). The average duration of the liabilities in this example is around 12.2 years in this instance, which would rise to 12.7 years if all of the liabilities were RPI linked.

A.5.5 One can now see why cash flows as per (d) are so problematic — they require lots more detail to specify precisely. It may be possible to develop suitable approximations that simplify them into a form that is more



Source: Threadneedle

Figure 13. Illustrative cash flow projection, all cash flows discounted to the present time using a discount factor of 4% p.a., inflation assumed to be 3% p.a. in the future

easily specifiable. It might also, in practice, be possible to simplify away liabilities of the form described in (c) above. It is also worth noting that the cash flows are not deterministic in nature. If the number of members involved is quite small, then the random incidence of individual deaths will introduce uncertainty. For more sizeable schemes, the unpredictable nature of future changes in general levels of longevity is likely to be more significant (as is whether the mortality table in question is suitable for the actual type of individuals represented by the scheme membership).

A.5.6 Once the liabilities have been expressed in a suitably simplified form, it becomes possible to structure swaps that capture the main characteristics of these cash flows. Liabilities that are fixed in nominal terms would be matched using swaps that generate fixed cash flows, whilst those that are RPI linked would utilise inflation swaps. LPI linked liabilities can be catered for in a similar fashion, although often their costs seem high to clients. This seems to be because clients worry less than the market as a whole does about the possibility of inflation becoming negative.

A.5.7 Performance (and risk) measurement and attribution of the swaps portfolio can then also be carried out by reference to the simplified cash flows, discounted (probably) at swap rates, versus mark to market movements in the value of the swaps.

A.5.8 There is a link between liability driven investment and fair valuation principles. The actuary will, typically, have placed some value on the liability cash flows. Assuming that the liability cash flow projections are truly correct (and ignoring some of the niceties surrounding credit risk on cash deposits, etc.), we might ask how we can tell if this sum would actually be sufficient to provide all of the projected cash flows. This depends on whether the actuary's valuation is bigger or smaller than the *fair value* of the liabilities derivable from the mark to market value of the swaps. It is *not* sufficient merely to compare the return on the liability driven portfolio with the movement in value placed on these liabilities by the actuary. The movement needs to be unbundled into its various parts, including, potentially, a part relating to the difference between the fair valuation and the actuary's valuation.

A.5.9 Even the above analysis involves simplifications. For example, there is an implicit assumption in the above that the fund's mortality experience can be well predicted at outset, but merely differentiating between nominal, real and LPI linked increases provides no protection against unexpected improvements in mortality. There may be future discretionary benefit improvements. Active members' liabilities are particularly difficult to project reliably in this context, given their sensitivity to uncertain future member specific salary increases. For a full picture, one would, in principle, differentiate between each such risk, as per Section 4. In practice this is likely to be challenging, although at least thinking about such matters may help to highlight what sorts of risks a liability driven investment portfolio does, or does not, hedge against.

A.6 *Alternative Approaches*

A.6.1 The above overlay approach clearly demarcates who is responsible for what, but trustees might prefer merely to set their investment manager a liability driven benchmark akin to the one described above, and say: "Get on with it", with the investment manager free to use whatever instruments it likes (including swaps and other derivatives), and whenever it likes, to match the liabilities or, preferably, to add value versus them.

A.6.2 Key requirements for such an approach are for the trustees and their consultants to craft very carefully an appropriate liability driven benchmark as above, for the fund manager to have good systems for measuring, at all times, how far its portfolio deviates from this benchmark, and for it to be very clear exactly what is expected of the fund manager. The bespoke nature of such a service is likely to make it practical only for larger accounts. It is worth noting that, if the fund manager cannot practically hedge a particular part of the liability benchmark, then there will be a 'random' element to his performance. The fund manager may stress this whenever he thinks it has worked to his disadvantage, and the trustees may do the opposite whenever they think it has worked in the fund manager's favour. Unfortunately, there is almost certain to be disagreement about which is the case, unless the whole arrangement is very carefully managed. An advantage of the swaps overlay approach, described above, is that it airs and manages these potential disagreements at outset, via the discussions needed around the formulation of the swaps overlay.

A.6.3 The trustees may deliberately want to adopt a strategy that deviates from the most precise liability driven benchmark. In these circumstances, a clear liability driven benchmark might still be defined, but then deliberately modified to focus on what the trustees want.

A.6.4 For example, the trustees may feel that banks might be quoting excessive prices for buying cash flows that embed option like inflation characteristics, such as those implicit in LPI linked benefits. Yet, they may still want some hedging of such risks. They might then ask the fund manager to hedge these risks in a more approximate way, using dynamic hedging, to avoid ceding this supposed profit margin to the bank. This could, perhaps, most easily be achieved by giving the investment manager a benchmark that changes in a dynamic fashion as the underlying economic parameters change. The aim would be to mimic the economic sensitivity of the fair value of the option-like characteristics, insofar as far as these depend on the parameters in question. A perfect hedging algorithm, were one to exist, would, of course, also depend on volatility, which would require the use of more complicated derivatives (but this would then defeat the point of seeking to avoid the use of such derivatives, because they are believed to offer poor value-for-money).

A.6.5 Some modification to the swaps overlay approach may be needed for smaller schemes. A single swap might be easier to have ‘segregated’ in this context than a whole bond portfolio, but there are still implicit lower limits on the sizes at which they become practical. A better alternative may be to create specially tailored long duration pooled bond funds. Several investment managers appear to be designing such products. In real life, a portfolio of pension liabilities typically gets shorter over time, so any pooled approach is unlikely to match any particular scheme’s liabilities as well as a more bespoke approach.

APPENDIX B

PERFORMANCE MEASUREMENT AND ATTRIBUTION

B.1 *The Main Steps involved in Performance Attribution*

B.1.1 The main purpose of investment performance measurement and attribution is to determine, in a quantitative sense, how well a portfolio has performed and where that performance has come from. Mathematically, performance measurement is relatively straightforward compared with risk measurement, although careful attention to accounting detail is required. Different audiences may want to see attribution subdivisions in different ways. Because results are often highly sensitive to the accuracy of input data, it can also provide a useful check of the accuracy of the underlying accounting processes. Performance attribution involves calculating the total returns for both fund and benchmark (for the relevant period), creating suitably accurate models of how these total returns can be built up from the various constituent parts, and then decomposing the differences in a way that is illuminating to the various audiences. For a hedge fund or a trading account, there might be no explicit benchmark as such, so performance attribution might, instead, concentrate on a cash benchmark.

B.1.2 The modelling process will subdivide time into various periods. Returns do not compound additively over time, but geometrically. The root time period can be as short as a single day, although such a short period can create extra work without necessarily offering any material improvement in accuracy. Even over very short periods, it may be necessary to make assumptions or approximations, or, equivalently, you may have to accept that there will be residuals that need explaining or quantifying.

