CHAPTER 11

Assessing the Quality of Risk Measures

VaR has been subjected to much criticism. In the previous chapter, we reviewed the sharpest critique: that the standard normal return model underpinning most VaR estimation procedures is simply wrong. But there are other lines of attack on VaR that are relevant even if VaR estimates are not based on the standard model. This chapter discusses three of these viewpoints:

1. The devil is in the details: Subtle and not-so-subtle differences in how VaR is computed can lead to large differences in the estimates.
2. VaR cannot provide powerful tests of its own accuracy.
3. VaR is "philosophically" incoherent: It cannot do what it purports to be able to do, namely, rank portfolios in order of riskiness.

We will also discuss a pervasive basic problem with all models, including risk models: The fact that they can err or be used inappropriately. A further major critique, the putative potential for VaR to exacerbate systemic risk, is discussed in Chapter 14.

11.1 MODEL RISK

In Chapter 10, we focused on the basic modeling problem facing VaR, that the actual distribution of returns doesn't conform to the model assumption of normality under which VaR is often computed. Using a VaR implementation that relies on normality without appreciating the deviations of the model from reality is an example of *model risk*. Models are used in risk measurement as well as in other parts of the trading and investment process. The term "model risk" describes the possibility of making incorrect trading or risk management decisions because of errors in models and how they are

applied. Model risk can manifest itself and cause losses in a number of ways. The *consequences* of model error can be trading losses, as well as adverse legal, reputational, accounting, and regulatory results.

All social science models are "wrong," in the sense that model assumptions are always more or less crude approximations to reality. In Friedman's (1953) view on the methodology of economics, deviation from reality is a virtue in a model, because the model then more readily generates testable hypotheses that can be falsified empirically, adding to knowledge. We encountered an example of this in the previous chapter. The so-called Black-Scholes biases provide very useful insights into return behavior, and yet are defined as violations of the model predictions. A model may, however, be inherently wrong, in that it is based on an incorrect overall view of reality. The data inputs can be inaccurate, or may be inappropriate to the application.

Error can be introduced into models in any number of ways. A seemingly trivial channel, but one that can have large consequences, is that the programming of a model algorithm can contain bugs. An example occurred in the ratings process for structured credit products, and was revealed during the subprime crisis. The press reported in May 2008 that Moody's had incorrectly, given their own ratings methodology, assigned AAA ratings to certain structured credit products using materially flawed programming. Another example occurred when AXA Rosenberg Group LLC, an asset-management subsidiary of the French insurance company AXA, using a quantitative investment approach, discovered a programming error in its models that had likely induced losses for some investors.[1]

These episodes also provide examples of the linkages between different types of risk. In the Moody's case, the model risk was closely linked to the reputational and liquidity risks faced by Moody's. The error had been discovered by Moody's before being reported in the press, but had coincided with changes in the ratings methodology for the affected products, and had not resulted in changes in ratings while still known only within the firm. Moody's therefore, once the bugs became public knowledge, came under suspicion of having tailored the ratings model to the desired ratings, tarnishing its reputation as an objective ratings provider. Within a few days of the episode being reported, S&P placed Moody's-issued commercial paper on negative watch, illustrating the economic costs that reputational risk events can cause. In the AXA Rosenberg episode, the discovery of the error had not been communicated in a timely fashion to investors, resulting in

[1]On Moody's, see Sam Jones, Gillian Tett, and Paul J. Davies, "CPDOs expose ratings flaw at Moody's," *Financial Times*, May 20, 2008. On AXA Rosenberg, see Jean Eaglesham and Jenny Strasburg, "Big Fine Over Bug in 'Quant' Program," Wall Street Journal, Feb. 4, 2011.

loss of assets under management, an SEC fine, and considerable overall reputational damage.

Even when software is correctly programmed, it can be used in a way that is inconsistent with the model that was intended to be implemented in the software. One type of inconsistency that arises quite frequently concerns the mapping of positions to risk factors, which we'll discuss in a moment. Such inconsistencies can contribute to differences in VaR results.

11.1.1 Valuation Risk

Model errors can occur in the valuation of securities or in hedging. Errors in valuation can result in losses that are hidden within the firm or from external stakeholders. A portfolio can be more exposed to one or more risk factors than the portfolio manager realizes because of hedging errors.

Valuation errors due to inaccurate models are examples of market risk as well as of operational risk. As a market risk phenomenon, they lead, for example, to buying securities that are thought to be cheaply priced in the market, but are in fact fairly priced or overpriced. As an operational risk phenomenon, the difficulty of valuing some securities accurately makes it possible to record positions or trades as profitable that have in fact lost money.

Model errors can, in principle, be avoided and valuation risk reduced, by relying on market prices rather than model prices. There are several problems with this approach of always marking-to-market and never *marking-to-model*. Some types of positions, such as longer-term bank commercial loans, have always been difficult to market-to-market because they do not trade frequently or at all, and because their value is determined by a complex internal process of monitoring by the lender. Accounting and regulatory standards mandating marking such positions to market have been held responsible by some for exacerbating financial instability, an issue we discuss in Chapter 14.

11.1.2 Variability of VaR Estimates

VaR also faces a wide range of practical problems. To understand these better, we'll first briefly sketch the implementation process for risk computation. This entire process and its results are sometimes referred to as the firm's "VaR model." We'll then discuss how implementation decisions can lead to differences in VaR results.

Risk management is generally carried out with the aid of computer systems that automate to some extent the process of combining data and computations, and generating reports. Risk-measurement systems are available commercially. Vendor systems are generally used by smaller financial firms. Large firms generally build their own risk-measurement systems, but may purchase some components commercially.

One particular challenge of implementing risk-measurement systems is that of data preparation. Three types of data are involved:

Market data are time series data on asset prices or other data that we can use to forecast the distribution of future portfolio returns. Obtaining appropriate time series, purging them of erroneous data points, and establishing procedures for handling missing data, are costly but essential for avoiding gross inaccuracies in risk measurement. Even with the best efforts, appropriate market data for some exposures may be unobtainable.

Security master data include descriptive data on securities, such as maturity dates, currency, and units. Corporate securities such as equities and, especially, debt securities present particular challenges in setting up security master databases. To name but one, issuer hierarchy data record which entity within a large holding company a transaction is with. Such databases are difficult to build and maintain, but are extremely important from a credit risk management point of view. Netting arrangements, for example, may differ for trades with different entities. Such issues become crucial if counterparties file for bankruptcy. Chapter 6 discussed one important example from the subprime crisis: Recovery by Lehman's counterparties depended in part on which Lehman subsidiary they had faced in the transactions.

Position data must be verified to match the firm's books and records. Position data may have to be collected from many trading systems and across a number of geographical locations within a firm.

To compute a risk measure, software is needed to correctly match up this data, and present it to a calculation engine. The engine incorporates all the formulas or computation procedures that will be used, calling them from libraries of stored procedures. The calculations have to be combined with the data appropriately. Results, finally, must be conveyed to a reporting layer that manufactures documents and tables that human managers can read. All of these steps can be carried out in myriad ways. We focus on two issues, the variability of the resulting measures, and the problem of using data appropriately.

