SESSION 3

ISSUES IN VALUE-AT-RISK MODELING AND EVALUATION

Papers by

Jon Danielsson, Casper G. de Vries, and Bjorn N. Jorgensen

Peter F. Christoffersen, Francis X. Diebold, and Til Schuermann

Jose A. Lopez

Commentary by

Beverly Hirtle

The Value of Value at Risk: Statistical, Financial, and Regulatory Considerations

Summary of Presentation

Jon Danielsson, Casper G. de Vries, and Bjørn N. Jørgensen

Value at risk (VaR) has emerged as a major tool for measuring market risk, and it is used internally by banks for risk management and as a regulatory tool for ensuring the soundness of the financial system. A large amount of research work into VaR has emerged, and various aspects of VaR have been extensively documented. There are two areas of VaR-related research that we feel have been relatively neglected: the relationship of VaR to statistical theory and the financial-economic foundations of VaR. Most VaR methods are based on normality, however; as stated by Alan Greenspan (1997), "the biggest problems we now have with the whole evaluation of risk is the fat-tailed problem, which is really creating very large conceptual difficulties."

Common methods for measuring VaR fall into two major categories—parametric modeling of the conditional (usually normal) distribution of returns and nonparametric methods. Parametric modeling methods have been adapted from well-known forecasting technologies to the problem of VaR prediction. As a result, they seek to forecast the entire return distribution, from which only the tails are used for VaR inference.

Value at risk, however, is not about common observations. Value at risk is about extremes. For most parametric methods, the estimation of model parameters is weighted to the center of the distribution and, perversely, a method that is specifically designed to predict common events well is used to predict extremes, which are neglected in the estimation. Nonparametric historical simulation, where current portfolio weights are applied to past observations of the returns on the assets in the portfolio, does not suffer from these deficiencies. However, it suffers from the problem of tail discreteness and from the inability to provide predictions beyond the size of the data window used.

Danielsson and de Vries (1997) apply semiparametric extreme value theory to the problem of value at risk, where only the tail events are modeled parametrically, while historical simulation is used for common observations. Extreme value theory is especially designed for extremum problems, and hence their semiparametric method combines the advantages of parametric modeling of tail events and nonparametric modeling of common observations. Danielsson and de Vries (1997) develop estimators for both daily and multiday VaR predictions, and demonstrate that for their sample of U.S.

Jon Danielsson is a professor in the Department of Accounting and Finance at the London School of Economics and a contributor to the Institute of Economic Studies at the University of Iceland. Casper G. de Vries is a research fellow of Tinbergen Institute and a professor at Erasmus University. Bjørn N. Jørgensen is an assistant professor at Harvard Business School.

stock returns, the conditional parametric methods underestimate VaR and hence extreme risk, which, according to historical simulation, suffers from undesirable statistical properties in the tails. The semiparametric method, however, performs better than either a parametric conditional variance-covariance method or nonparametric historical simulation.

Conditional parametric methods typically depend on the conditional normality for the derivation of multiperiod VaR estimates, primarily because of the self-additivity of the normal distribution. The Basle Accord suggests using the so-called square-root-of-time rule to obtain multiday VaR estimates from one-day VaR values, where multiperiod volatility predictions are obtained by multiplying one-day volatility by the square root of the time horizon. However, relaxation of the normality assumption results in this scaling factor becoming incorrect. Danielsson and de Vries (1997) argue that the appropriate method for scaling up a single-day VaR to a multiday VaR is an alpha-root rule, where alpha is the number of finite-bounded moments, also known as the tail index. This eventually leads to lower multiday VaRs than would be obtained from the normal rule. Hence, the normality assumption may be, counterintuitively, overly conservative in a multiperiod analysis.

Danielsson, Hartmann, and de Vries (1998) examine the impact of these conclusions in light of the current market risk capital requirements and argue that most current methodologies underestimate the VaR, and are therefore ill-suited for market risk capital. Better VaR methods are available, such as the tail-fitting method proposed by Danielsson and de Vries (1997). However, financial institutions may be reluctant to use these methods because current market risk regulations may, perversely, provide incentives for banks to underestimate the VaR.

Danielsson, Jørgensen, and de Vries (1998) investigate the question of why regulators are interested in imposing VaR regulatory measures. Presumably, VaR reporting is meant to counter systemic risk caused by asymmetric information, that is, in a perfect market there is no need for VaR reports. But, as we argue, even if VaR reveals some hidden information, VaR-induced recapitalization may not improve the value of the firm. In our opinion, the regulatory basis for VaR is not well understood and merits further study.

REFERENCES

The authors' research papers are available on the World Wide Web at http://www.hag.hi.is/~jond/research.

Danielsson, J., and C. G. de Vries. 1997. "Value at Risk and Extreme Returns." London School of Economics, Financial Markets Group Discussion Paper no. 273.

Danielsson, J., P. Hartmann, and C. G. de Vries. 1998. "The Cost of Conservatism: Extreme Value Returns, Value-at-Risk, and the Basle 'Multiplication Factor.'" RISK, January.

Danielsson, J., B. N. Jørgensen, and C. G. de Vries. 1998. "On the (Ir)relevancy of Value-at-Risk." London School of Economics, mimeo.

Greenspan, Alan. 1997. Discussion at Federal Reserve Bank of Kansas City symposium "Maintaining Financial Stability in a Global Economy."

Horizon Problems and Extreme Events in Financial Risk Management

Peter F. Christoffersen, Francis X. Diebold, and Til Schuermann

I. INTRODUCTION

There is no one "magic" relevant horizon for risk management. Instead, the relevant horizon will generally vary by asset class (for example, equity versus bonds), industry (banking versus insurance), position in the firm (trading desk versus chief financial officer), and motivation (private versus regulatory), among other things, and thought must be given to the relevant horizon on an application-by-application basis. But one thing is clear: in many risk management situations, the relevant horizons are long—certainly longer than just a few days—an insight incorporated, for example, in Bankers Trust's RAROC system, for which the horizon is one year.

Simultaneously, it is well known that short-horizon asset return volatility fluctuates and is highly forecastable, a phenomenon that is very much at the center of modern risk management paradigms. Much less is known, however, about the forecastability of long-horizon volatility, and the speed and pattern with which forecastability decays as the horizon lengthens. A key question arises: Is volatility forecastability important for long-horizon risk management, or is a traditional constant-volatility assumption adequate?

In this paper, we address this question, exploring the interface between long-horizon financial risk management and long-horizon volatility forecastability and, in particular, whether long-horizon volatility is forecastable enough such that volatility models are useful for long-horizon risk management. In particular, we report on recent relevant work by Diebold, Hickman, Inoue, and Schuermann (1998); Christoffersen and Diebold (1997); and Diebold, Schuermann, and Stroughair (forthcoming).

To assess long-horizon volatility forecastability, it is necessary to have a measure of long-horizon volatility, which can be obtained in a number of ways. We proceed in Section II by considering two ways of converting short-horizon volatility into long-horizon volatility: scaling and formal model-based aggregation. The defects of those procedures lead us to take a different approach in Section III, estimating volatility forecastability directly at the horizons of interest, without making assumptions about the nature of the volatility process, and arriving at a surprising conclusion: Volatility forecastability seems to decline quickly with horizon, and seems to have largely vanished beyond horizons of ten or fifteen trading days.

Peter F. Christoffersen is an assistant professor of finance at McGill University. Francis X. Diebold is a professor of economics and statistics at the University of Pennsylvania, a research fellow at the National Bureau of Economic Research, and a member of the Oliver Wyman Institute. Til Schuermann is head of research at Oliv•, Wyman & Company.