B.1.3 Ideally, any performance attribution should start with the contributions to performance arising from each individual line of stock for both the fund and the benchmark. These would then be grouped together in some suitable fashion, e.g. a country/sector classification/portfolio design structure (for equity and managed funds) and/or using 'factor' exposures such as duration (for bond funds). This may involve a hierarchical structure, drilling down, potentially, several levels. Sometimes cash is kept separate, and sometimes it is aggregated with the rest of the portfolio. Security classifications need to be maintained (including relevant factor exposures). The classification of a given security and its factor exposures may change over time. If the portfolio contains derivatives or similar instruments, their values may need to be divided between two or more characteristics/factors simultaneously, often positive to one characteristic/factor and negative to another, see e.g. Kemp (1997), LIFFE (1992a) or LIFFE (1992b). Carrying out the same calculations for large numbers of funds simultaneously is facilitated by giving careful consideration to how to store all of these data in a suitable fashion, and how to process it efficiently. Many of the same data management issues also arise in practical risk management systems.

B.2 *The Mathematics behind Multi-Period Performance Attribution*

B.2.1 Suppose that we are interested in calculating the rate of return on a portfolio from time zero to time one using some suitable units of time. Suppose that there are n new money payments into or out of the portfolio in the period, of value C_j (positive for inflows, negative for outflows) occurring at times t_j, for $j = 1$ to n. The t_j are assumed to be ordered so that $0 = t_0 \leq t_1 \leq \ldots \leq t_n = 1$. The market values at the corresponding points in time (immediately after receipt of the new money) are M_j. Dividend/interest payments are treated as outflows from the relevant stock/bond sector and inflows into the cash sector, and so net to zero at the total fund level (unless the income is paid away).

B.2.2 The *time weighted rate of return* for the period is then $g = \prod_1^n (1+g_j) - 1$ where $1 + g_j = \left(M_j - C_j\right)/M_{j-1}$. The time weighted rate of return is effectively equivalent to the growth in a unit net asset value price (were the fund to be unitised and were it to accumulate income internally, ignoring complications such as bid/offer spreads, etc.) The positive or negative impact of money arriving or being withdrawn from the portfolio at opportune or inopportune times is stripped out of the calculation. Time weighted rates of return naturally compound up over time, i.e. if the time weighted rate of return in one period is g_a and in the next is g_b, then the time weighted rate of return for the combined period is g, where $1 + g = (1 + g_a) \times (1 + g_b)$.

B.2.3 The *money weighted* or *internal rate of return* on a fund over the same period is defined as the 'sensible' solution for r to the following equation (if the C_j are of differing signs, then there will usually be more than one solution, although, normally, only one would be remotely sensible):

$$M_{Start}(1+r) + \sum_{j=1}^{n} C_j(1+r)^{(1-t_j)} = M_{End} \quad \text{where} \quad M_{Start} \equiv M_0, M_{End} \equiv M_n.$$

B.2.4 One nearly always assumes that $(1+r)^t \approx 1 + tr$. The internal rate of return can therefore be approximated by the formula $r = CR/MF$, where the *contribution to return* (CR), *net new money* (NMM), *time weighted net investment* ($TWNI$), and *mean fund* (MF), are defined as follows:

$$CR = M_{End} - M_{Start} - NNM, NMM = \sum_{j=1}^{n} C_j,$$

$$MF = M_{start} + TWNI, TWNI = \sum_{j=1}^{n} C_j(1 - t_j).$$

B.2.5 The internal rate of return is the (constant) interest rate that a bank account would need to provide (possibly negative) to return the same amount at the end of the period as the portfolio, given the same new money flows and the same start market value. Money weighted rates of return do not naturally compound up over time.

B.2.6 Calculating the time weighted rate of return, in principle, involves valuations whenever there is a cash flow. This can be time consuming, unless you have an exceptionally good valuation engine (and even then is potentially impossible if you wish to value at the exact intra-day point of time at which a particular trade takes place).

B.2.7 In practice, therefore, performance measurers often merely chain-link internal rates of return. This is because the money weighted and time weighted rates of return are the same if there are no intra-period new money flows. So, if you calculate internal rates of return sufficiently often, and chain-link them together, then the result will always tend to the time weighted rate of return.

B.2.8 In certain other special circumstances, the money weighted and time weighted rates of return are also identical. Normally, cash flows and market values will be expressed in some base currency, but suppose that we generalise the calculation of money weighted rates of return so that it can include an arbitrary *calculation numeraire*, which is worth f_j in the base currency at time t_j. The money weighted rate of return then becomes r, where $(1+r) = (1+s) \times f_n/f_0$ and where s is the solution to:

$$\frac{M_{Start}}{f_0}(1+s) + \sum_{j=1}^{n} \frac{C_j}{f_j}(1+s)^{(1-t_j)} = \frac{M_{End}}{f_n}.$$

B.2.9 The money weighted rate of return, described above, is then merely a special case of this calculation with a constant (in base currency) numeraire. Suppose that we choose $f_j = (1+g_j)$, where g_j is the true cumulative time weighted return from time 0 to time t_j. Then $s = 0$, and the money weighted rate of return r will (in this numeraire) be *identical* to the time weighted rate of return g. If f_j closely approximates to $(1+g_j)$, then s will closely approximate to 0, and the approximation $(1+s)^t \approx 1+ts$ will be very good. The money weighted rate of return, using such a numeraire, will then be very similar to the true time weighted rate of return. If the new money flows are small in relation to start *and* end market values, then the money weighted rate of return will also be very similar to the true time weighted rate of return, irrespective of the calculation numeraire.

B.2.10 The calculation numeraire can be differentiated from the *presentation numeraire* used to express the results of the calculation, which will normally be the base currency of the portfolio. If the presentation numeraire is h_j, then the rates of return would be restated to be a_j, where $(1+a) = (1+r) \times h_0/h_n$.

B.2.11 The above approach requires, not only fund holdings and price data, but also information on the prices at which individual transactions were carried out. If these are difficult to obtain, then an alternative, less exact, methodology involves *buy and hold* attribution. In this methodology, the return on each line of stock is imputed merely from market data over a given period (usually daily), on the assumption that no transactions have taken place. Such an approach produces the same answer as a true transactions based analysis, either if no transactions occur or if they occur at the prices assumed in the algorithm. Unfortunately this approximation can lead to significant residuals for funds with high turnover or subject to significant dealing costs.