The computation process we've just described applies to any risk measure, not just to VaR, but for concreteness, we focus on VaR. The risk manager has a great deal of discretion in actually computing a VaR. The VaR techniques we described in Chapter 3—modes of computation and the user-defined parameters—can be mixed and matched in different ways. Within each mode of computation, there are major variants, for example, the so-called "hybrid" approach of using historical simulation with

exponentially weighted return observations. This freedom is a mixed blessing. On the one hand, the risk manager has the flexibility to adapt the way he is calculating VaR to the needs of the firm, its investors, or the nature of the portfolio. On the other hand, it leads to two problems with the use of VaR in practice:

1. There is not much uniformity of practice as to confidence interval and time horizon; as a result, intuition on what constitutes a large or small VaR is underdeveloped.
2. Different ways of measuring VaR would lead to different results, even if there were standardization of confidence interval and time horizon. There are a number of computational and modeling decisions that can greatly influence VaR results, such as
   - Length of time series used for historical simulation or to estimate moments
   - Technique for estimating moments
   - Mapping techniques and the choice of risk factors, for example, maturity bucketing
   - Decay factor if applying EWMA
   - In Monte Carlo simulation, randomization technique and the number of simulations

Dramatic changes in VaR can be obtained by varying these parameters. In one well-known study (Beder, 1995), the VaRs of relatively simple portfolios consisting of Treasury bonds and S&P 500 index options were computed using different combinations of these parameters, all of them well within standard practice. For example, 100 or 250 days of historical data might be used to compute VaR via historical simulation, or Monte Carlo VaR might be computed using different correlation estimates. For a given time horizon and confidence level, VaR computations differed by a factor of six or seven times. Other oddities included VaR estimates that were higher for shorter time horizons.

A number of large banks publish VaR estimates for certain of their portfolios in their annual reports, generally accompanied by backtesting results. These VaR estimates are generated for regulatory purposes, as discussed in Chapter 15. Perusing these annual reports gives a sense of how different the VaR models can be, as they use inconsistent parameters and cannot be readily compared.

11.1.3 Mapping Issues

Mapping, the assignment of risk factors to positions, can also have a large impact on VaR results. We discussed mapping, and the broad choices risk

managers can make, in Chapter 5. Some decisions about mapping are pragmatic choices among alternatives that each have their pros and cons. An example is the choice between cash flow versus duration-convexity mapping for fixed-income. Cash flow mappings are potentially more accurate than duration mappings, since, in the former, each cash flow is mapped to a fixed income security with a roughly equal discount factor, to which the latter is clearly only an approximation. But cash flow mapping requires using many more risk factors and more complex computations, which are potentially more expensive and entail risks of data errors and other model risks.

In other cases, it may be difficult to find data that address certain risk factors. Such mapping problems may merely mirror the real-world difficulties of hedging or expressing some trade ideas. An example is the practice, said to be widespread prior to the subprime crisis, of mapping residential mortgage-backed securities (RMBS) and other securitized credit products to time series for corporate credit spreads with the same rating. Market data on securitization spreads generally is sparse, available only for very generic types of bonds and hard to update regularly from observed market prices. Figure 14.14 and the discussion in Chapter 14 illustrate how misleading such a mapping to a proxy risk factor could be. Prior to the crisis, the spread volatility of investment-grade securitizations was lower than those of corporate bonds with similar credit ratings. Yet during the financial crisis, spreads on securitizations widened, at least relatively, far more than corporate spreads. This episode illustrates not only the model risks attendant on proxy mapping, but also the inefficacy of VaR estimates in capturing large moves in market prices and the importance of stress testing.

Another example is convertible bond trading. As we saw in Chapter 10, convertible bonds can be mapped to a set of risk factors including, among others, implied volatilities, interest rates, and credit spreads. Such mappings are based on the theoretical price of a convertible bond, which is arrived at using its replicating portfolio. However, at times theoretical and market prices of converts can diverge dramatically, as can be seen in Figure 12.2. These divergences are liquidity risk events that are hard to capture with market data, so VaR based on the replicating portfolio alone can drastically understate risk. This problem can be mitigated through stress testing, which is discussed in Chapter 13.

In some cases, a position and its hedge might be mapped to the same risk factor or set of risk factors. The mapping might be justified on the grounds that the available data do not make it possible to discern between the two closely related positions. The result, however, will be a measured VaR of zero, even though there is a significant *basis risk*; that is, risk that the hedge will not provide the expected protection. Risk modeling of securitization exposures provides a pertinent example of basis risk, too. Securitizations

are often hedged with similarly-rated corporate CDS indexes. If both the underlying exposure and its CDX hedge are mapped to a corporate spread time series, the measured risk disappears. We discuss basis risk further in Chapter 13.

For some strategies, VaR can be misleading for reasons over and above the distribution of returns and VaR's dependence on specific modeling choices. For some strategies, outcomes are close to binary. One example is *event-driven* strategies, a broad class of strategies that includes trades that depend on the occurrence of terms of a corporate acquisition or merger, the outcome of bankruptcy proceedings, or of lawsuits. For many such trades, there is no historical time series of return data that would shed light on the range of results. Another example are dynamic strategies, in which the risk is generated by the trading strategy over time rather than the set of positions at a point in time. We present some tools for treating the risks of such strategies in Chapter 13.

11.1.4 Case Study: The 2005 Credit Correlation Episode

An episode of volatility in the credit markets that occurred in the late spring of 2005 provides a case study of model risk stemming from misinterpretation and misapplication of models. Some traders suffered large losses in a portfolio credit trade in which one dimension of risk was hedged in accordance with a model, while another dimension of risk was neglected. We start by reviewing the mechanics of the trade, which involved credit derivatives based on CDX.NA.IG, the investment grade CDS index.

Description of the Trade and Its Motivation A widespread trade among hedge funds, as well as proprietary trading desks of banks and brokerages, was to sell protection on the equity tranche and buy protection on the junior mezzanine tranche of the CDX.NA.IG. The trade was thus long credit and credit-spread risk through the equity tranche and short credit and credit-spread risk through the mezzanine. It was executed using several CDX.NA.IG series, particularly the IG3 introduced in September 2004 and the IG4 introduced in March 2005.

The trade was designed to be default-risk-neutral at initiation, by sizing the two legs of the trade so that their credit spread sensitivities were equal. The motivation of the trade was not to profit from a view on credit or credit spreads, though it was primarily oriented toward market risk. Rather, it was intended to achieve a positively convex payoff profile. The portfolio of two positions would then benefit from credit spread volatility. In addition, the portfolio had positive carry; that is, it earned a positive net spread. Such

trades are highly prized by traders, for whom they are akin to delta-hedged long option portfolios in which the trader receives rather than paying away time value. Compare this to the situation depicted in Figure 4.3. As we'll see, the trade was also a short credit dispersion trade, analogous to the equity dispersion trades discussed in the last chapter.

To understand the trade and its risks, we can draw on the tools we developed in Chapter 9. The securities in the extened example of that chapter are similar enough in structure to the standard tranches of the CDX.NA.IG that we can mimic the trade and understand what went wrong. Let's set up a trade in tranches of Chapter 9's illustrative CLO that is similar in structure and motivation to the standard tranche trade we have been describing. The trade takes a long credit risk position in the equity tranche and an offsetting short credit position in the mezzanine bond. Bear in mind that we would unlikely be able, in actual practice, to take a short position in a cash securitization, since the bond would be difficult to locate and borrow. We might be able to buy protection on the mezzanine tranche through a CDS, but the dealer writing it would probably charge a high spread to compensate for the illiquidity of the product and the difficulty of hedging it, in addition to the default and correlation risk. The standard tranches are synthetic CDS and their collateral pools also consist of CDS. They are generally more liquid than most other structured products, so it is easier to take short as well as long positions in them.

To determine the hedge ratio, that is, the amount of the mezzanine we are to short, we use the default sensitivities, the default01s. These are credit-risk sensitivities, while the 2005 CDX trade employed market-risk sensitivities, the spread01s. But the mechanics of hedging are similar. We assume that, at the time the trade is initiated, the expected default rate and implied correlation are $\pi = 0.03$ and $\rho = 0.30$. The default01 of a \$1,000,000 notional position in the equity is −\$6,880. The default01 of the mezzanine is −0.07212 times the notional value, so the default01 of a \$1,000,000 notional position is −\$721. These values can be read off of Figure 9.5 or Table 9.5. With a hedge ratio of about 9.54—that is, by shorting \$9,540,000 of par value of the mezzanine for every \$1,000,000 notional of long equity—we create a portfolio that, at the margin, is default-risk neutral.