If volatility forecastability is not important for risk management beyond horizons of ten or fifteen trading days, then what *is* important? The really big movements such as the U.S. crash of 1987 are still poorly understood, and ultimately the really big movements are the most important for risk management. This suggests the desirability of directly modeling the extreme tails of return densities, a task potentially facilitated by recent advances in extreme value theory. We explore that idea in Section IV, and we conclude in Section V.

II. Obtaining Long-Horizon Volatilities from Short-Horizon Volatilities: Scaling and Formal Aggregation[1]

Operationally, risk is often assessed at a short horizon, such as one day, and then converted to other horizons, such as ten days or thirty days, by scaling by the square root of horizon {for instance, as in Smithson and Minton (1996a, 1996b) or J.P. Morgan (1996)}. For example, to obtain a ten-day volatility, we multiply the one-day volatility by $\sqrt{10}$. Moreover, the horizon conversion is often significantly longer than ten days. Many banks, for example, link trading volatility measurement to internal capital allocation and risk-adjusted performance measurement schemes, which rely on annual volatility estimates. The temptation is to scale one-day volatility by $\sqrt{252}$. It turns out, however, that scaling is both inappropriate and misleading.

Scaling Works in iid Environments

Here we describe the restrictive environment in which scaling *is* appropriate. Let v_t be a log price at time t, and suppose that changes in the log price are independently and identically distributed,

$$v_t = v_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2)\,.$$

Then the one-day return is

$$v_t - v_{t-1} = \varepsilon_t\,,$$

with standard deviation σ. Similarly, the h-day return is

$$v_t - v_{t-h} = \sum_{i=0}^{h-1} \varepsilon_{t-i}\,,$$

with variance $h\sigma^2$ and standard deviation $\sqrt{h}\sigma$. Hence, the "$\sqrt{h}$ rule": to convert a one-day standard deviation to an h-day standard deviation, simply scale by $\sqrt{h}$. For some applications, a percentile of the distribution of h-day returns may be desired; percentiles also scale by $\sqrt{h}$ if log changes are not only iid, but also normally distributed.

Scaling Fails in Non-iid Environments

The scaling rule relies on one-day returns being iid, but high-frequency financial asset returns are distinctly *not* iid. Even if high-frequency portfolio returns are conditional-mean independent (which has been the subject of intense debate in the efficient markets literature), they are certainly not conditional-variance independent, as evidenced by hundreds of recent papers documenting strong volatility persistence in financial asset returns.[2]

To highlight the failure of scaling in non-iid environments and the nature of the associated erroneous long-horizon volatility estimates, we will use a simple GARCH(1,1) process for one-day returns,

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \\ \varepsilon_t &\sim NID(0, 1), \end{aligned}$$

$t = 1, \ldots, T$. We impose the usual regularity and covariance stationarity conditions, $0 < \omega < \infty$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta < 1$. The key feature of the GARCH(1,1) process is that it allows for time-varying conditional volatility, which occurs when α and/or β is nonzero. The model has been fit to hundreds of financial series and has been tremendously successful empirically; hence its popularity. We hasten to add, however, that our general thesis—that scaling fails in the non-iid environments associated with high-frequency asset returns—does not depend in any way on a GARCH(1,1) structure. Rather, we focus on the GARCH(1,1) case because it has been studied the most intensely, yielding a wealth of results that enable us to illustrate the failure of scaling both analytically and by simulation.

Drost and Nijman (1993) study the temporal aggregation of GARCH processes.[3] Suppose we begin with a sample path of a one-day return series, $\{y_{(1)t}\}_{t=1}^{T}$, which follows the GARCH(1,1) process above.[4] Then Drost and Nijman show that, under regularity conditions, the corresponding sample path of h-day returns, $\{y_{(h)t}\}_{t=1}^{T/h}$, similarly follows a GARCH (1,1) process with

$$\sigma^2_{(h)t} = \omega_{(h)} + \beta_{(h)}\sigma^2_{(h)t-1} + \alpha_{(h)}y^2_{(h)t-1},$$

where

$$\omega_{(h)} = h\omega\frac{1-(\beta+\alpha)^h}{1-(\beta+\alpha)}$$

$$\alpha_{(h)} = (\beta+\alpha)^h - \beta_{(h)},$$

and $|\beta_{(h)}| < 1$ is the solution of the quadratic equation,

$$\frac{\beta_{(h)}}{1+\beta^2_{(h)}} = \frac{a(\beta+\alpha)^h - b}{a(1+(\beta+\alpha)^{2h}) - 2b},$$

where

$$a = h(1-\beta)^2 + 2h(h-1)\frac{(1-\beta-\alpha)^2(1-\beta^2-2\beta\alpha)}{(\kappa-1)(1-(\beta+\alpha)^2)} + 4\frac{(h-1-h(\beta+\alpha)+(\beta+\alpha)^h)(\alpha-\beta\alpha(\beta+\alpha))}{1-(\beta+\alpha)^2}$$

$$b = (\alpha-\beta\alpha(\beta+\alpha))\frac{1-(\beta+\alpha)^{2h}}{1-(\beta+\alpha)^2},$$

and κ is the kurtosis of y_t. The Drost-Nijman formula is neither pretty nor intuitive, but it is important, because it is the key to correct conversion of one-day volatility to h-day volatility. It is painfully obvious, moreover, that the $\sqrt{h}$ scaling formula does not look at all like the Drost-Nijman formula.

Despite the fact that the scaling formula is incorrect, it would still be very useful if it was an accurate approximation to the Drost-Nijman formula, because of its simplicity and intuitive appeal. Unfortunately, such is not the case. As $h \to \infty$, the Drost-Nijman results, which build on those of Diebold (1988), reveal that $\alpha_{(h)} \to 0$ and $\beta_{(h)} \to 0$, which is to say that temporal aggregation produces gradual disappearance of volatility fluctuations. Scaling, in contrast, *magnifies* volatility fluctuations.

A Worked Example

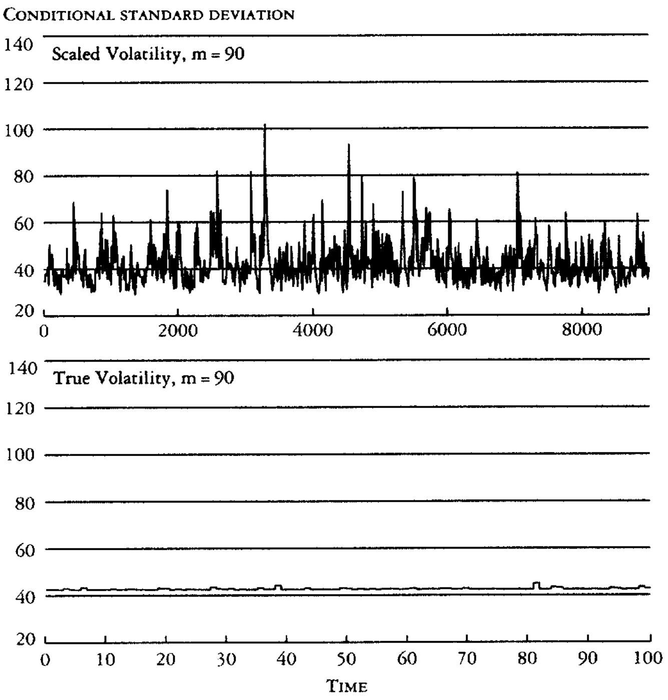
Let us examine the failure of scaling by $\sqrt{h}$ in a specific example. We parameterize the GARCH(1,1) process to be realistic for daily returns by setting $\alpha = 0.10$ and $\beta = 0.85$, which are typical of the parameter values obtained for estimated GARCH(1,1) processes. The choice of ω is arbitrary; we set $\omega = 1$.