B.2.12 Portfolios will, typically, contain several sectors, in which case, given the same linear approximation as used above, that the total fund and benchmark returns, r and R respectively, and their difference will be as follows, where $w_i =$ mean fund weighting for sector i, $r_i =$ return for that individual sector, etc., $b_i =$ benchmark weighting for sector i and $q_i =$ return on benchmark for sector i (since $\sum_i w_i = \sum_i b_i = 1$):

$$r = \frac{CR}{MF} = \sum_i w_i r_i \quad R = \sum_i b_i q_i \quad \text{where } r_i = \frac{CR_i}{MF_i} \text{ and } w_i = \frac{MF_i}{MF}$$

$$\Rightarrow r - R = \sum_i (w_i - b_i)(q_i - R) + \sum_i b_i(r_i - q_i) + \sum_i (w_i - b_i)(r_i - q_i)$$

$$= \sum_i AA_i + \sum_i SS_i + \sum_i IE_i \qquad \text{say.}$$

B.2.13 The AA_i are the contributions from ‘asset allocation’, the SS_i are the contributions from ‘stock selection’ and the IE_i are the contributions from an ‘interaction effect’. The interaction effect is the cross product term that arises from the fact that the value added by stock selection is based on the amount of assets involved. Typically, the interaction effect is added into stock selection if you are a ‘top-down’ manager and into asset allocation if you are a ‘bottom-up’ manager.

B.2.14 The above analysis concentrates on *additive* attribution. To make the contributions from asset allocation and stock selection chain link, they can be restated in a geometric fashion, as follows: $GAA_i = (1+g)^{(AA_i/ARR)} - 1$ and $GSS_i = (1+g)^{(SS_i/ARR)}$, where $g =$ geometric relative return at total assets level, $ARR =$ additive relative return at total assets level and AA_i and SS_i are the additive asset allocation contribution and additive stock selection contribution from sector i, or, one can use natural logarithms, using, say, $LAA_i = AA_i \log(1+g)/ARR$, so that $GAA_i = \exp(LAA_i) - 1$. The total logarithmic contribution to return from a particular source over several periods can then be found merely by adding these terms together over.

B.2.15 Decomposing returns by ‘factors’ is conceptually quite similar. However, we also need:

(a) for both fund and benchmark, the average exposure to each factor involved in the decomposition, say$(a_{fund,i,1}, a_{fund,i,2}, \ldots)$ and $(a_{bench,i,1}, a_{bench,i,2}, \ldots)$; and

(b) for the benchmark only, the return for a zero factor exposure and the extra return for a unit exposure to each individual factor, say: $(z_{bench,i,0}, z_{bench,i,1}, z_{bench,i,2}, \ldots)$, so that $R_i = z_{i,0} + \sum_{k=1} a_{bench,i,k} z_{i,k}$ and the relative return can then be decomposed into:

$$r - R = \sum_i (w_i - b_i)(q_i - R) + \sum_{i,k \geq 1} w_i \left(a_{fund,i,k} - a_{bench,i,k}\right) z_{i,k} + \sum_i w_i \left(r_i - \left(z_{i,0} + \sum_k a_{fund,i,k} z_{i,k} \right) \right).$$

B.2.16 The first term is the contribution from asset allocation, the second the component of the stock selection explained by the various factors, and the third the residual component of stock selection which is unexplained by the various factors. The second term would normally be shown decomposed by both sector and factor. The sector analysis described above is a special case, with $z_{i,k} = 0$ and with more than one value for k. We would, ideally, want to build up the $a_{fund,i,k}$ by calculating the corresponding factor exposures by line of stock, and then aggregating to the sector level. We might also do this for the benchmark as well, or we might use a separate summarised data source.

B.2.17 Currency effects can be accommodated within this framework by including, as separate ‘sectors’, any currency hedges away from the fund’s base position. If the base position is a hedged benchmark, then there would be notional reverse hedges to reintroduce exposure to that currency. Performance measurers have developed lots of other ways of taking currency into account, although many only seem particularly relevant for certain ways in which currency decisions might be taken *vis-à-vis* sector or security selection decisions.

APPENDIX C

INCORPORATING NON-NORMAL DISTRIBUTIONS

C.1 *The Cornish-Fisher Asymptotic Approximation*

C.1.1 One way of taking into account non-Normality, and thus, by implication, moments higher than the second moment, is to use the Cornish-Fisher asymptotic expansion, see Abramowitz & Stegun (1970). Let the cumulative distribution function of $Y = \sum_{i=1}^{n} Y_i$ be denoted by $F(y)$. Then the (Cornish-Fisher) asymptotic expansion (with respect to n) for the value of y_p, such that $F(y_p) = 1 - p$, is $y_p \sim m + \sigma w$, where:

(a) $w = x + \left[\gamma_1 h_1(x)\right] + \left\lfloor \gamma_2 h_2(x) + \gamma_1^2 h_{11}(x) \right\rfloor + \ldots$ (terms in brackets are terms of the same order with respect to n);

(b) m is the mean and σ the standard deviation of the distribution;

(c) κ_r are the distribution's *cumulants*, i.e. the coefficients of the power series expansion for $\ln \phi(t) = \sum_{n=0}^{\infty} \kappa_n (it)^n / n!$ (i.e. $(\phi(t))$ is the distribution's *characteristic function*);

(d) $\gamma_{r-2} = \kappa_r / \kappa_2^{r/2}$ (for $r = 3, 4, \ldots$), which means, for example, that γ_1 is the skewness and γ_2 is the (excess) kurtosis;

(e) x is the relevant cumulative Normal distribution point, i.e. $1/\sqrt{2\pi} \int_x^\infty e^{-t^2/2} dt = p$; and

(f) $h_1(x) = \frac{1}{6} He_2(x)$, $h_2(x) = \frac{1}{24} He_3(x)$, $h_{11}(x) = -\frac{1}{36}(2He_3(x) + He_1(x)), \ldots,$ where $He_n(x)$ are the Hermite polynomials

$$He_n(x) = n! \sum_{m=0}^{\mathrm{int}(n/2)} \frac{(-1)^m}{2^m m!(n-2m)!} x^{n-2m}.$$

C.1.2 Exactly how much better it typically is to use a Cornish-Fisher expansion is not something which I have seen analysed in detail. In the situation where the population is actually Normal (and the sample is large), then the Cornish Fisher expansion for the 50th percentile should be similar to the mean (maybe not exactly equal, since the sample skew, etc. may not be zero), and the sample 50th percentile would be the median. The ratio of the variance of the median to the variance of the mean is 157%, so, using the Cornish-Fisher expansion, might in this case, be 37% 'better' than using sample percentiles to determine the 50th percentile point of such an underlying population distribution. One suspects that in the tail the efficiency should be higher still, but this is almost certainly highly dependent on the distributional form.

C.2 *Copulas*

C.2.1 If we have several variables, each of which can no longer be characterised purely by their first and second moments, then the co-dependency between the variables can no longer, in general, be described solely via a correlation matrix. The most common approach that seems to be used in practice (particularly for modelling credit risk) is to use *copulas*, see e.g. Schönbucher (2003).