Figure 11.1 illustrates how the trade was set up. At a default rate of 0.003, the portfolio has zero sensitivity to a small rise or decline in defaults. But the trade has positive convexity. The equity cheapens at a declining rate in response to spread widening. A noteworthy feature is that, because at low default rates, the mezzanine tranche has *negative* convexity, the short position *adds* positive convexity to the portfolio. The trade benefits from changes in the default rate in either direction. The actual CDX trade benefitted from large credit spread changes. It behaved, in essence, like an option

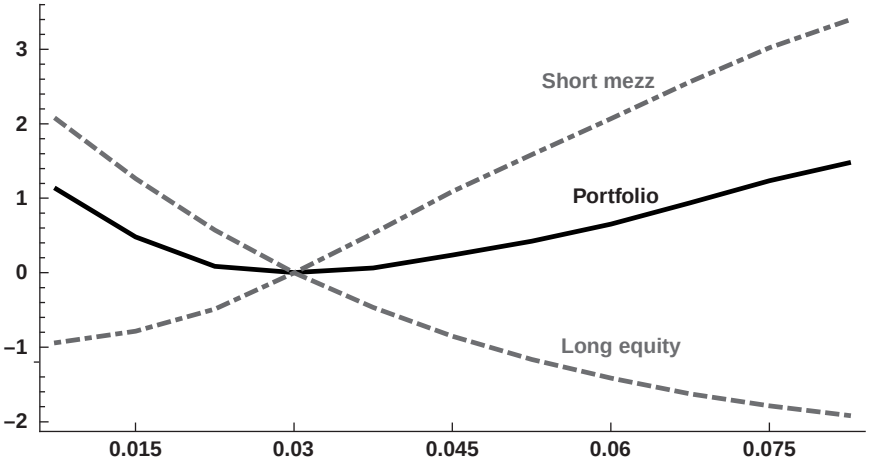

FIGURE 11.1 Convexity of CLO Liabilities
The graph plots the P&L, for varying default rates, of a portfolio consisting of (1) a long credit position in the equity tranche of the CLO described in Chapter 9 with a notional amount of $1,000,000, and (2) a short credit position in the mezzanine tranche of the same CLO with a notional amount of $1,000,000 times the hedge ratio of 9.54, that is, a par value of $9,540,000. The P&Ls of the constituent positions are also plotted. The default rates vary in the graph, but the correlation is fixed at 0.30. That is, the hedge ratio is set at a default rate of 3 percent, and a correlation of 0.30, but only the default rate is permitted to vary in the plot. The default rates are measured on the horizontal axis as decimals. The P&L is expressed on the vertical axis in millions of dollars.

straddle on credit spreads. In contrast to a typical option, however, this option, when expressed using the CDX standard tranches at the market prices prevailing in early 2005, paid a premium to its owner, rather than having negative net carry.

In the actual standard tranche trade, the mechanics were slightly different. Since the securities were synthetic CDO liabilities, traders used spread sensitivities; that is, spread01s or risk-neutral default01s, rather than actuarial default01s. The sensitivities used were not to the spreads of the underlying constituents of the CDX.NA.IG, but to the tranche spread. The hedge ratio in the actual trade was the ratio of the P&L impact of a 1bp widening of CDX.NA.IG on the equity and on the junior mezzanine tranches. The hedge ratio was between 1.5 and 2 at the beginning of 2005, lower than our example's 9.54, and at the prevailing tranche spreads, resulted in a net flow of spread income to the long equity/short mezz trade. However, the trade was set up at a particular value of implied correlation. As we will see, this was the critical error in the trade.

One additional risk should be highlighted, although it did not in the end play a crucial role in the episode we are describing: The recovery amount was at risk. In the event of a default on one or more of the names in the index, the recovery amount was not fixed but a random variable.

The Credit Environment in Early 2005 In the spring of 2005, the credit markets came under pressure, focused on the automobile industry, but not limited to it. The three large U.S.-domiciled original equipment manufacturers (OEMs), Ford, General Motors (GM), and Chrysler, had long been troubled. For decades, the OEMs had been among the most important companies in the U.S. investment-grade bond market, both in their share of issuance and in their benchmark status. The possibility of their being downgraded to junk was new and disorienting to investors. They had never been constituents of the CDX.NA.IG, but two "captive finance" companies, General Motors Acceptance Co. (GMAC) and Ford Motor Credit Co. (FMCC), were.

A third set of companies at the core of the automotive industries were the auto parts manufacturers. Delphi Corp. had been a constituent of IG3, but had been removed in consequence of its downgrade below investment grade. American Axle Co. had been added to IG4.

From a financial standpoint, the immediate priority of the OEMs had been to obtain relief from the UAW auto workers union from commitments to pay health benefits to retired workers. The "hot" part of the 2005 crisis began with two events in mid-April, the inability of GM and the UAW to reach an accord on benefits, and the announcement by GM of large losses. On May 5, GM and Ford were downgraded to junk by S&P. Moody's did the same soon after. The immediate consequence was a sharp widening of some corporate spreads, including GMAC and FMCC and other automotive industry names. Collins and Aikman, a major parts manufacturer, filed for Chapter 11 protection from creditors in May. Delphi and Visteon, another large parts manufacturer, filed later in 2005.

The two captive finance arms and the two auto parts manufacturers American Axle and Lear together constituted 4 out of the 125 constituents of the IG4. The market now contemplated the possibility of experiencing several defaults in the IG3 and IG4. The probability of extreme losses in the IG3 and IG4 standard equity tranches had appeared to be remote; it now seemed a distinct possibility. Other credit products also displayed sharp widening; the convertible bond market, in particular, was experiencing one of its periodic selloffs, as seen in Figure 12.2.

The automotive and certain other single-name spreads widened sharply, among them GMAC and FMCC. The IG indexes widened in line with the widening in their constituents, many of which did not widen at all. The pricing of the standard tranches, however, experienced much larger changes,

brought about by the panicky unwinding of the equity-mezzanine tranche trade. Figure 11.3 shows the behavior of credit spreads and the price of the standard equity tranche during the episode.

- The mark-to-market value of the equity tranche dropped sharply. This can be seen in the increase in points upfront that buyers of protection had to pay.
- The implied correlation of the equity tranche dropped sharply. Stated equivalently, its mark-to-market value dropped more and its points upfront rose more sharply than the widening of the IG 4 spread alone would have dictated.
- The junior mezzanine tranche experienced a small widening, and at times even some tightening, as market participants sought to cover positions by selling protection on the tranche, that is, taking on long credit exposures via the tranche.
- The relative value trade as a whole experienced large losses.

The implied correlation fell for two reasons. The automotive parts supplier bankruptcies had a direct effect. All were in the IG4, which meant that about 10 percent of that portfolio was now near a default state. But the correlation fell also because the widening of the IG 4 itself was constrained by hedging. The short-credit position via the equity tranche could be hedged by selling protection on a modest multiple of the mezzanine tranche, or a large multiple of the IG4 index. Although spreads were widening and the credit environment was deteriorating, at least some buyers of protection on the IG4 index found willing sellers among traders long protection in the equity tranche who were covering the short leg via the index as well as via the mezzanine tranche itself.

Modeling Issues in the Setup of the Trade The relative value trade was set up in the framework of the standard copula model, using the analytics described in Chapter 9. These analytics were simulation-based, using risk-neutral default probabilities or hazard-rate curves derived from single-name CDS. The timing of individual defaults was well modeled. Traders generally used a normal copula. The correlation assumption might have been based on the relative frequencies of different numbers of joint defaults, or, more likely, on equity return correlations or prevailing equity implied correlations, as descriped at the end of Chapter 10.