The GARCH(1,1) process governs one-day volatility; now let us examine ninety-day volatility. In Chart 1, we show ninety-day volatilities computed in two different ways. We obtain the first (incorrect) ninety-day volatility by scaling the one-day volatility, σ_t, by $\sqrt{90}$. We obtain the second (correct) ninety-day volatility by applying the Drost-Nijman formula.

It is clear that although scaling by $\sqrt{h}$ produces volatilities that are correct on average, it magnifies the volatility fluctuations, whereas they should in fact be damped. That is, scaling produces erroneous conclusions of large fluctuations in the conditional variance of long-horizon returns, when in fact the opposite is true. Moreover, we cannot claim that the scaled volatility estimates are "conservative" in any sense; rather, they are sometimes too high and sometimes too low.

Chart 1

Ninety-Day Volatility, Scaled and Actual



Formal Aggregation Has Problems of Its Own

One might infer from the preceding discussion that formal aggregation is the key to converting short-horizon volatility estimates into good, long-horizon volatility estimates, which could be used to assess volatility forecastability. In general, such is not the case; formal aggregation has at least two problems of its own. First, temporal aggregation formulae are presently available only for restrictive classes of models; the literature has progressed little since Drost and Nijman. Second, the aggregation formulae assume the truth of the fitted model, when in fact the fitted model is simply an approximation, and the best approximation to h-day volatility dynamics is not likely to be what one gets by aggregating the best approximation (let alone a mediocre approximation) to one-day dynamics.

III. Model-Free Assessment of Volatility Forecastability at Different Horizons[5]

The model-dependent problems of scaling and aggregating daily volatility measures motivate the model-free investigation of volatility forecastability in this section. If the true process is GARCH(1,1), we know that volatility is forecastable at all horizons, although forecastability will decrease with horizon in accordance with the Drost-Nijman formula. But GARCH is only an approximation, and in this section we proceed to develop procedures that allow for assessment of volatility forecastability across horizons with no assumptions made on the underlying volatility model.

The Basic Idea

Our model-free methods build on the methods for evaluation of interval forecasts developed by Christoffersen (forthcoming). Interval forecasting is very much at the heart of modern financial risk management. The industry standard value-at-risk measure is effectively the boundary of a one-sided interval forecast, and just as the adequacy of a value-at-risk forecast depends crucially on getting the volatility dynamics right, the same is true for interval forecasts more generally.

Suppose that we observe a sample path $\{y_t\}_{t=1}^{T}$ of the asset return series y_t and a corresponding sequence of one-step-ahead interval forecasts, $\{(L_{t|t-1}(p), U_{t|t-1}(p))\}_{t=1}^{T}$, where $L_{t|t-1}(p)$ and $U_{t|t-1}(p)$ denote the lower and upper limits of the interval forecast for time t made at time $t-1$ with desired coverage probability p. We can think of $L_{t|t-1}(p)$ as a value-at-risk measure, and $U_{t|t-1}(p)$ as a measure of potential upside. The interval forecasts are subscripted by t as they will vary through time in general: in volatile times a good interval forecast should be wide and in tranquil times it should be narrow, keeping the coverage probability, p, fixed.

Now let us formalize matters slightly. Define the *hit sequence*, I_t, as

$$I_t = \begin{cases} 1, \textit{ if } y_t \in [L_{t|t-1}(p), U_{t|t-1}(p)] \\ 0, \textit{ otherwise}, \end{cases}$$

for $t = 1, 2, \ldots, T$. We will say that a sequence of interval forecasts has correct *unconditional coverage* if $E[I_t] = p$ for all t, which is the standard notion of "correct coverage."

Correct unconditional coverage is appropriately viewed as a necessary condition for adequacy of an interval forecast. It is not sufficient, however. In particular, in the presence of conditional heteroskedasticity and other higher order dynamics, it is important to check for adequacy of conditional coverage, which is a stronger concept. We will say that a sequence of interval forecasts has correct *conditional coverage with respect to an information set* Ω_{t-1} if $E[I_t|\Omega_{t-1}] = p$ for all t. The key result is that if $\Omega_{t-1} = \{I_{t-1}, I_{t-2}, \ldots, I_1\}$, then correct conditional coverage is equivalent to $\{I_t\} \overset{iid}{\sim}$ Bernoulli(p), which can readily be tested.

Consider now the case where no volatility dynamics are present. The optimal interval forecast is then constant, and given by $\{(L(p), U(p))\}, t = 1, \ldots, T$. In that case, testing for correct conditional coverage will reveal no evidence of dependence in the hit sequence, and it is exactly the independence part of the iid Bernoulli(p) criterion that is designed to pick up volatility dynamics. If, however, volatility dynamics *are* present but ignored by a forecaster who erroneously uses the constant $\{L(p), U(p)\}$ forecast, then a test for dependence in the hit sequence will reject the constant interval as an appropriate forecast: the ones and zeros in the hit sequence will tend to appear in time-dependent clusters corresponding to tranquil and volatile times.

It is evident that the interval forecast evaluation framework can be turned into a framework for assessing volatility forecastability: if a naive, constant interval forecast produces a dependent hit sequence, then volatility dynamics are present.

Measuring and Testing Dependence in the Hit Sequence

Now that we have established the close correspondence between the presence of volatility dynamics and dependence in the hit sequence from a constant interval forecast, it is time to discuss the measurement and testing of this dependence. We discuss two approaches.

First, consider a runs test, which is based on counting the number of strings, or *runs,* of consecutive zeros and ones in the hit sequence. If too few runs are observed (for example, 0000011111), the sequence exhibits positive correlation. Under the null hypothesis of independence, the exact finite sample distribution of the number of runs in the sequence has been tabulated by David (1947), and the corresponding test has been shown by Lehmann (1986) to be uniformly most powerful against a first-order Markov alternative.

We complement the runs test by a second measure, which has the benefit of being constrained to the interval [-1,1] and thus easily comparable across horizons and sequences. Let the hit sequence be first-order Markov with an arbitrary transition probability matrix. Then dependence is fully captured by the nontrivial eigenvalue, which is simply $S \equiv \pi_{11} - \pi_{01}$, where π_{ij} is the probability of a j following an i in the hit sequence. S is a natural persistence measure and has been studied by Shorrocks (1978) and Sommers and Conlisk (1979). Note that under independence $\pi_{01} = \pi_{11}$, so $S = 0$, and conversely, under strong positive persistence π_{11} will be much larger than π_{01}, so S will be large.

An Example: The Dow Jones Composite Stock Index

We now put the volatility testing framework to use in an application to the Dow Jones Composite Stock Index, which comprises sixty-five major stocks (thirty industrials, twenty transportations, and fifteen utilities) on the New York Stock Exchange. The data start on January 1, 1974, and continue through April 2, 1998, resulting in 6,327 daily observations.