C.2.2 The definition of a copula is a function $C : [0, 1]^N \rightarrow [0, 1]$ where:

(a) there are random variables $U_1, \ldots, U_N$ taking values in $[0, 1]$ such that C is their distribution function; and

(b) C has uniform marginal distributions, i.e. for all $\leq N$, $u_i \in [0, 1]$, we have: $C(1, \ldots 1, u_i, 1 \ldots 1) = u_i$.

C.2.3 The basic rationale for copulas is that any joint distribution F of a set of random variables $X_1, \ldots, X_N$, i.e. $F(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N)$, can be separated into two parts. The first is the marginal distribution functions, or *marginals*, for each random variable in isolation, i.e. $F_i(.)$ where $F_i(x) = P(X_i x)$. The second is the *copula* that describes the dependence structure between the random variables. Mathematically, this decomposition relies on Sklar's theorem, which states that, if $X_1, \ldots, X_N$ are random variables with marginal distribution functions $F_1, \ldots, F_N$ and joint distribution function F, then there exists an N-dimensional copula C such that, for all $\mathbf{x} \in \Re^N$:

$$F(\mathbf{x}) = C(F_1(x_1), F_2(x_2), \ldots, F_N(x_N)) = C(\mathbf{F}(\mathbf{x}))$$

i.e. C is the joint distribution function of the unit random variables $(F_1(x_1), F_2(x_2), \ldots, F_N(x_N))$. If $F_1, \ldots, F_N$ are continuous, then C is unique.

C.2.4 A particularly simple copula is the *product* (or *independence*) copula $\Pi^N(v) = \prod_{i=1}^N v_i$. It is the copula of independent random variables. Indeed, because the copula completely specifies the dependency structure of a set of random variables, random variables $X_1, \ldots, X_N$ are independent if, and only if, their N dimensional copula is the product copula. The copula most commonly used in practice is probably the *Gaussian* copula (for a given correlation matrix). It is the copula applicable to a multivariate Normal distribution with that correlation matrix.

APPENDIX D

RISK ATTRIBUTION

D.1 *Risk Attribution: Historic Risk*

D.1.1 To highlight the similarities between risk attribution and return attribution, we start by considering how historic risk statistics might be attributed between different sources, even though risk attribution is more normally a forward looking exercise. Usually, such an attribution analysis would focus on a suitable decomposition of the variance of the historic relative returns, as it is then relatively straightforward to get the sum of the parts to equal the whole.

D.1.2 Suppose $r_t = r_{1t} + r_{2t} + \ldots + e_t$ and $b_t = b_{1t} + b_{2t} + \ldots,$ where r_{it} and b_{it} are the contributions to the fund return and benchmark return respectively due to the ith *factor*, and e_t is the residual contribution to the fund return not explained by any factor. The factors here might be market exposure for long/short hedge funds, duration and convexity for bond funds, and fundamental factors/sector exposures for equity funds, etc., i.e. any 'factors' that might otherwise be used in a performance attribution analysis. They might also include any other elements that might add or subtract to the relative performance, e.g. asset allocation stances and/or expenses, tax, etc. Suppose also that there are n time periods, each is given an equal weight in the computation of the variance (which we assume is taken as the 'population' rather than the 'sample' variance), and that we measure historic risk using arithmetic variances rather than geometric or logarithmic variances. Then:

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum_t (r_t - b_t)^2$$

$$= \frac{1}{n} \times \begin{cases} \displaystyle\sum_i \sum_t (r_{it} - b_{it})^2 & \text{factor contributions} \\ \displaystyle\sum_{i,j \neq i} \sum_t (r_{it} - b_{it})(r_{jt} - b_{jt}) & \text{cross factor contributions} \\ \displaystyle\sum_t e_t^2 & \text{contribution from residual term} \\ \displaystyle\sum_i \sum_t (r_{it} - b_{it})e_t & \text{cross factor residual contributions.} \end{cases}$$

D.1.3 So variance of historic relative returns can be decomposed into various terms akin to those appearing in a performance attribution analysis, the only differences being:

(a) the analysis concentrates on second moments, i.e. terms in $(r_{it} - b_{it})^2$ and $(r_{it} - b_{it})(r_{jt} - b_{jt})$, rather than on first moments, i.e. terms in just $(r_{it} - b_{it})$;

(b) as a result, there are cross factor terms linked to the correlation between different factors; and

(c) there are also (for historic risk attribution) contributions from cross correlations between factors and the residual term (because the observed correlation between them is not necessarily exactly zero), as well as a contribution deriving exclusively from the residual term.

D.2 *Risk Attribution: Prospective Risk as Measured by Variance of Expected Future Relative Return*

D.2.1 More common, in practice, is to attribute forward looking risk measures. Within the asset management community, it is again most common to attribute the projected variance of returns rather than any other sort of risk measure, since the same sort of additive decomposition as above then applies.

D.2.2 The only differences, from a mathematical perspective, versus historic risk attribution are:

(a) Factor and cross factor contributions again arise (as does a contribution from the residual term), but they are now derived directly from the covariance matrix assumed to underlie the risk model, i.e. from $(\mathbf{p}-\mathbf{a})^T\mathbf{V}(\mathbf{p}-\mathbf{a})$, where $\mathbf{p}$ is the vector of portfolio exposures, $\mathbf{a}$ the vector of benchmark exposures and $\mathbf{V}$ is the covariance matrix.

(b) The cross factor residual term disappears, given the usual assumption that the residual terms are uncorrelated with any factor term.

D.3 *Risk Attribution: Other Prospective Risk Measures*

D.3.1 For other risk measures such VaR (or even tracking error), risk attribution can be developed as follows. Suppose that the risk measure is defined as a function of the active positions, say, $f(\mathbf{x})$, where $\mathbf{x}=\mathbf{p}-\mathbf{a}$. We can, subject to suitable regularity conditions on f, always expand this as a Taylor series for marginal changes to $\mathbf{x}$:

$$f(\mathbf{x}+d\mathbf{x})=f(\mathbf{x})+\sum_i \frac{\partial f}{\partial x_i} dx_i.$$

D.3.2 We can always calculate marginal contributions to the risk measure using this sort of decomposition and a suitably normalised way of defining dx_i, but what will not necessarily happen is that the sum of these marginal contributions adds up to the total. Instead, the total might need to be reapportioned in proportion to the individual marginal elements to force additivity in the presentation.

D.3.3 For example, Heywood *et al.* (2003) describe a way of decomposing tracking errors using *marginal contributions to tracking error* (MCR) based on the following formula: $MCR_i = 1/\sigma_p \sum_j w_j \sigma_{ij}$. This is equivalent to the above approach with $f(x) = \sigma_p = \sqrt{\sum_{ij} w_i w_j \sigma_{ij}}$ and normalising the dx_i to be unit active money positions.