In any event, the correlation assumption was static. This was the critical flaw, rather than using the "wrong" copula function, or even the "wrong" value of the correlation. The deltas used to set the proportions of the trade were partial derivatives that did not account for changing correlation. Changing correlation drastically altered the hedge ratio between the

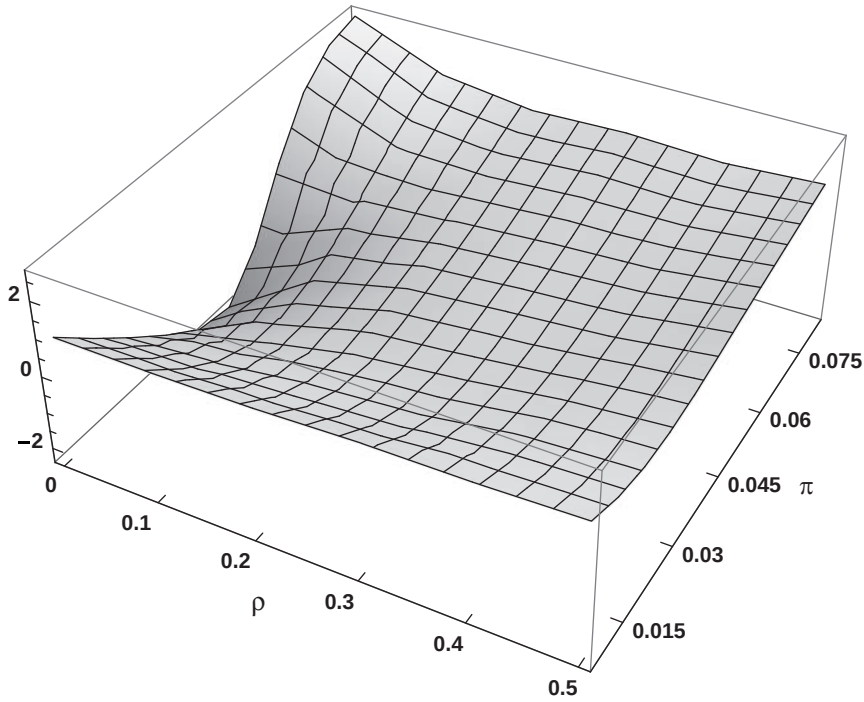

FIGURE 11.2 Correlation Risk of the Convexity Trade
The graph plots the P&L of the convexity trade for default rates from 0.0075 to 0.0825 per annum and constant pairwise Gaussian copula correlations from 0.0 to 0.5. The P&L is expressed on the vertical (z) axis in millions of dollars.

equity and mezzanine tranches, which more or less doubled to nearly 4 by July 2005. In other words, traders needed to sell protection on nearly twice the notional value of the mezzanine tranche in order to maintain spread neutrality in the portfolio.

Figure 11.2 displays the P&L profile of the trade for different spreads and correlations, again using the CLO example of Chapter 9. The portfolio P&L plotted as a solid line in Figure 11.1 is a cross-section through Figure 11.2 at a correlation of 0.30. Figure 11.2 shows that the trade was profitable for a wide range of spreads, but only if correlation did not fall. If correlation fell abruptly, and spreads did not widen enough, the trade would become highly unprofitable.

The model did not ignore correlation, but the trade thesis focused on anticipated gains from convexity. The flaw in the model could have been readily corrected if it had been recognized. The trade was put on at a time

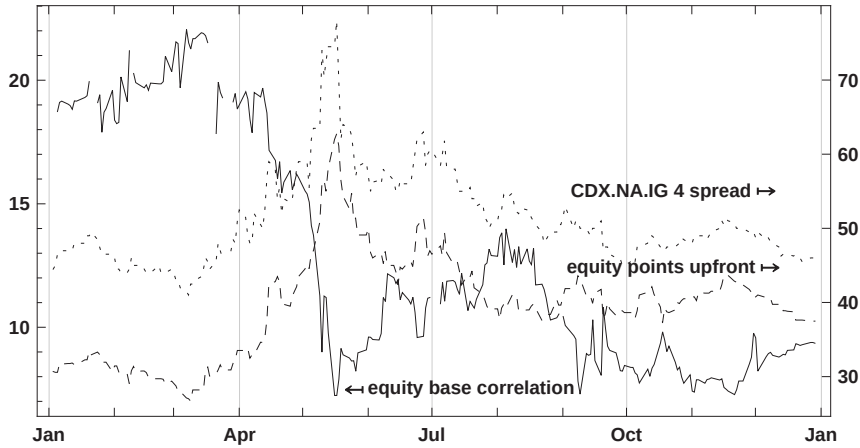

FIGURE 11.3 Implied Correlation in the 2005 Credit Episode
The graph plots the implied or base correlation of the equity (0–3 percent) tranche (solid line, percent, left axis), the price of the equity tranche (dashed line, points upfront, right axis), and the CDX IG 4 spread (dotted line, basis points, right axis). *Source:* JPMorgan Chase.

when copula models and the concept of implied correlation generally had only recently been introduced into discussions among traders, who had not yet become sensitized to the potential losses from changes in correlation. Stress testing correlation would have revealed the risk. The trade could also have been hedged against correlation risk by employing an overlay hedge: that is, by going long single-name protection in high default-probability names. In this sense, the "arbitrage" could not be captured via a two-leg trade, but required more components.

11.1.5 Case Study: Subprime Default Models

Among the costliest model risk episodes was the failure of subprime residential mortgage-based security (RMBS) valuation and risk models. These models were employed by credit-rating agencies to assign ratings to bonds, by traders and investors to value the bonds, and by issuers to structure them. While the models varied widely, two widespread defects were particularly important:

- In general, the models assumed positive future house price appreciation rates. In the stress case, house prices might fail to rise, but would not actually drop. The assumption was based on historical data, which was sparse, but suggested there had been no extended periods of falling

house prices on a large scale in any relevant historical period. As can be seen in Figure 15.1, house prices did in fact drop very severely starting in 2007. Since the credit quality of the loans depended on the borrowers' ability to refinance the loans without additional infusions of equity, the incorrect assumption on house price appreciation led to a severe underestimate of the potential default rates in underlying loan pools in an adverse economic scenario.

- Correlations among regional housing markets were assumed to be low. Bonds based on pools of loans from different geographical regions were therefore considered well-diversified. In the event, while house prices fell more severely in some regions than others, they fell—and loan defaults were much higher than expected in a stress scenario—in nearly all.

Together, these model errors or inappropriate parameters led to a substantial underestimation of the degree of systematic risk in subprime RMBS returns. Once the higher-than-expected default rates began to materialize, the rating agencies were obliged to downgrade most RMBS. The large-scale downgrades of AAA RMBS were particularly shocking to the markets, as it was precisely these that revealed the extent to which systemic risk had been underestimated and mispriced. As of the end of 2009, about 45 percent of U.S. RMBS with original ratings of AAA had been downgraded by Moody's.[2]

The inaccuracy of rating agency models for subprime RMBS is a complex phenomenon with a number of roots. As noted in Chapter 6, some observers have identified the potential conflict of interest arising from compensation of rating agencies by bond issuers as a factor in driving ratings standards lower. Others have focused on reaching for yield and the high demand for highly rated bonds with even modestly higher yields.

As we saw earlier in this chapter, a number of instances of mapping problems, contributing to seriously misleading risk measurement results, arose in securitization and structured credit products. Up until relatively recently, little time-series data was available covering securitized credit products. Highly rated securitized products were often mapped to time series of highly rated corporate bond spread indexes in risk measurement systems, or, less frequently, to the ABX index family, introduced in 2006. VaR measured using such mappings would have indicated that the bonds were unlikely under any circumstances to lose more than a few points of value. As can, however, be seen in Figure 11.4, the ABX index of the most highly rated RMBS lost 70 percent of their value during the subprime crisis. Somewhat lower, but

[2]See Moody's Investors Service (2010), p. 19.