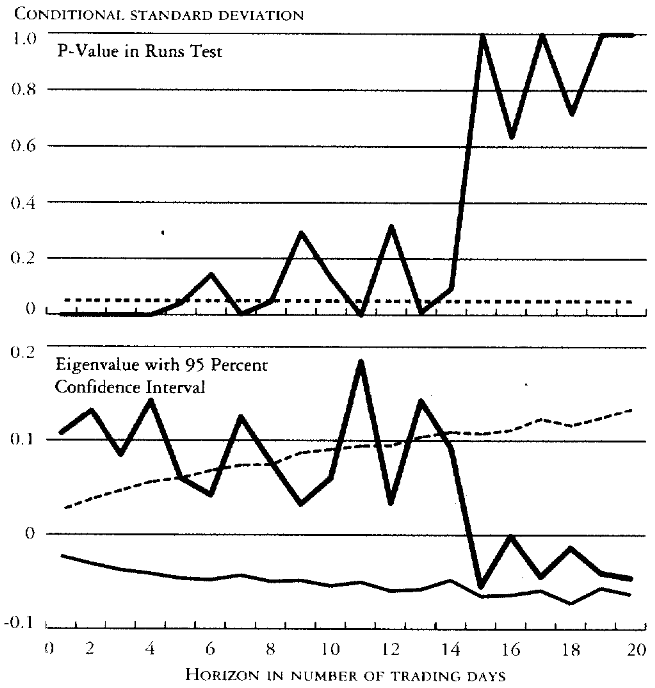
We examine asset return volatility forecastability as a function of the horizon over which the returns are calculated. We begin with daily returns and then aggregate to obtain nonoverlapping h-day returns, $h = 1, 2, 3, \ldots, 20$. We set $\{L(p), U(p)\}$ equal to ±2 standard deviations and then compute the hit sequences. Because the standard deviation varies across horizons, we let the interval vary correspondingly. Notice that p might vary across horizons, but such variation is irrelevant: we are interested only in dependence of the hit sequence, not its mean.

At each horizon, we measure volatility forecastability using the P-value of the runs test—that is, the probability of obtaining a sample that is less likely to conform to the null hypothesis of independence than does the sample at hand. If the P-value is less than 5 percent, we reject the null of independence at that particular horizon. The top panel of Chart 2 on the next page shows the P-values across horizons of one through twenty trading days. Notice that despite the jaggedness of the line, a distinct pattern emerges: at short horizons of up to a week, the P-value is very low and thus there is clear evidence of volatility forecastability. At medium horizons of two to three weeks, the P-value jumps up and down, making reliable inference difficult. At longer horizons, greater than three weeks, we find no evidence of volatility forecastability.

We also check the nontrivial eigenvalue. In order to obtain a reliable finite-sample measure of statistical significance at each horizon, we use a simulation-based resampling procedure to compute the 95 percent confidence interval under the null hypothesis of no dependence in the hit sequence (that is, the eigenvalue is zero). In the bottom panel of Chart 2, we plot the eigenvalue at each

Chart 2

Volatility Persistence across Horizons in the Dow Jones Composite Index



Notes: The hit sequence is defined relative to a constant ±2 standard deviation interval at each horizon. The top panel shows the P-value for a runs test of the hypothesis that the hit sequence is independent. The horizontal line corresponds to a 5 percent significance level. The bottom panel shows the nontrivial eigenvalue from a first-order Markov process fit to the hit sequence. The 95 percent confidence interval is computed by simulation.

horizon along with its 95 percent confidence interval. The qualitative pattern that emerges for the eigenvalue is the same as for the runs test: volatility persistence is clearly present at horizons less than a week, probably present at horizons between two and three weeks, and probably not present at horizons beyond three weeks.

Multi-Country Analysis of Equity, Foreign Exchange, and Bond Markets

Christoffersen and Diebold (1997) assess volatility forecastability as a function of horizon for many more assets and countries. In particular, they analyze stock, foreign exchange, and bond returns for the United States, the United Kingdom, Germany, and Japan, and they obtain results very similar to those presented above for the Dow Jones composite index of U.S. equities.

For all returns, the finite-sample P-values of the runs tests of independence tend to rise with the aggregation level, although the specifics differ somewhat depending on the particular return examined. As a rough rule of thumb, we summarize the results as saying that for aggregation levels of less than ten trading days we tend to reject independence, which is to say that return volatility is significantly forecastable, and conversely for aggregation levels greater than ten days.

The estimated transition matrix eigenvalues tell the same story: at very short horizons, typically from one to ten trading days, the eigenvalues are significantly positive, but they decrease quickly, and approximately monotonically, with the aggregation level. By the time one reaches ten-day returns—and often substantially before—the estimated eigenvalues are small and statistically insignificant, indicating that volatility forecastability has vanished.

IV. Forecasting Extreme Events[6]

The quick decay of volatility forecastability as the forecast horizon lengthens suggests that, if the risk management horizon is more than ten or fifteen trading days, less energy should be devoted to modeling and forecasting volatility and more energy should be devoted to modeling directly the extreme tails of return densities, a task potentially facilitated by recent advances in extreme value theory (EVT).[7] The theory typically requires independent and identically distributed observations, an assumption that appears reasonable for horizons of more than ten or fifteen trading days.

Let us elaborate. Financial risk management is intimately concerned with tail quantiles (for example, the value of the return, y, such that $P(Y > y) = .05$) and tail probabilities (for example, $P(Y > y)$, for a large value y). *Extreme* quantiles and probabilities are of particular interest, because the ability to assess them accurately translates into the ability to manage extreme financial risks effectively, such as those associated with currency crises, stock market crashes, and large bond defaults.

Unfortunately, traditional parametric statistical and econometric methods, typically based on estimation of entire densities, may be ill-suited to the assessment of extreme quantiles and event probabilities. Traditional parametric methods implicitly strive to produce a good fit in regions where most of the data fall, potentially at the expense of a good fit in the tails, where, by definition, few observations fall. Seemingly sophisticated nonparametric methods of density estimation, such as kernel smoothing, are also well known to perform poorly in the tails.

It is common, moreover, to require estimates of quantiles and probabilities not only *near* the boundary of the range of observed data, but also *beyond* the boundary. The task of estimating such quantiles and probabilities would seem to be hopeless. A key idea, however, emerges from EVT: one can estimate extreme quantiles and probabilities by fitting a "model" to the empirical survival function of a set of data using only the extreme event data rather than all the data, thereby fitting the tail and only the tail.[8] The approach has a number of attractive features, including:

- the estimation method is tailored to the object of interest—the tail of the distribution—rather than the center of the distribution, and
- an arguably reasonable functional form for the tail can be formulated from a priori considerations.

The upshot is that the methods of EVT offer hope for progress toward the elusive goal of reliable estimates of extreme quantiles and probabilities.

Let us briefly introduce the basic framework. EVT methods of tail estimation rely heavily on a power law assumption, which is to say that the tail of the survival function is assumed to be a power law times a slowly varying function:

$$P(Y>y) = k(y)\ y^{-\alpha},$$

where the "tail index," α, is a parameter to be estimated. That family includes, for example, α-stable laws with $\alpha < 2$ (but not the Gaussian case, $\alpha = 2$).

Under the power law assumption, we can base an estimator of α directly on the extreme values. The most popular, by far, is due to Hill (1975). It proceeds by ordering the observations with $y_{(1)}$ the largest, $y_{(2)}$ the second largest, and so on, and forming an estimator based on the difference between the average of the m largest log returns and the m-th largest log return:

$$\alpha = \left(\left(\frac{1}{m}\sum_{i=1}^{m} ln(y_{(i)})\right) - ln(y_{(m)})\right)^{-1}.$$

It is a simple matter to convert an estimate of α into estimates of the desired quantiles and probabilities. The Hill estimator has been used in empirical financial settings, ranging from early work by Koedijk, Schafgans, and de Vries (1990) to more recent work by Danielsson and de Vries (1997). It also has good theoretical properties; it can be shown, for example, that it is consistent and asymptotically normal, assuming the data are iid and that m grows at a suitable rate with sample size.

But beware: if tail estimation via EVT offers opportunities, it is also fraught with pitfalls, as is any attempt to estimate low-frequency features of data from short historical samples. This has been recognized in other fields, such as the empirical finance literature on long-run mean reversion in asset returns (for instance, Campbell, Lo, and MacKinlay [1997, Chapter 2]). The problem as relevant for the present context—applications of EVT in financial risk management—is that for performing statistical inference on objects such as a "once every hundred years" quantile, the relevant measure of sample size is likely better approximated by the number of nonoverlapping hundred-year intervals in the data set than by the actual number of data points. From that perspective, our data samples are terribly small relative to the demands placed on them by EVT.