D.4 *Risk/Return Attribution in Manager Selection*

D.4.1 The approach set out above might be described as the classical way of 'attributing' risk, just as the approach set out in Appendix B might be described as the classical way of 'attributing' return. Implicit is the assumption that you know the factors contributing to risk (or return) and the exposure of the portfolio to them.

D.4.2 An apparently somewhat different methodology may be more relevant if you are also trying to ascertain the fund's exposures to difference factors merely from the observed returns. I say 'apparently' because there are strong parallels here with the apparently different risk methodologies, described in Section 6.3, that we discovered were less different than appeared at first sight.

D.4.3 Our first task is to ensure that we have the true underlying return series. If we are analysing a unitised fund, then the quoted unit return may not derive directly from the mid-market values of the underlying assets. There may be a bid/offer or swing mechanism (or a fair valuation adjustment) applied by the fund manager that is not relevant to the underlying reference series. We would, ideally, want to analyse separately such adjustments, as well as the impact of other extraneous factors like fund expenses.

D.4.4 However, stripping out such effects may not be enough. The 'quoted' mid-market values placed on illiquid instruments have a tendency to exhibit a smoother trajectory than they would do if the instruments were freely traded in the market. Dishonest fund managers could, in principle, manipulate the prices of illiquid securities, to make their fund appear less volatile, or to hide incipient underperformance, but even when fund prices have been honestly struck, they can exhibit artificial smoothness because of unconscious behavioural biases that creep into the pricing process. For example, there is a natural tendency to benchmark what is considered a sensible price quotation by reference to the last transaction in the instrument. This may be a particular issue for a hedge fund of a fund manager seeking to analyse candidate hedge funds, as some hedge fund strategies involve extensive use of less liquid instruments. It is a well-known feature of surveyors' valuations of property, see Booth & Marcato (2004). This incidentally demonstrates that the problem is not necessarily solved merely by having a third party carry out the valuations.

D.4.5 Such price smoothing shows up as autocorrelation in the return series. It can therefore be unwound by *de-correlating* the return series r_t, e.g. by assuming that there is some underlying 'true' return series s_t, and that the observed series derives from it via, say, the formula $r_t = (1 - \alpha)s_t + \alpha s_{t-1}$, estimating α (from, in effect, the autocorrelation of r_t) and then backing out s_t.

D.4.6 One can then regress the assumed true underlying return series against the different reference factors to identify the fund's apparent exposure to each factor. The exposures can be equated with the regression betas, and the value added arising from 'non-market' exposures with the regression alpha. The more variables against which the return series is regressed, the better will be the regression fit, but not necessarily its predictive capability.

D.4.7 There are, of course, lots of potential reference factors that could be used (e.g. large cap, small cap indices, value indices, growth indices, etc.). The selection of which ones to use (and how many of them to use) could be found by stepwise regression or by some sort of criterion that balanced model fit versus model complexity, e.g. the Akaike Information Criterion, Schwarz's Bayesian Information Criterion or the Empirical Information Criterion, see Billah *et al.* (2003). We also note that a perfect regression fit will be achieved if there are at least as many independent factors as there are return observations to fit, in much the same way as there is a limit to the number of non-zero eigenvalues for an observed covariance matrix.

APPENDIX E

QUANTITATIVE RETURN FORECASTING

E.1 *Quantitative Return Forecasting*

E.1.1 Many different techniques exist for trying to *predict* or *forecast* the future movements of investment markets. These range from purely judgmental to purely quantitative approaches, and from ones that concentrate on individual stocks to ones that are applied to entire markets. Quantitative return forecasting is a special case of time series analysis. Time series analysis can, in turn, be split into two main types, both of which are typically analysed in a mathematical context using *regression* techniques:

(a) Analysis of the interdependence of two or more variables *measured at the same time*, e.g. whether high inflation is associated with high asset returns. In an investment context, the aim would not be primarily to predict future asset returns directly from current inflation levels. Instead, it is assumed that, in some other way, we form an opinion on what inflation will be, which we use to determine the most appropriate investment stance to adopt. This sort of analysis is closely linked to risk modelling, see Section 6.

(b) Analysis of the interdependence of one or more variables *measured at different times*. Usually, some intuitive justification for any such interdependence will be sought, to reassure sceptical colleagues. Such links (if they can be found) can be used directly to identify profitable investment strategies.

E.1.2 A simple example of problem (a) might involve univariate linear least squares regression involving two time series x_i and y_i (for $i = 1$ to n), which satisfy the linear relationship $y_i = a + bx_i + e_i$, where the e_i are random errors each with mean zero, and a and b are unknown constants. The y_i are known as the *dependent* variables and the x_i as the *independent* variables, as y depends on x. If the e_i are independent identically distributed Normal random variables with the same variance (and same zero mean), then the maximum likelihood estimators of a and b are the values that minimise the sum of the squared forecast error, i.e. $\sum (y_i - (a - bx_i))^2$. These are also known as their *least squares estimators*.

E.1.3 For problem (b), we would incorporate a time lag in the above relationship, i.e. we would assume that stocks, markets and/or factors driving them exhibit *autoregression*. The mathematical framework involved can most easily be explained using vectors, see below. Mathematically, we assume that there is some equation governing the behaviour of the system $y_t = f(y_{t-1})$ (where y_t is in general a vector rather than a scalar quantity, and some of the ys may be unobserved *state* variables). Traditional time series analysis would assume that f is a linear function $y_t = f(y_{t-1})$, typically exhibiting *time stationarity*. We shall see later, though, that such models can only describe a relatively small number of possible market dynamics, in effect just *regular* cyclicality and purely exponential growth or decay. Sadly, traditional linear regression techniques seem to work rather poorly for direct identification of profitable investment strategies. Investment markets do show cyclical behaviour, but the frequencies of the cycles are often far from regular. It is easy to postulate variables that ought to influence markets, but much more difficult to identify ones that seem to do so consistently, whilst at the same time offering significant predictive power. Relationships that work well over some time periods often seem to work less well over others. Perhaps this is not too surprising. If successful forecasting techniques were easy to find, then, presumably, market prices would have already reacted, reducing or eliminating their potential to add value in the future. So, in this field, as in other aspects of active investment management, it is necessary to stay one step ahead of others.