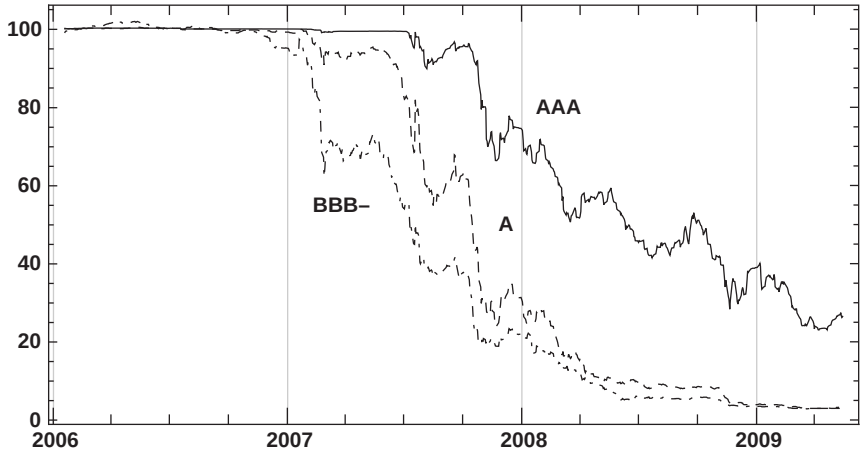

FIGURE 11.4 ABX Index of RMBS Prices
Rolling indexes of AAA, A, and BBB– ABX. For each index, the graph displays the most recent vintage.
Source: JPMorgan Chase.

still investment-grade RMBS lost almost all their value. As we will see in Chapter 14, securitizations suffered far greater losses than corporate bonds. Losses varied greatly by asset class, the year in which they were issued, or "vintage," and position in the capital structure. The corporate-bond and ABX mappings were highly misleading and would have understated potential losses by several orders of magnitude for investment-grade bonds. Similar issues arose for CMBS, and their relationship to the ratings curves and the CMBX, an index of CMBS prices analogous to the ABX.

11.2 BACKTESTING OF VAR

Assessing the accuracy of VaR estimates is important not only because firms may rely on it to assess risk, but also because, as we describe in more detail in Chapter 15, the international regulatory framework relies on it to set bank capital requirements, and to assess how accurately banks assess risk. We discuss the role of VaR in the regulatory framework in that chapter. For now, we set out some of the issues involved in assessing VaR accuracy and give some examples of the statistical techniques used.

We focus on the standard model, in which the portfolio return is normally distributed with a mean of zero. We can test VaR estimates from two points of view:

1. The first is to test the estimated standard deviation of the portfolio return, treating it as a type of parameter test. This approach is relevant only if we have estimated a parametric or Monte Carlo VaR. If we have estimated VaR by historical simulation, there is no return volatility to test.

   Statistically, setting the mean return to zero is a strong assumption, but it allows one to focus on the well-known problem of putting a confidence interval around a standard deviation estimate, rather than the much less well-known problem of jointly testing estimates of the mean and standard deviation of a distribution.
2. The second approach to assessing VaR accuracy studies the performance of the VaR rather than the accuracy of the parameters. This *backtesting* approach focuses on how often the portfolio return falls below the VaR. Such an event is often called an *excession*. The backtesting approach focuses on the VaR model itself, rather than its constituent hypotheses. In this context, "VaR model" doesn't mean the distributional hypothesis underpinning the VaR, say, normally distributed returns. Rather, it refers to the entire process, described earlier, from data gathering and position capture to implementation and reporting. Backtesting is therefore applicable to VaR estimates derived using historical as well as Monte Carlo simulation.

Several standard backtests are available. They are developed in the context of classical statistical hypothesis testing, summarized in Appendix A.4. In our context, the null hypothesis is a specific statement about the statistical distribution of excessions. The null hypothesis underpinning most backtesting, and the regulatory framework, is based on the idea that, if the model is accurate, then the proportion of excessions should be approximately equal to one minus the confidence level of the VaR.

Suppose we have a VaR estimation procedure that produces τ-period VaR estimates with a confidence level α. To simplify, we'll set $\tau = \frac{1}{\sqrt{252}}$, that is, a one-day VaR. We also assume that the VaR estimates are being generated in such a way that the estimates made in different periods are independent draws from the same distribution. That is, we assume not only an unchanging distribution of risk factor returns, but also that we're not changing the VaR estimation procedure in a way that would change the distribution of results. Excessions are then binomially distributed. The confidence level parameter α takes on the role of the event probability.

If the VaR model is accurate, then, by definition, the probability p of an excession in each period is equal to $1 - \alpha$, where α is the confidence level of the VaR, say, 99 percent. Therefore, if the VaR model is accurate, the

probability of observing x excessions in T periods, given by the binomial distribution, is

$$\binom{T}{x}(1-\alpha)^x\alpha^{T-x}$$

and one would expect the proportion of exceedances in the sequence of VaR estimates to equal $1-\alpha$:

$$\frac{x}{T} \approx 1-\alpha = p$$

A formal test takes as the null hypothesis $\mathfrak{H}_0 : p = 1-\alpha$. The log likelihood ratio test statistic is

$$2\left\{\log\left[\left(\frac{x}{T}\right)^x\left(1-\frac{x}{T}\right)^{T-x}\right] - \log\left[(1-\alpha)^x\alpha^{T-x}\right]\right\}$$

What does this expression mean? If the null hypothesis is true, then, on the one hand, we expect $\frac{x}{T}$ to be fairly close to $1-\alpha$ and the test statistic to be fairly close to zero. On the other hand, we recognize that because of random error, it's very unlikely for $\frac{x}{T}$ to be exactly $1-\alpha$. Under the null hypothesis, this test statistic is asymptotically distributed as a χ^2-variate with one degree of freedom (for the one parameter α). This distribution has most of its weight near zero, but a long right tail, as can be seen in Figure 11.5; increasing the degrees of freedom even moderately pushes the distribution away from zero. Using this distribution to carry out the test is intuitive, since a fraction of excessions different from the expected value, equal to one minus the confidence level, pushes our test statistic away from zero.

Some examples of specific interesting questions we can try to answer in this framework are:

- We can set a specific probability of a *Type I error* (false positive), that is, rejecting $\mathfrak{H}_0$ even though it is true, and determine an acceptance/non-rejection region for x, the number of excessions we observe. The probability of a Type I error is set to a level "we can live with." The acceptance region will depend on the sample size, that is the number of periods we observe the VaR model at work. The region has to be a range of integers.

  For example, if $\alpha = 0.99$, and we observe the VaR for 1,000 trading days, we would expect to see about 10 excessions. If we set the probability of a Type I error at 5 percent, the acceptance region for $\mathfrak{H}_0$ is $x \in (4, 17)$.

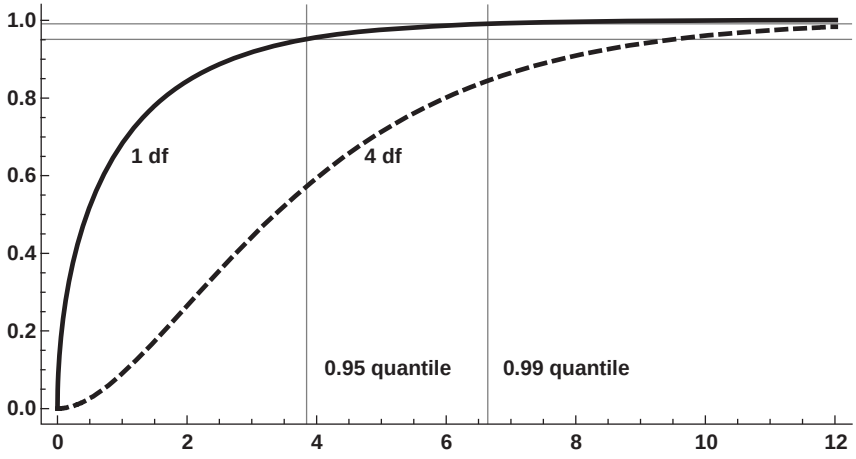

FIGURE 11.5 Chi-Square Distribution
Cumulative distribution functions of chi-square distributions with 1 and 4 degrees of freedom. The grid lines mark the 95th and 99th percentiles of the chi-square distribution with 1 df, equal to 3.841 and 6.635.