Thus, we believe that best-practice applications of EVT to financial risk management will benefit from awareness of its limitations, as well as the strengths. When the smoke clears, the contribution of EVT remains basic and useful: it helps us to draw smooth curves through the extreme tails of empirical survival functions in a way that is consistent with powerful theory. Our point is simply that we should not ask more of the theory than it can deliver.

V. Concluding Remarks

If volatility is forecastable at the horizons of interest, then volatility forecasts are relevant for risk management. But our results indicate that if the horizon of interest is more than ten or fifteen trading days, depending on the asset class, then volatility is effectively not forecastable. Our results question the assumptions embedded in popular risk management paradigms, which effectively assume much greater volatility forecastability at long horizons than appears consistent with the data, and suggest that for improving long-horizon risk management, attention is better focused elsewhere. One such area is the modeling of extreme events, the probabilistic nature of which remains poorly understood, and for which recent developments in extreme value theory hold promise.

ENDNOTES

We thank Beverly Hirtle for insightful and constructive comments, but we alone are responsible for remaining errors. The views expressed in this paper are those of the authors and do not necessarily reflect those of the International Monetary Fund.

[1] This section draws on Diebold, Hickman, Inoue, and Schuermann (1997, 1998).

[2] See, for example, the surveys of volatility modeling in financial markets by Bollerslev, Chou, and Kroner (1992) and Diebold and Lopez (1995).

[3] More precisely, they define and study the temporal aggregation of *weak* GARCH processes, a formal definition of which is beyond the scope of this paper. Technically inclined readers should read "weak GARCH" whenever they encounter the word "GARCH" in this paper.

[4] Note the new and more cumbersome, but necessary, notation: the subscript, which keeps track of the aggregation level.

[5] This section draws on Christoffersen and Diebold (1997).

[6] This section draws on Diebold, Schuermann, and Stroughair (forthcoming).

[7] See the recent book by Embrechts, Klüppelberg, and Mikosch (1997), as well as the papers introduced by Paul-Choudhury (1998).

[8] The survival function is simply 1 minus the cumulative density function, $1 - F(y)$. Note, in particular, that because $F(y)$ approaches 1 as y grows, the survival function approaches 0.

REFERENCES

Andersen, T., and T. Bollerslev. Forthcoming. "Answering the Critics: Yes, ARCH Models Do Provide Good Volatility Forecasts." INTERNATIONAL ECONOMIC REVIEW.

Bollerslev, T., R. Y. Chou, and K. F. Kroner. 1992. "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence." JOURNAL OF ECONOMETRICS 52: 5-59.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay. 1997. THE ECONOMETRICS OF FINANCIAL MARKETS. Princeton.: Princeton University Press.

Christoffersen, P. F. Forthcoming. "Evaluating Interval Forecasts." INTERNATIONAL ECONOMIC REVIEW.

Christoffersen, P. F., and F. X. Diebold. 1997. "How Relevant Is Volatility Forecasting for Financial Risk Management?" Wharton Financial Institutions Center Working Paper no. 97-45.

Danielsson, J., and C. G. de Vries. 1997. "Tail Index and Quantile Estimation with Very High Frequency Data." JOURNAL OF EMPIRICAL FINANCE 4: 241-57.

David, F. N. 1947. "A Power Function for Tests of Randomness in a Sequence of Alternatives." BIOMETRIKA 34: 335-9.

Diebold, F. X. 1988. EMPIRICAL MODELING OF EXCHANGE RATE DYNAMICS. New York: Springer-Verlag.

Diebold, F. X., A. Hickman, A. Inoue, and T. Schuermann. 1997. "Converting 1-Day Volatility to h-Day Volatility: Scaling by $\sqrt{h}$ is Worse Than You Think." Wharton Financial Institutions Center Working Paper no. 97-34.

———. 1998. "Scale Models." RISK 11: 104-7. (Condensed and retitled version of Diebold, Hickman, Inoue, and Schuermann [1997].)

Diebold, F. X., and J. Lopez. 1995. "Modeling Volatility Dynamics." In Kevin Hoover, ed., MACROECONOMETRICS: DEVELOPMENTS, TENSIONS AND PROSPECTS, 427-72. Boston: Kluwer Academic Press.

Diebold, F. X., T. Schuermann, and J. Stroughair. Forthcoming. "Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management." In P. Refenes, ed., COMPUTATIONAL FINANCE. Boston: Kluwer Academic Press.

Drost, F. C., and T. E. Nijman. 1993. "Temporal Aggregation of GARCH Processes." ECONOMETRICA 61: 909-27.

Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. MODELLING EXTREMAL EVENTS. New York: Springer-Verlag.

Hill, B.M. 1975. "A Simple General Approach to Inference About the Tail of a Distribution." ANNALS OF STATISTICS 3: 1163-74.

Koedijk, K. G., M. A. Schafgans, and C. G. de Vries. 1990. "The Tail Index of Exchange Rate Returns." JOURNAL OF INTERNATIONAL ECONOMICS 29: 93-108.

Lehmann, E. L. 1986. TESTING STATISTICAL HYPOTHESES. 2d ed. New York: John Wiley.

Morgan, J.P. 1996. "RiskMetrics Technical Document." 4th ed.

Paul-Choudhury, S. 1998. "Beyond Basle." RISK 11: 89. (Introduction to a symposium on new methods of assessing capital adequacy, RISK 11: 90-107.)

Shorrocks, A.F. 1978. "The Measurement of Mobility." ECONOMETRICA 46: 1013-24.

Smithson, C., and L. Minton. 1996a. "Value at Risk." RISK 9: January.

———. 1996b. "Value at Risk (2)." RISK 9: February.

Sommers, P. M., and J. Conlisk. 1979. "Eigenvalue Immobility Measures for Markov Chains." JOURNAL OF MATHEMATICAL SOCIOLOGY 6: 253-76.

Methods for Evaluating Value-at-Risk Estimates

Jose A. Lopez

I. Current Regulatory Framework

In August 1996, the U.S. bank regulatory agencies adopted the market risk amendment (MRA) to the 1988 Basle Capital Accord. The MRA, which became effective in January 1998, requires that commercial banks with significant trading activities set aside capital to cover the market risk exposure in their trading accounts. (For further details on the market risk amendment, see *Federal Register* [1996].) The market risk capital requirements are to be based on the value-at-risk (VaR) estimates generated by the banks' own risk management models.

In general, such risk management, or VaR, models forecast the distributions of future portfolio returns. To fix notation, let y_t denote the log of portfolio value at time t. The k-period-ahead portfolio return is $\varepsilon_{t+k} = y_{t+k} - y_t$. Conditional on the information available at time t, ε_{t+k} is a random variable with distribution f_{t+k}. Thus, VaR model m is characterized by f_{mt+k}, its forecast of the distribution of the k-period-ahead portfolio return.

VaR estimates are the most common type of forecast generated by VaR models. A VaR estimate is simply a specified quantile (or critical value) of the forecasted f_{mt+k}. The VaR estimate at time t derived from model m for a k-period-ahead return, denoted $VaR_{mt}(k,\alpha)$, is the critical value that corresponds to the lower α percent tail of f_{mt+k}. In other words, VaR estimates are forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level.