E.2 *Traditional Time Series Analysis*

E.2.1 Consider, first, a situation where we only have one time series, where we are attempting to forecast future values from observed past values. For example, the time series followed by a given variable might be governed by the following relationship, where the value at time t of the variable is denoted by $y_t = cy_{t-1} + w_t$. This is a *linear first order difference equation*. A *difference equation* is an expression relating a variable y_t to its previous values. The above equation is *first order*, because only the first lag (y_{t-1}) appears on the right hand side of the equation. It is *linear*, because it expresses y_t as a linear function of y_{t-1} and the *innovations* w_t. w_t are often treated as random variables, but we do not always need to do this. It is an autoregressive model, with a unit time lag, and is, therefore, typically referred to as an $AR(1)$ model. It is also time stationary, since c is constant. Nearly all linear time series analysis assumes time invariance. We could, however, introduce secular changes, by assuming that one of the variables on which the time series is based is a dummy variable linked to time. An example, commonly referred to in the quantitative investment literature, is a dummy variable set equal to one in January, but zero otherwise, to identify whether there is any 'January' effect.

E.2.2 If we know the value y_{-1} at time $t = -1$, then we find, using *recursive substitution*, that $y_t = c^{t+1}y_{-1} + \sum_{j=0}^{t} c^{t-j}w_j$. We can also determine the effect of each individual w_t on, say, the jth further into the future value of y_t, i.e. y_{t+j}. This is sometimes called the *dynamic multiplier* $\partial y_{t+j}/\partial w_t = c^j$. If $|c| < 1$, then such a system is stable, in that the consequences of a given change in w_t will eventually die out. It is unstable if $|c| > 1$. An interesting possibility is the borderline case where $c = 1$, when the output variable y_{t+j} is the sum of its initial starting value and historical inputs.

E.2.3 We can generalise the above dynamic system to be a linear pth order difference equation by making it depend on the first p lags along with the current value of the innovation (input value) w_t, i.e. $y_t = c_1 y_{t-1} + c_2 y_{t-2} + \ldots + c_p y_{t-p} + w_t$. This can be rewritten in vector/matrix form as a first order difference equation, but relating to a *vector*, if we define the vector as follows:

$$\mathbf{g}_t \equiv \begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \ldots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} c_1 & c_2 & \ldots & c_{p-1} & c_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \ldots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \\ 0 \\ \ldots \\ 0 \end{pmatrix} \equiv \mathbf{F}.\mathbf{g}_{t-1} + \mathbf{v}_t \text{ say}$$

$$\Rightarrow \mathbf{g}_t = \mathbf{F}^{t+1}\mathbf{g}_{-1} + \sum_{j=0}^{t} \mathbf{F}^{t-j}\mathbf{v}_j.$$

E.2.4 These sorts of dynamic systems have richer structures than simple scalar difference equations. For a pth order equation, we have:

$$y_{t+j} = \sum_{k=1}^{p} f_{1,k}^{(j+1)} y_{t-k} + \sum_{k=0}^{j} f_{1,1}^{(j-k)} w_{t+k}$$

(if $f_{i,k}^{(j)}$ is the element in the ith row and kth column of $\mathbf{F}^j$). To analyse the characteristics of such a system in more detail, we first need to identify the *eigenvalues* of $\mathbf{F}$. These are the values of λ for which $|\mathbf{F} - \lambda\mathbf{I}| = 0$ where $\mathbf{I}$ is the identity matrix. They are the roots to the following equation: $\lambda^p - c_1\lambda^{p-1} - c_2\lambda^{p-2} - \ldots - c_{p-1}\lambda - c_p = 0$. A pth order equation always has p roots, but some of these may be complex numbers rather than real ones, even if (as would be the case, in practice, for investment time series) all the c_j are real numbers. Complex roots correspond to cyclical (sinusoidal) behaviour. We can, therefore, have combinations of exponential decay, exponential growth and sinusoidal (perhaps damped or inflating) behaviour. For such a system to be stable, we require all the eigenvalues λ to satisfy $|\lambda| < 1$, i.e. for their absolute values all to be less than unity.

E.2.5 Eigenvalues are closely associated with *principal components analysis*. All non-negative definite symmetric $n \times n$ matrices $\mathbf{V}$ will have n non-negative eigenvalues λ_1 to λ_n and associated eigenvectors $\mathbf{x}_1$ to $\mathbf{x}_n$ (the eigenvectors can sometimes be degenerate) that satisfy $\mathbf{V}\mathbf{x}_i = \lambda_i\mathbf{x}_i$. The eigenvalues can be the same, in which case the eigenvectors can be degenerate. The eigenvectors are orthogonal (or can be chosen to be orthogonal if they are degenerate), so that any n-vector can be written as $\mathbf{p} = p_1\mathbf{x}_1 + p_2\mathbf{x}_2 + \ldots + p_n\mathbf{x}_n$.

E.2.6 The principal components are the eigenvectors of the relevant covariance matrix corresponding to the largest eigenvalues, since they explain the greatest amount of variance when averaged over all possible positions. This is because $\mathbf{x}^T\mathbf{V}\mathbf{x} = p_1^2\lambda_1 + p_2^2\lambda_2 \ldots + p_n^2\lambda_n$. However, as explained in Section 6.7, there is no fundamental reason why all stocks should be given equal weight in this averaging process. Different weighting schemas result in different vectors being deemed 'principal'.

E.3 *The Spectrum and z-Transform of a Time Series, and AR, MA and ARMA Models*

E.3.1 An equivalent way of analysing a time series is via its *spectrum,* since we can transform a time series into a frequency spectrum (and vice versa) using Fourier transforms. Take, for example, another sort of prototypical time series model, i.e. the *moving average* or *MA* model. This assumes that the output depends purely on an input series (without autoregressive components), i.e.: $y_t = \sum_{n=0}^{M} b_n w_{t-n}$.

E.3.2 There are three equivalent characterisations of a MA model:

(a) In the *time domain* — i.e. directly via the b_n.

(b) In the form of *autocorrelations*, i.e. $\rho_\tau = E((x_t - \mu)(x_{t-\tau} - \mu))\big/\sigma^2$ (where $E(x)$ means the expected value of x and $\mu = E(x_t), \sigma^2 = E\big((x_t - \mu)^2\big)$. If the input to the system is a stochastic process with input values at different times being uncorrelated (i.e. $E(x_i x_j) = 0$ for $i \neq j$) then the autocorrelation coefficients become:

$$\rho_\tau = \begin{cases} \sum_{n=\tau}^{N} b_n b_{n-|\tau|} \Big/ \sum_{n=0}^{N} b_n^2 & |\tau| \leq N \\ 0 & |\tau| \leq N. \end{cases}$$

(c) In the *frequency domain.* If the input to a MA model is an impulse, then the spectrum of the output (i.e. the result of applying the discrete Fourier transform to the time series) is given by:

$$S = \left|1 + b_1 e^{-2\pi i.1f} + b_2 e^{-2\pi i.2f} + \ldots + b_N e^{-2\pi i.Nf}\right|^2.$$

E.3.3 It is possible to show that an *AR* model of the form described earlier has a power spectrum of the following form:

$$S = 1\Big/1\left|1 - c_1 e^{-2\pi i.1f} - c_2 e^{-2\pi i.2f} - \ldots - c_p e^{-2\pi i.pf}\right|^2.$$