If instead we set $\alpha = 0.95$ and $T = 1{,}000$, but keep the probability of a Type I error at 5 percent, the acceptance region for $\mathfrak{H}_0$ is $x \in (37, 65)$. We expect to see a higher fraction of excessions—days on which trading losses exceed the VaR—at a lower VaR confidence level, but the range within which we feel comfortable concluding that the null has not been rejected is also wider.

- We can assess the probability of a *Type II error* (false negative), that is, nonrejection of $\mathfrak{H}_0$ even though a specific alternative hypothesis $\mathfrak{H}_1$ is true. This test will be a function of the probability of a Type I error, α, T, and the alternative hypothesis. For example, if the probability of a Type I error is fixed at 5 percent, $\alpha = 0.99$, $T = 1{,}000$, and $\mathfrak{H}_1 : p = 0.02$, then there is a 21.8 percent probability of a Type II error.

These examples show how inaccurate VaR can be. Setting $T = 1{,}000$ corresponds to almost four years of data. And after four years, we cannot reject the accuracy of a 99 percent daily VaR (Type I error) even if we observe as few as 60 percent less or as many as 70 percent more than the expected 10 excessions. And if the true probability of an excession is 2 percent, there is still over a 20 percent chance of mistakenly accepting the lower probability (Type II error), even after watching our VaR model in action for four years.

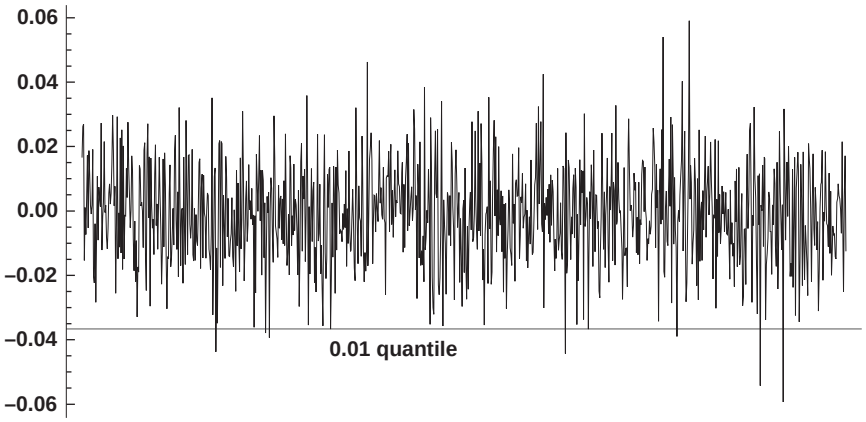

FIGURE 11.6 Backtest of a Normal Distribution
Simulation of 1,000 sequential steps of a geometric Brownian motion process with $\mu = 0$ and $\sigma = 0.25$ at an annual rate. The horizontal grid line marks the 0.01 quantile. There should be about 10 occurrences of returns below the grid lines. In this simulation, there are nine.

To provide more intuition on this test, let's look at two examples. In both examples, we look at the return series rather than a VaR, but this is just a convenient simplification: It would be trivial to multiply each return series by the number of units of a linear asset to get a VaR. In the terminology of Chapter 3, we are looking here at the VaR shock rather than the VaR.

The first example is a normal distribution. We know for this distribution that the null hypothesis is true. Figure 11.6 shows 1,000 simulation from a normal distribution with mean zero and an annual volatility of 25 percent. There happen to be nine outliers. The value of the likelihood ratio test statistic is 0.105, quite close to zero and well below either the 95th or 99th percentiles of the χ^2 distribution. We therefore do not reject the null hypothesis that the probability of a loss exceeding the VaR shock equals the confidence level of the VaR. If there had been 20 excessions, the test statistic would equal 7.827, well in excess of the critical values, and we would reject the null.

The second example is a VaR estimated using the historical simulation approach and time series data for the dollar-yen exchange rate and the S&P 500. We compute VaR shocks for each return series on its own, not as a portfolio. We use daily return data from January 2, 1996, through November 10, 2006, so the number of VaR observations is 2,744. Figure 11.7

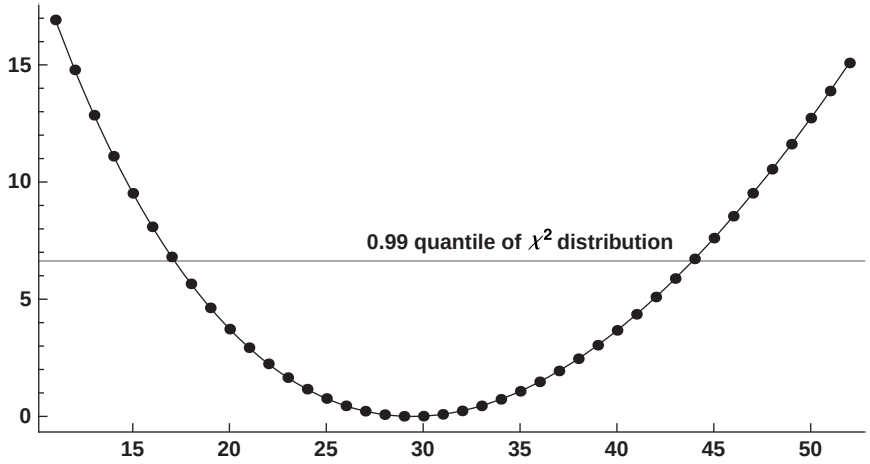

FIGURE 11.7 Likelihood-Ratio Test
Value of the likelihood ratio test statistic for the null hypothesis $\mathfrak{H}_0 : p = 0.01$ and with $T = 2{,}744$. The horizontal gridline marks the 0.99 quantile of the χ^2 distribution with one degree of freedom, denoted $\chi^2_{1,0.01} = 6.635$. It intersects the plot of the test statistic at the critical values for tests of the null at a confidence level of 99 percent.

plots the log likelihood ratio test statistic as a function of the number of excessions for $T = 2{,}744$. The number of excessions can only take on non-negative integer values, so the possible values of the test statistic are marked by dots. The horizontal grid lines mark the critical region for the 99 percent confidence level.

Figure 11.8 illustrates the tests at a confidence level of 99 percent. We can see that the S&P estimate fails the backtest (rejection of $\mathfrak{H}_0$), since there are 49 excessions. The value of the test statistic is 13.89, exceeding the upper critical value of 6.635. However, at the 95 percent confidence level, there are 141 excessions of the 0.05 quantile (not shown in Figure 11.8). Of course, these are more numerous than excessions of the 0.01 quantile, but they amount to 5.14 percent of the observations, close to the expected 5 percent if the null hypothesis is true. The test statistic is 0.113, and the critical value is 3.84, so we don't reject at the 95 percent confidence level.

For USD-JPY, we have 35 excessions and a test statistic value of 1.941, so we do not reject $\mathfrak{H}_0$ at a 99 percent confidence level. There are 130 excessions of the 0.05 quantile, or 4.74 percent of the observations. The test statistic is 0.399, and the critical value is 3.84, so we don't reject at the 95 percent confidence level, either.