Under the "internal models" approach embodied in the MRA, regulatory capital against market risk exposure is based on VaR estimates generated by banks' own VaR models using the standardizing parameters of a ten-day holding period ($k = 10$) and 99 percent coverage ($\alpha = 1$). A bank's market risk capital charge is thus based on its own estimate of the potential loss that would not be exceeded with 1 percent certainty over the subsequent two-week period. The market risk capital that bank m must hold for time $t+1$, denoted MCR_{mt+1}, is set as the larger of $VaR_{mt}(10,1)$ or a multiple of the average of the previous sixty $VaR_{mt}(10,1)$ estimates, that is,

$$MRC_{mt+1} = max\left[VaR_{mt}(10,1); \quad S_{mt} \times \frac{1}{60}\sum_{i=0}^{59} VaR_{mt-i}(10,1)\right] + SR_{mt},$$

where S_{mt} is a multiplication factor and SR_{mt} is an additional capital charge for the portfolio's idiosyncratic credit risk. Note that under the current framework $S_{mt} \geq 3$.

The S_{mt} multiplier explicitly links the accuracy of a bank's VaR model to its capital charge by varying over time. S_{mt} is set according to the accuracy of model m's VaR

Jose A. Lopez, formerly an economist at the Federal Reserve Bank of New York, is now an economist at the Federal Reserve Bank of San Francisco.

estimates for a one-day holding period ($k = 1$) and 99 percent coverage, denoted $VaR_{mt}(1,1)$ or simply VaR_{mt}. S_{mt} is a step function that depends on the number of exceptions (that is, occasions when the portfolio return ε_{t+1} is less than VaR_{mt}) observed over the last 250 trading days. The possible number of exceptions is divided into three zones. Within the green zone of four or fewer exceptions, a VaR model is deemed "acceptably accurate," and S_{mt} remains at its minimum value of three. Within the yellow zone of five to nine exceptions, S_{mt} increases incrementally with the number of exceptions. Within the red zone of ten or more exceptions, the VaR model is deemed to be "inaccurate," and S_{mt} increases to its maximum value of four.

II. Alternative Evaluation Methods

Given the obvious importance of VaR estimates to banks and now their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed: the binomial method, currently the quantitative standard embodied in the MRA, and the interval forecast method proposed by Christoffersen (forthcoming). For these tests, the null hypothesis is that the VaR estimates in question exhibit a specified property characteristic of accurate VaR estimates. If the null hypothesis is rejected, the VaR estimates do not exhibit the specified property, and the underlying VaR model can be said to be "inaccurate." If the null hypothesis is not rejected, then the model can be said to be "acceptably accurate."

However, for these evaluation methods, as with any hypothesis test, a key issue is their statistical power, that is, their ability to reject the null hypothesis when it is incorrect. If the hypothesis tests exhibit low power, then the probability of misclassifying an inaccurate VaR model as "acceptably accurate" will be high. This paper examines the power of these tests within the context of a simulation exercise.

In addition, an alternative evaluation method that is not based on a hypothesis-testing framework, but instead uses standard forecast evaluation techniques, is proposed. That is, the accuracy of VaR estimates is gauged by how well they minimize a loss function that represents the regulators' concerns. Although statistical power is not relevant for this evaluation method, the related issues of comparative accuracy and model misclassification are examined within the context of a simulation exercise. The simulation results are presented below, after the three evaluation methods are described. (See Lopez [1998] for a more complete discussion.)

Evaluation of VaR Estimates Based on the Binomial Distribution

Under the MRA, banks will report their VaR estimates to their regulators, who observe when actual portfolio losses exceed these estimates. As discussed by Kupiec (1995), assuming that the VaR estimates are accurate, such exceptions can be modeled as independent draws from a binomial distribution with a probability of occurrence equal to 1 percent. Accurate VaR estimates should exhibit the property that their unconditional coverage $\alpha^* = x/250$, where x is the number of exceptions, equals 1 percent. Since the probability of observing x exceptions in a sample of size 250 under the null hypothesis is

$$Pr(x) = \binom{250}{x} 0.01^x \times 0.99^{250-x},$$

the appropriate likelihood ratio statistic for testing whether $\alpha^* = 0.01$ is

$$LR_{uc} = 2[\log(\alpha^{*x}(1-\alpha^*)^{250-x}) \\ - \log(0.01^x \times 0.99^{250-x})].$$

Note that the LR_{uc} test is uniformly most powerful for a given sample size and that the statistic has an asymptotic $\chi^2(1)$ distribution.

Evaluation of VaR Estimates Using the Interval Forecast Method

VaR estimates are also interval forecasts of the lower 1 percent tail of f_{t+1}, the one-step-ahead return distribution. Interval forecasts can be evaluated conditionally or unconditionally, that is, with or without reference to the information available at each point in time. The LR_{uc} test is an unconditional test since it simply counts exceptions over the entire period. However, in the presence of variance dynamics, the conditional accuracy of interval forecasts is an important issue. Interval forecasts that ignore variance dynamics may have correct unconditional coverage, but at any given time, they will have incorrect conditional coverage. In such cases, the LR_{uc} test is of limited use since it will classify inaccurate VaR estimates as "acceptably accurate."

The LR_{cc} test, adapted from the more general test proposed by Christoffersen (forthcoming), is a test of correct conditional coverage. Given a set of VaR estimates, the indicator variable I_{mt+1} is constructed as

$$I_{mt+1} = \begin{cases} 1 \text{ if } \varepsilon_{t+1} < VaR_{mt} \\ 0 \text{ if } \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Since accurate VaR estimates exhibit the property of correct conditional coverage, the I_{mt+1} series must exhibit both correct unconditional coverage and serial independence. The LR_{cc} test is a joint test of these two properties. The relevant test statistic is $LR_{cc} = LR_{uc} + LR_{ind}$, which is asymptotically distributed $\chi^2(2)$. The LR_{ind} statistic is the likelihood ratio statistic for the null hypothesis of serial independence against the alternative of first-order Markov dependence.

Evaluation of VaR Estimates Using Regulatory Loss Functions

The loss function evaluation method proposed here is not based on a hypothesis-testing framework, but rather on assigning to VaR estimates a numerical score that reflects specific regulatory concerns. Although this method forgoes the benefits of statistical inference, it provides a measure of relative performance that can be used to monitor the performance of VaR estimates.

To use this method, the regulatory concerns of interest must be translated into a loss function. The general form of these loss functions is

$$C_{mt+1} = \begin{cases} f(\varepsilon_{t+1}, VaR_{mt}) \text{ if } \varepsilon_{t+1} < VaR_{mt} \\ g(\varepsilon_{t+1}, VaR_{mt}) \text{ if } \varepsilon_{t+1} \geq VaR_{mt} \end{cases},$$

where $f(x,y)$ and $g(x,y)$ are functions such that $f(x,y) \geq g(x,y)$ for a given y. The numerical scores are constructed with a negative orientation, that is, lower values of C_{mt+1} are preferred since exceptions are given higher scores than nonexceptions. Numerical scores are generated for individual VaR estimates, and the score for the complete regulatory sample is

$$C_m = \sum_{i=1}^{250} C_{mt+i}.$$

Under very general conditions, accurate VaR estimates will generate the lowest possible numerical score. Once a loss function is defined and C_m is calculated, a benchmark can be constructed and used to evaluate the performance of a set of VAR estimates. Although many regulatory loss functions can be constructed, two are described below (see diagram).

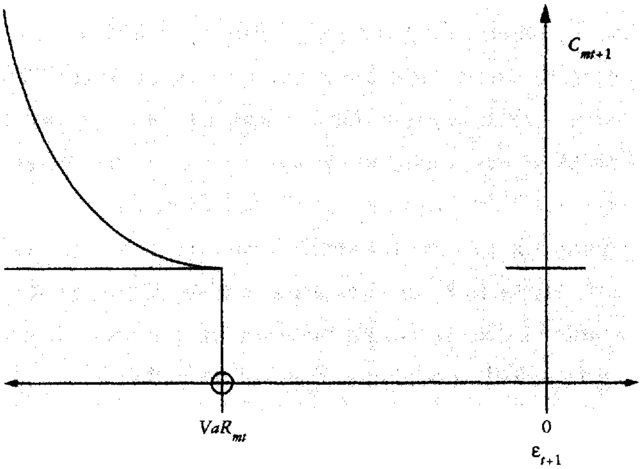
Loss Function Implied by the Binomial Method

The loss function implied by the binomial method is

$$C_{mt+1} = \begin{cases} 1 \text{ if } \varepsilon_{t+1} < VaR_{mt} \\ 0 \text{ if } \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Note that the appropriate benchmark is the expected value of C_{mt+1}, which is $E[C_{mt+1}] = 0.01$, and for the full sample, $E[C_m] = 2.5$. As before, only the number of exceptions is of interest, and the same information contained in the binomial method is included in this loss function.