The obvious next step in complexity is to have both AR and MA components in the same model, e.g. an ARMA(M,N) model, of the following form:

$$y_t = \sum_{m=1}^{M} c_m y_{t-m} + \sum_{n=0}^{N} b_n w_{t-n}.$$

E.3.4 The output of an ARMA model is most easily understood in terms of the *z-transform*, which generalises the discrete Fourier transform to the complex plane: $X(z) \equiv \sum_{t=-\infty}^{\infty} x_t z^t$. On the unit circle in the complex plane, the z-transform reduces to the discrete Fourier transform. Off the unit circle, it measures the rate of divergence or convergence of a series. Convolution of two series in the time domain corresponds to the multiplication of their z-transforms. Therefore, the z-transform of the output of an ARMA model is:

$$Y(z) = C(z)Y(z) + B(z)W(z) = \frac{B(z)}{1 - C(z)} W(z).$$

E.3.5 This has the form of an input z-transform $W(z)$ multiplied by a *transfer function* unrelated to it. The transfer function is zero at the zeros of the MA term, i.e. where $B(\mathrm{z}) = 0$, and diverges to infinity, i.e. has poles (in a complex number sense), where $C(\mathrm{z}) = 1$, unless these are cancelled by zeros in the numerator. The number of poles and zeros in this equation determines the number of *degrees of freedom* in the model. Since only a ratio appears, there is no unique ARMA model for any given system. In extreme cases, a finite order AR model can always be expressed by an infinite order MA model, and vice versa.

E.3.6 There is no fundamental reason to expect an arbitrary model to be able to be described in an ARMA form. However, if we believe that a system is linear in nature, then it is reasonable to attempt to approximate its true transfer function by a ratio of polynomials, i.e. as an ARMA model. This is a problem in function approximation. It can be shown that a suitable sequence of ratios of polynomials (called *Padé approximants*) converges faster than a power series for an arbitrary function, but this still leaves unresolved the question of what the *order* of the model should be, i.e. what values of M and N to adopt. This is, in part, linked to how best to approximate the z-transform. There are several heuristic algorithms for finding the 'right' order, for example the Akaike Information Criterion, see Billah *et al.* (2003). These heuristic approaches usually rely very heavily on the model being linear, and can also be sensitive to the assumptions adopted for the error terms.

E.3.7 If we have some a priori knowledge about the nature of the linear relationship, then our best estimate at any point in time will be updated as more knowledge becomes available in a Bayesian fashion. Updating estimates of the linear parameters in this manner is usually called applying a *Kalman filter* to the process, a technique that is also used in general insurance claims reserving.

E.4 *Generalising Linear Regression Techniques*

E.4.1 There are several ways in which we can generalise linear regression, including:

(a) *multiple regression*, in which the dependent variables (the ys in the above example) depend on several different independent variables simultaneously;

(b) *heteroscedasticity*, in which we assume that the e_i have different (known) standard deviations; we then adjust the weightings assigned to each term in the sum, giving greater weight to the terms in which we have greater confidence;

(c) *autoregression*, in which the dependent data series depends, not just on other independent data sets, but also on prior values of itself;

(d) *autoregressive heteroscedasticity*, in which the standard deviations of the e_i vary in some sort of autoregressive manner;

(e) *generalised least squares regression*, in which we assume that the dependent variables are linear combinations of functions of the x_i; least squares regression is merely a special case of this, consisting of a linear combination of two functions $f_1(x_i) = 1$ and $f_2(x_i) = x_i$; and

(f) *non-Normal random terms*, where we no longer assume that the random terms are distributed as Normal random variables. This is sometimes called *robust regression*. This may involve distributions where the maximum likelihood estimators minimise $\sum \left| y_i - (a - bx_i) \right|$, in which case the formulae for the estimators then involve medians rather than means. We can, in principle, estimate the form of the dependency by the process of *box counting*, which has close parallels with the mathematical concept of entropy, see e.g. Press *et al.* (1992) or Abarbanel (1993).

E.4.2 In all of the above refinements, if we know the form of the error terms and heteroscedasticity, then we can always transform the relationship back to a *generalised linear regression* framework by transforming the dependent variable to be linear in the independent variables. It is, thus, rather important to realise that only certain sorts of time series can be handled successfully within a linear framework, however complicated are the adjustments that we might apply as above. All such linear models are ultimately characterised by a spectrum (or to be more a precise z-transform) that, in general, involves merely rational polynomials. Thus the output of all such systems is still characterised by noise superimposed on combinations of exponential decay, exponential growth, and regular sinusoidal behaviour.

E.4.3 We can, in principle, identify the dynamics of such systems by identifying the eigenvalues and eigenvectors of the corresponding matrix equations. If noise does not overwhelm the system dynamics, we should expect the spectrum/z-transform to have a small number of distinctive peaks corresponding to relevant zeros or poles applicable to the AR or MA elements. We can postulate that these correspond to the underlying dynamics of the time series. Noise will result in the spreading out of the power spectrum around these peaks. The noise can be 'removed' by replacing the observed power spectrum with one that has sharp peaks, albeit not with perfect accuracy (since we will not know exactly where the sharp peak should be positioned). For these sorts of time series problems, the degree of external noise present is, in some sense, linked to the degree of spreading of the power spectrum around its peaks.

E.4.4 However, the converse is not true. Merely because the power spectrum is broad (and without sharp peaks) does not mean that its broadband component is all due to external noise. Irregular behaviour can still appear in a perfectly deterministic framework, if the framework is *chaotic*.

E.5 *Chaotic Market Behaviour*

E.5.1 To achieve *chaotic behaviour* (at least chaotic as defined mathematically), we need to drop the assumption of time stationarity, in some shape or form. This does not mean that we need to drop time predictability. Instead, it means that the equation governing the behaviour of the system $y_t = f(y_{t-1})$ involves a non-linear function f.

E.5.2 This change can create quite radically different behaviour. Take, for example, the *logistic map* or *quadratic map*: $y_t = cy_{t-1}(1 - y_{t-1})$. In this equation y_t depends deterministically on y_{t-1} and c is a parameter that controls the qualitative behaviour of the system, ranging from $c = 0$ which generates a fixed point ($y_t = 0$) to $c = 4$, where each iteration, in effect, destroys one bit of information. In this latter case, if we know the location within ε (ε small) at one iteration, then we will only know the position within 2ε at the next iteration. This exponential increase in uncertainty or divergence of nearby trajectories is what is generally understood by the term *deterministic chaos*. This behaviour is quite different to that produced by traditional linear models. Any broadband component in the power spectrum output of a traditional linear model has to come from external noise. With non-linear systems, such output can be purely deterministically driven (and therefore, in some cases, predictable). The above example also shows that the systems do not need to be complicated to generate chaotic behaviour.