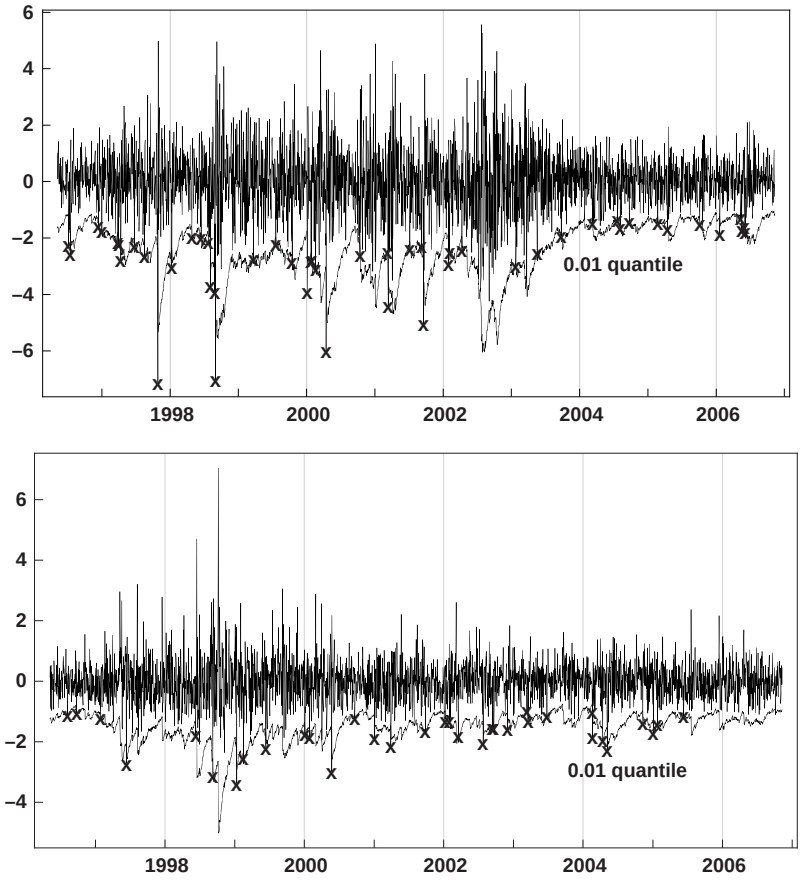

FIGURE 11.8 Historical Backtesting
Upper panel: S&P 500 index. There are 49 excessions or 1.79 percent of the sample size. The null is rejected at the 99 percent confidence level.
Lower panel: USD-JPY exchange rate. There are 35 excessions or 1.28 percent of the sample size. The null is not rejected at the 99 percent confidence level.
Both panels display return data from January 2, 1996, through November 10, 2006.
Source: Bloomberg Financial L.P.

There are a number of related tests that are discussed in the references at the end of the chapter. One similar test, for example, is based on the same model but exploiting the fact that in the model, the time to the first excession is a random variable with a known distribution.

11.3 COHERENCE OF VAR ESTIMATES

VaR has been criticized on the grounds that it cannot be grounded axiomatically and therefore lacks scientific underpinning. This is often stated as the notion that VaR is not a *coherent* risk measure.

To explain what coherence means, we need to explain the general idea of a *risk measure* as a function. As in other areas of finance theory, we start by defining a finite set Ω of I possible future states of the world. We can think of the return on a portfolio as a random variable X defined on Ω and call the set of all possible random returns X the *risk set*, denoted $\mathfrak{G}$. We can think of the X as portfolio returns; the important point here is that each $X \in \mathfrak{G}$ is a random variable with I possible outcomes. Now we can define a risk measure $\rho : \mathfrak{G} \to \mathbb{R}$ as a particular method for assigning a "single-number" measure of risk to a portfolio.

The risk measure ρ is called coherent if it has the following properties:

Monotonicity. If $X, Y \in \mathfrak{G}$, $X \geq Y$, then

$$\rho(X) \leq \rho(Y)$$

The notation $X \geq Y$ means that the value of X is at least as great as that of Y in every state $\omega \in \Omega$. The property means that if one portfolio never has a smaller return than another, it must have a smaller risk measure.

Homogeneity of degree one. For any $X \in \mathfrak{G}$ and any positive number h, we have

$$\rho(hX) = h\rho(X)$$

If you just double all the positions in a portfolio, or double the return for each outcome, the risk measure of the new portfolio must also double. Appendix A.6 explains what homogeneous functions are. We use this property in defining risk capital measures in Chapter 13.

Subadditivity. $X, Y, X + Y \in \mathfrak{G} \Rightarrow \rho(X + Y) \leq \rho(X) + \rho(Y)$

A portfolio consisting of two subportfolios can have a risk measure no greater, and possibly lower, than the sum of the risk measures of the two subportfolios. In other words, you can't reduce the risk by breaking a portfolio into pieces and measuring them separately.

Translation invariance. Let r represent the risk-free return, and let a be an amount invested in the risk-free security. Then for $X \in \mathfrak{G}$, $a \in \mathbb{R}$ we have

$$\rho(X + a \cdot r) = \rho(X) - a$$

This property means that adding a risk-free return equal to $a \cdot r$ to every possible outcome for a portfolio reduces its risk by a. In other words, adding cash to a portfolio doesn't essentially change its risk measure; it does, however, add a capital buffer against losses and reduces the risk measure by that amount.

An additional axiom is not part of the definition of coherence of a risk measure:

Relevance. If $X \in \mathfrak{G}$, $X \leq 0$, then

$$\rho(X) > 0$$

This property says that if a portfolio's return is never positive, and is negative in at least one state $\omega \in \Omega$, the risk measure must be positive. It guarantees that really bad portfolios have a large risk measure.

VaR does not have the subadditivity property. There are cases in which the VaR of a portfolio is greater than the sum of the VaRs of the individual securities in the portfolio. We provide a market risk and a credit risk example.

Example 11.1 (Failure of Subadditivity of VaR) A classic counterexample to subadditivity is a portfolio consisting of two one-day options: a short out-of-the-money put and a short out-of-the-money call. We assume logarithmic returns on the underlying security are normally distributed with a known drift and volatility.

The options expire tomorrow, so the P&L of each is equal to the accrual of a one-day option premium less tomorrow's intrinsic value if the option expires in-the-money. The options are so short-dated that there is no trading P&L from vega. We set the exercise price of each option so that its overnight, 99 percent VaR is barely zero. To do so, we need to set the exercise prices so there is a 1 percent probability of ending slightly in-the-money or better. Then the terminal intrinsic value lost exactly offsets or exceeds the option premium or time decay earned. For the put we find the exercise price X_1

such that

$$\mathbf{P}\left[S_{t+\tau} \leq X_1 - p(S_t, \tau, X_1, \sigma, r_t, q_t)\right] = 0.01$$

and for the call we set X_2 such that

$$\mathbf{P}\left[S_{t+\tau} \leq X_2 + c(S_t, \tau, X_2, \sigma, r_t, q_t)\right] = 0.99$$

The notation and parameters for the example are

Initial underlying price	$S_t = 100$
VaR confidence level is set at 99 percent	$z_* = -2.33$
Time horizon of the VaR and the option maturity	$\tau = \frac{1}{252}$
Underlying security's volatility	$\sigma = 0.30$
Overnight risk-free rate	$r_t = 0.03$
Underlying security's dividend or cash-flow rate	$q_t = 0.00$
Put exercise price	$X_1 = 95.6993$
Call exercise price	$X_2 = 104.4810$
Put fair value at time t	$p(S_t, \tau, X_1, \sigma, r_t, q_t) = 0.00615$
Call fair value at time t	$c(S_t, \tau, X_2, \sigma, r_t, q_t) = 0.00680$

The premium of the options is negligible, since they are so far out-of-the-money and close to expiry. If the put (call) exercise price were any lower (higher), the VaR would be zero at a higher confidence level.