Loss Functions of Interest



Notes: The diagram graphs both the binomial and the magnitude loss functions. The binomial loss function is equal to 1 for $\varepsilon_{t+1} < VaR_{mt}$ and zero otherwise. For the magnitude loss function, a quadratic term is added to the binomial loss function for $\varepsilon_{t+1} < VaR_{mt}$.

Loss Function That Addresses the Magnitude of the Exceptions

As noted by the Basle Committee on Banking Supervision (1996), the magnitude as well as the number of exceptions are a matter of regulatory concern. This concern can be readily incorporated into a loss function by introducing a magnitude term. Although several are possible, a quadratic term is used here, such that

$$C_{mt+1} = \begin{cases} 1 + (\varepsilon_{t+1} - VaR_{mt})^2 & if\ \varepsilon_{t+1} < VaR_{mt} \\ 0 & if\ \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Thus, as before, a score of one is imposed when an exception occurs, but now, an additional term based on its magnitude is included. The numerical score increases with the magnitude of the exception and can provide additional information on how the underlying VaR model forecasts the lower tail of the underlying f_{t+1} distribution. Unfortunately, the benchmark based on the expected value of C_{mt+1} cannot be determined easily, because the f_{t+1} distribution is unknown. However, a simple, operational benchmark can be constructed and is discussed in Section III.

Simulation Exercise

To analyze the ability of the three evaluation methods to gauge the accuracy of VaR estimates and thus avoid VaR model misclassification, a simulation exercise is conducted. For the two hypothesis-testing methods, this amounts to analyzing the power of the statistical tests, that is, determining the probability with which the tests reject the null hypothesis when it is incorrect. With respect to the loss function method, its ability to evaluate VaR estimates is gauged by how frequently the numerical score for VaR estimates generated from the true data-generating process (DGP) is lower than the score for VaR estimates from alternative models. If the method is capable of distinguishing between these scores, then the degree of VaR model misclassification will be low.

In the simulation exercise, the portfolio value y_{t+1} is specified as $y_{t+1} = y_t + \varepsilon_{t+1}$, where the portfolio return ε_{t+1} is generated by a GARCH(1,1)-normal process. That is, h_{t+1}, the variance of ε_{t+1}, has dynamics of the form $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. The true DGP is one of eight VaR models evaluated and is designated as the "true" model, or model 1.

The next three alternative models are homoskedastic VaR models. Model 2 is simply the standard normal distribution, and model 3 is the normal distribution with a variance of 1½. Model 4 is the *t*-distribution with six degrees of freedom, which has fatter tails than the normal distribution and an unconditional variance of 1½.

The next three models are heteroskedastic VaR models. For models 5 and 6, the underlying distribution is the normal distribution, and h_{mt+1} evolves over time as an exponentially weighted moving average of past squared returns, that is,

$$h_{mt+1} = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i\varepsilon_{t-i}^2 = \lambda h_{mt} + (1-\lambda)\varepsilon_t^2.$$

This type of VaR model, which is used in the well-known RiskMetrics calculations (see J.P. Morgan [1996]), is calibrated here by setting λ equal to 0.94 and 0.99 for models 5 and 6, respectively. Model 7 has the same variance dynamics as the true model, but instead of using the normal distribution, it uses the *t*-distribution with six degrees of freedom. Model 8 is the VaR model based on historical simulation using 500 observations, that is, using the past 500 observed returns, the α percent VaR estimate is observation number $5^*\alpha$ of the sorted returns.

In the table, panel A presents the power analysis of the hypothesis-testing methods. The simulation results indicate that the hypothesis-testing methods can have relatively low power and thus a relatively high probability of misclassifying inaccurate VaR estimates as "acceptably accurate." Specifically, the tests have low power against the calibrated normal models (models 5 and 6) since their smoothed variances are quite similar to the true GARCH variances. The power against the homoskedastic alternatives is quite low as well.

For the proposed loss function method, the simulation results indicate that the degree of model misclassification generally mirrors that of the other methods, that is, this method has a low-to-moderate ability to distinguish between the true and alternative VaR models. However, in certain cases, it provides additional useful information on

SIMULATION RESULTS FOR GARCH(1,1)-NORMAL DGP
Units: percent

	Models						
	Homoskedastic			Heteroskedastic			Historical
	2	3	4	5	6	7	8
PANEL A: POWER OF THE LR_{UC} AND LR_{CC} AGAINST ALTERNATIVE VAR MODELS[a]							
LR_{uc}	52.3	21.4	30.5	5.1	10.3	81.7	23.2
LR_{cc}	56.3	25.4	38.4	6.7	11.9	91.6	33.1
PANEL B: ACCURACY OF VAR ESTIMATES USING REGULATORY LOSS FUNCTIONS[b]							
Loss function							
Binomial	91.7	41.3	18.1	52.2	48.9	0	38.0
Magnitude	96.5	56.1	29.1	75.3	69.4	0	51.5

Notes: The results are based on 1,000 simulations. Model 1 is the true data-generating process, $\varepsilon_{t+1}|\Omega_t \sim N(0,h_{t+1})$, where $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. Models 2, 3, and 4 are the homoskedastic models $N(0, 1)$, $N(0,1.5)$, and t (6), respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1) -t (6) model with the same parameter values as model 1. Model 8 is the historical simulation model based on the previous 500 observations.

[a] The size of the tests is set at 5 percent using finite-sample critical values.

[b] Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the "true" model, that is, the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

the accuracy of the VaR estimates under the defined loss function. For example, note that the magnitude loss function is relatively more correct in classifying VaR estimates than the binomial loss function. This result is not surprising given that it incorporates the additional information on the magnitude of the exceptions into the evaluation. The ability to use such additional information, as well as the flexibility with respect to the specification of the loss function, makes a reasonable case for the use of the loss function method in the regulatory evaluation of VaR estimates.

III. IMPLEMENTATION OF THE LOSS FUNCTION METHOD

Under the current regulatory framework, regulators observe $\{\varepsilon_{t+i}, VaR_{mt+i}\}_{i=1}^{250}$ for bank m and thus can construct, under the magnitude loss function, C_m. However, for a realized value C_m*, aside from the number of exceptions, not much inference on the performance of these VaR estimates is available. It is unknown whether C_m* is a "high" or "low" number.

To create a comparative benchmark, the distribution of C_m, which is a random variable due to the random observed portfolio returns, can be constructed. Since each observation has its own distribution, additional assumptions must be imposed in order to analyze $f(C_m)$, the distribution of C_m. Specifically, the observed returns can be assumed to be independent and identically distributed (iid); that is, $\varepsilon_{t+1} \sim f$. This is quite a strong assumption, especially given the heteroskedasticity often found in financial time series. However, the small sample size of 250 mandated by the MRA allows few other choices.

Having made the assumption that the observed returns are iid, their empirical distribution $\hat{f}(\varepsilon_{t+1})$ can be estimated parametrically, that is, a specific distributional form is assumed, and the necessary parameters are estimated from the available data. For example, if the returns are assumed to be normally distributed with zero mean, the variance can be estimated such that $\hat{f}(\varepsilon_{t+1})$ is $N(0,\hat{\sigma}^2)$.