E.5.3 The main advantages of such non-linear models are that many factors influencing market behaviour can be expected to do so in a non-linear fashion, and the resultant behaviour matches observations, e.g. markets often seem to exhibit cyclical behaviour, but with the cycles having no set lengths, and markets are often relatively little affected by certain drivers in some circumstances, but affected much more by the same drivers in other circumstances.

E.5.4 The main disadvantages of non-linear models are:

(a) the mathematics is more complex;

(b) modelling underlying market dynamics in this way will make the modelling process less efficient if the underlying dynamics are, in fact, linear in nature; and

(c) if markets are chaotic, then this typically places fundamental limits on the ability of any approach to predict more than a few time steps ahead. This is because chaotic behaviour is characterised by small disturbances being magnified over time in an exponential fashion, eventually swamping the predictive power of any model that can be built up. Of course, in these circumstances, using linear approaches may be even less effective! There are purely deterministic non-linear models that are completely impossible to use for predictive purposes, even one step ahead. Take, for example, a situation in which there is a hidden state variable developing according to the following formula $x_t = 2x_{t-1} (\text{mod } 1)$, but we can only observe y_t, the integer nearest to x_t. The action of the map is most easily understood by writing x_t in a binary fractional expansion, i.e. $x_t = 0.d_1 d_2 \ldots = d_1/2 + d_2/2^2 + \ldots$). Each iteration shifts every digit to the right, so $y_t = d_t$. Thus, this system successively reveals each digit in turn. Without prior knowledge of the seeding value the output will appear to be completely random, and the past values of y_t available at time t tell us nothing at all about values at later times!

E.6 *Neural Networks*

E.6.1 Mathematicians first realised the fundamental limitations of traditional time series analysis two or three decades ago. This coincided with a time when computer scientists were particularly enthusiastic about the prospects of developing artificial intelligence. The combination led to the development of *neural networks*. A neural network is a mathematical algorithm that takes a series of inputs and produces some output dependent on these inputs. The inputs cascade through a series of steps that are conceptually modelled on the apparent behaviour of neurons in the brain. Each step (‘neuron’) takes as its input signals one or more of the input feeds (and potentially one or more of the output signals generated by other steps), and generates an output signal that would normally involve a non-linear function of the inputs (e.g. a logistic function). Typically, some of the steps are intermediate.

E.6.2 Essentially, any function of the input data can be replicated by a sufficiently complicated neural network. So, it is not enough merely to devise a single neural network. What you actually need to do is to create lots of potential alternative neural networks, and then develop some *evolutionary* or *genetic* algorithm that is used to work out which is the best one to use for a particular problem, or, more usually, you define a much narrower class of neural networks that are suitably parameterised (maybe even just one class, with a fixed number of neurons and predefined linkages between these neurons, but where the non-linear functions within each neuron are parameterised in a suitable fashion). You then *train* the neural network, by giving it some historic data, adopting a *training algorithm* that you hope will home in on an appropriate choice of parameters that are likely to work well when attempting to predict the future.

E.6.3 There was an initial flurry of interest within the financial community in neural networks, but this interest seems to have subsided. It is not that the brain does not in some respects, seem to work in the way that neural networks postulate. Rather, computerised neural networks generally proved rather poor at the sorts of tasks which they were being asked to perform.

E.7 *Locally Linear Time Series Analysis*

E.7.1 One possible reason why neural networks were found to be relatively poor at financial problems is that the effective signal to noise ratio involved in such problems may be much lower than for other types of problem, where they have proved more successful. In other words there is so much random behaviour that cannot be explained by the inputs that they struggle to make much sense of it.

E.7.2 However, even if this is not the case, it seems to me that disillusionment with neural networks was almost inevitable. Mathematically, our forecasting problem involves attempting to predict the immediate future from some past history. You must implicitly believe that the past does offer *some* guide to the future, otherwise the task is doomed to failure. If the whole of the past is uniformly relevant to predicting the immediate future, then, as we have noted above, a suitable transformation of variables moves us back into the realm of traditional linear time series, which we might, in this context, call *globally linear time series analysis*. To get the sorts of broadband characteristics that real time series return forecasting problems seem to exhibit you must, therefore, be assuming that some parts of the past are a better guide for forecasting the immediate future than other parts of the past.

E.7.3 So, it seems to me that you ought anyway, in some sense, to do the following:

(a) identify the relevance of a given element of the past to forecasting the immediate future, which one might quantify in the form of some mathematical measure of ‘distance’, where the ‘distance’ between a highly relevant element of past and the present is deemed to be small, whilst, for a less relevant element, the ‘distance’ is greater; and

(b) carry out what is now (up to a suitable transform) a *locally linear time series analysis* (only applicable to the current time), in which you give more weight to those elements of the past that are ‘closer’, in the sense of (a), to present circumstances, see e.g. Abarbanel (1993) or Weigend & Gershenfeld (1993).

E.7.4 Such an approach is *locally linear*, in the sense that it involves a linear time series analysis, but only using data that is 'local' (i.e. deemed relevant in a forecasting sense) to current circumstances. It is also implicitly how non-quantitative investment managers think. You often hear them saying that conditions are (or are not) similar to: "the bear market of 1973 to 1994", "the Russian Debt Crisis", "the Asian crisis", etc., the unwritten assumption being that what happened then is (or is not) some reasonable guide to what might happen now.

E.7.5 In addition, the approach also caters for any feature of investment markets that you think is truly applicable in all circumstances, since this is the special case where we deem the entire past to be 'local' to the present, in terms of its relevance to forecasting the future. The approach provides a true generalisation of traditional time series analysis into the chaotic domain.

E.7.6 It then becomes relatively easy to see why neural networks run into problems. Almost always, the initial definition of the neural network will be hugely over-parameterised. The training process significantly reduces this over-parameterisation, but by a difficult to determine extent. So, if you fortuitously choose a good structure that happens to start off fitting the underlying system dynamics well (or your training is fortuitous), then the neural network should perform well, but the odds of this are typically slim.

E.7.7 In contrast, a locally linear time series analysis approach should be more robust, because it starts off with far fewer parameters. If you are good at identifying the parts of the past that are particularly relevant to the present, then, suitably generalised, it should perform about as well as any possible forecasting methodology. Probably, however, the metric which you choose to define a given past's degree of relevance will identify some relevant past times more correctly than others, leading to some degradation in forecasting power. Maybe the neural networkers had it the wrong way round. Maybe the 'neural networks' within our brains are evolution's way of approximating to the locally linear framework referred to above. Or maybe 'consciousness', that elusive God given characteristic of humankind, will forever remain difficult to understand from a purely mechanical or mathematical perspective.

Printed by Bell & Bain Ltd., Glasgow