Now consider the portfolio consisting of *both* options. The probability that at least one of the options will end in-the-money enough to incur a loss is close to 2 percent, so the 99 percent VaR of the portfolio is not zero, but a positive number. Figure 11.9 illustrates. The two-tailed 99 percent confidence interval for the portfolio has endpoints that are considerably deeper in-the-money than the one-tailed VaR of each option. There is a 1 percent probability, for the portfolio, of an outcome at one of these endpoints or worse.

This example shows that cases in which VaR is not subadditive are uncommon, but not pathological. One would not expect to observe violations of subadditivity frequently, but they do crop up.

The practical effect of a phenomenon like this is that it creates an avenue for market participants to game a system of *VaR limits*, position size limits that are based on VaR. A bank in which proprietary option trading is carried out could reduce the VaR it reports, and thus its regulatory capital, by separating the two options into two different "departments" and adding them, rather than reporting the higher consolidated VaR. But in practice, it

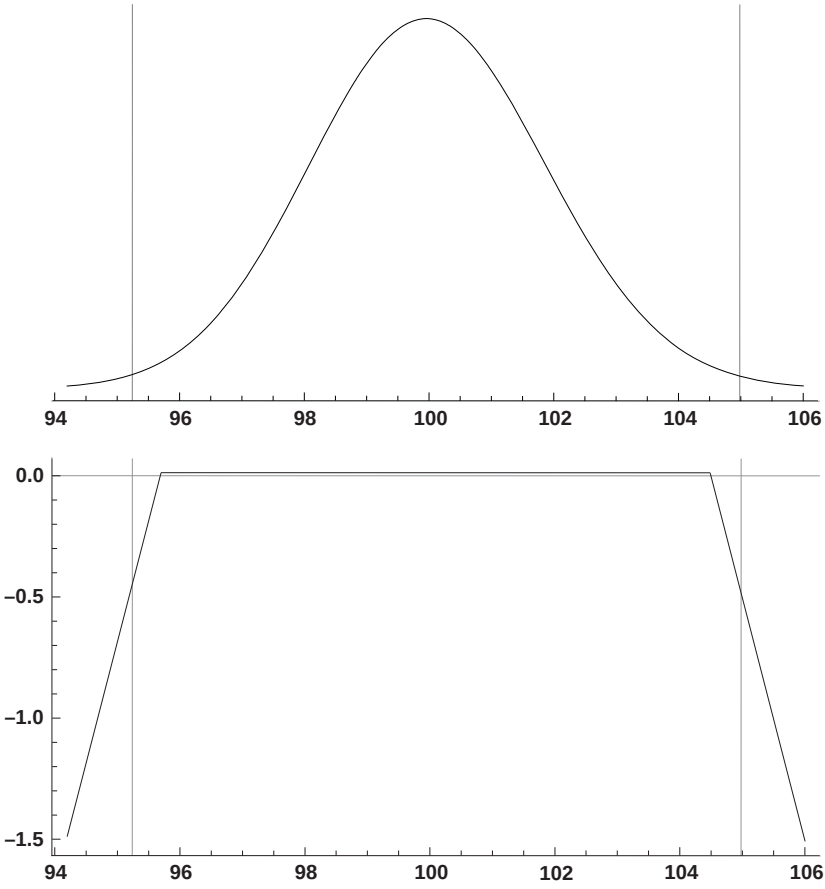

FIGURE 11.9 Failure of Subadditivity
The initial underlying price is 100.
Upper panel: One-day probability distribution of the underlying asset price.
Lower panel: Terminal P&L of the option portfolio. The intersection of the horizontal grid line at 0 with the P&L function shows the exercise prices at which each option individually breaks even with a probability of 1 percent.
The vertical grid lines mark the 99 percent confidence interval.

is unlikely that enough such anomalies could be found to have a meaningful impact.

While VaR as a measure of diversification appears unambiguously misleading in the option example, it may make more sense in others. Let's look at the apparently similar example of a credit portfolio.

Example 11.2 (Non-Subadditivity of Credit VaR) Suppose we have a set of one-year corporate bonds that pay a credit spread of 200 basis points. Let the risk-free rate be close enough to zero that it has little effect. The bonds will trade at approximately $S_t = 0.98$ dollars per dollar of par value. Let the issuers each have a default probability just under 1 percent, say, 99 basis points; defaults are independent. If there is no default, the bonds are redeemed at par, and if a default occurs, recovery is zero.

We'll look at the one year, 99 percent credit VaR of three possible portfolios, each with a current market value of \$98:

1. \$98 invested in one bond. The distribution of one-year net returns, in dollars, on this portfolio is

$$\begin{Bmatrix} 100-98 \\ 0-98 \end{Bmatrix} = \begin{Bmatrix} 2 \\ -98 \end{Bmatrix} \quad \text{w.p.} \quad \begin{Bmatrix} 0.9901 \\ 0.0099 \end{Bmatrix}$$

   Since a capital loss occurs with a probability less than 1 percent, the 99 percent VaR is zero.
2. \$49 invested in each of two different bonds. The distribution of returns on this portfolio is

$$\begin{Bmatrix} 100-98 \\ 50-98 \\ 0-98 \end{Bmatrix} = \begin{Bmatrix} 2 \\ -48 \\ -98 \end{Bmatrix} \quad \text{w.p.} \quad \begin{Bmatrix} 0.980298 \\ 0.019604 \\ 0.000098 \end{Bmatrix}$$

   The 99 percent VaR is \$48.
3. A highly granular portfolio consisting of tiny amounts invested in each of very many bonds with independent defaults, totalling \$98. There is no material risk in this portfolio. It has a virtually certain return of \$1.01, and its VaR is zero.

Subadditivity is violated, because diversifying the portfolio from one to two bonds increases the VaR massively.

But there is more to this example than meets the eye. Is the second portfolio really better for the investor? Suppose the investor is a pension fund and will be unable to meet its pension liabilities if it suffers a capital loss of \$48 or more. It is actually worse off with the second, "diversified" portfolio, because there is nearly a 2 percent probability of such a loss, even though the probability of a \$98 loss is now less than 1 basis point. With the single-bond portfolio, the probability of a catastrophic loss is only 99 basis points.

Some investors may have good reasons to be concerned with the probability as well as the size of a catastrophic loss.

FURTHER READING

Model risk is defined and discussed in Derman (1996, 2009). Crouhy, Mark, and Galai (2000b) is a textbook with an extensive treatment of model risk. See Plosser (2009) for discussion of some recent model risk episodes.

Important critiques of VaR include Danielsson (2002, 2008). Studies of variation in the implementation of VaR estimates include Beder (1995) and Marshall and Siegel (1997). Pritsker (1997) is worth reading as a survey of VaR techniques, as a study of the variability of VaR estimates, and as an approach to accuracy and backtesting of VaR.

Correlation trading is discussed in Collin-Dufresne (2009). See Coudert and Gex (2010) and Finger (2005) on the 2005 credit market episode. The references on credit correlation concepts at the end of Chapter 9 are also germane here.

Backtesting of VaR is discussed in Kupiec (1995b), Jorion (1996b), and Lopez (1999). Berkowitz (2001) proposes an approach to model testing that takes fat-tailed distributions and the potential for large losses into account. Kuester, Mittnik and Paolella (2006) employs backtesting of VaR as a means of testing volatility forecasting models. Berkowitz and O'Brien (2002) use supervisory data on bank P&L, rather than specified portfolios, to test the performance of VaR models empirically. See also Hendricks (1996).

Artzner, Delbaen, and Heath (1999) and Acerbi and Tasche (2002) discuss the concept of coherent risk measures.