Once $\hat{f}(\varepsilon_{t+1})$ has been determined, the empirical distribution of the numerical score C_m under the distributional assumptions, denoted $\hat{f}(C_m)$, can be generated since the distribution of the observed returns and the corresponding VaR estimates are now available. For example, if $\varepsilon_{t+1} \sim N(0,\hat{\sigma}^2)$, then the corresponding VaR estimates are $VaR_{\hat{f}t} = -2.32\hat{\sigma}$. Using this information, $\hat{f}(C_m)$ can then be constructed via simulation by forming 1,000 values of the numerical score C_m, each based on 250 draws from $\hat{f}(\varepsilon_{t+1})$ and the corresponding VaR estimates.

Once $\hat{f}(C_m)$ has been generated, the empirical quantile $\hat{q}_m = \hat{F}(C_m{*})$, where $\hat{F}(C_m)$ is the cumulative distribution function of $\hat{f}(C_m)$, can be calculated for the observed value C_m*. This empirical quantile provides a performance benchmark, based on the distributional assumptions, that can be incorporated into the evaluation of the underlying VaR estimates. In order to make this benchmark operational, the regulator should select a threshold quantile above which concerns regarding the performance of the VaR estimates are raised. This decision should be based both on the regulators' preferences and the severity of the distributional assumptions used. If $\hat{q}_m$ is below the threshold that regulators believe is appropriate, say, below 80 percent, then

C_m* is "typical" under both the assumptions on $\hat{f}(\varepsilon_{t+1})$ and the regulators' preferences. If $\hat{q}_m$ is above the threshold, then C_m* can be considered atypical, and the regulators should take a closer look at the underlying VaR model.

Note that this method for evaluating VaR estimates does not replace the hypothesis-testing methods, but instead provides complementary information, especially regarding the magnitude of the exceptions. In addition, the flexibility of this method permits many other concerns to be incorporated into the analysis via the choice of the loss function.

IV. Conclusion

As implemented in the United States, the market risk amendment to the Basle Capital Accord requires that commercial banks with significant trading activity provide their regulators with VaR estimates from their own internal models. The VaR estimates will be used to determine the banks' market risk capital requirements. This development clearly indicates the importance of evaluating the accuracy of VaR estimates from a regulatory perspective.

The binomial and interval forecast evaluation methods are based on a hypothesis-testing framework and are used to test the null hypothesis that the reported VaR estimates are "acceptably accurate," where accuracy is defined by the test conducted. As shown in the simulation exercise, the power of these tests can be low against reasonable alternative VaR models. This result does not negate their usefulness, but it does indicate that the inference drawn from this analysis has limitations.

The proposed loss function method is based on assigning numerical scores to the performance of the VaR estimates under a loss function that reflects the concerns of the regulators. As shown in the simulation exercise, this method can provide additional useful information on the accuracy of the VaR estimates. Furthermore, it allows the evaluation to be tailored to specific interests that regulators may have, such as the magnitude of the observed exceptions. Since these methods provide complementary information, all three could be useful in the regulatory evaluation of VaR estimates.

References

Basle Committee on Banking Supervision. 1996. "Supervisory Framework for the Use of 'Backtesting' in Conjunction with the Internal Models Approach to Market Risk Capital Requirements." Manuscript, Bank for International Settlements.

Christoffersen, P. F. Forthcoming. "Evaluating Interval Forecasts." INTERNATIONAL ECONOMIC REVIEW.

Federal Register. 1996. "Risk-Based Capital Standards: Market Risk." Vol. 61: 47357-78.

J.P. Morgan. 1996. RISKMETRICS TECHNICAL DOCUMENT. 4th ed. New York: J.P. Morgan.

Kupiec, P. 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models." JOURNAL OF DERIVATIVES 3: 73-84.

Lopez, J. A. 1998. "Methods for Evaluating Value-at-Risk Estimates." Federal Reserve Bank of New York Research Paper no. 9802.

Commentary

Beverly Hirtle

I am very pleased to speak here today and to comment on these three very interesting and constructive papers dealing with value-at-risk modeling issues. In my view, each paper is an excellent example of what academic research has to tell practitioners and supervisors about the practical problems of constructing value-at-risk models. Each paper examines a particular aspect of value-at-risk modeling or validation, and offers important insights into the very real issues that can arise when specifying these models and when considering their use for supervisory purposes. In that sense, the papers make important contributions to our understanding of how these models are likely to work in practice.

DANIELSSON, DE VRIES, AND JØRGENSEN

The Danielsson, de Vries, and Jørgensen paper examines some key issues surrounding the question of how well current state-of-the-art, value-at-risk models capture the behavior of the tails of the distribution of profit and loss, that is, those rare but important instances in which large losses are realized. As the paper points out, this question is a fundamental one in the world of value-at-risk modeling, since both risk managers and supervisors are presumably quite concerned about such events. In fact, one of the key motivations for the development of value-at-risk models was to be able to answer the question, If something goes really wrong, how much money am I likely to lose? Put more technically, risk managers and the senior management of financial institutions wanted to be able to assess both the *probability* that large losses would occur and the *extent* of losses in the event of unfortunate movements in markets. When supervisors began considering the use of these models for risk-based capital purposes, the fundamental questions were much the same. Thus, for all these reasons, the ability to model the tails of the distribution accurately is an important concern.

As the Danielsson et al. paper shows, this ability is especially key when there is suspicion that the distribution might feature "fat tails." As you know, the phrase fat tails refers to the situation in which the *actual* probability of experiencing a loss of a given size—generally, a large loss that would be considered to have a low probability of occurring—is greater than the probability predicted by the distribution *assumed* in the value-at-risk model. Obviously, this disparity would be a matter of concern for risk managers and for supervisors who would like to use value-at-risk models for risk-based capital purposes.

Beverly Hirtle is a vice president at the Federal Reserve Bank of New York.

The paper suggests a method for addressing this situation. I will not go into the details of the analysis, but the paper proposes a method of estimating the overall distribution of potential profits and losses that essentially combines fairly standard methods for specifying the middle of the distribution with an alternative approach for estimating the tails. The paper then tests this modeling approach using random portfolios composed of U.S. equities and concludes that, at least for these portfolios, the "tail estimator" approach outperforms value-at-risk models based on a normal distribution and historical simulation.

When thinking about the practical implications of the proposed tail estimator technique, at least one significant question occurs to me. The empirical experiments reported in the paper are based on a fairly large data sample of 1,500 trading-day observations, or about six years of historical data. While this long data history may be available for certain instruments, it strikes me that these are more data than are likely to be available for at least some of the key risk factors that could influence the behavior of many financial institutions' portfolios, particularly when regime shifts and major market breaks are taken into account. Thus, the question that arises is, How well would the proposed tail estimator approach perform relative to more standard value-at-risk techniques when used on an historical data set more typical of the size used by financial institutions in their value-at-risk models, say, one to three years of data? At its heart, the question I am asking is whether the tail estimator approach would continue to perform significantly better than other value-at-risk methods under the more typical conditions facing financial institutions, both in terms of data availability and in terms of more complex portfolios. This is a question on which future research in this area might focus.

Christoffersen, Diebold, and Schuermann

The Christoffersen, Diebold, and Schuermann paper addresses another key practical issue in value-at-risk modeling, namely, whether the volatility of important financial market variables such as stock price indices and exchange rates is *forecastable*. By asking whether volatility is forecastable, the paper essentially asks whether there is value to using recently developed econometric techniques—such as some form of GARCH estimation—to try to improve the forecast of the next period's volatility, or whether it makes more sense to view volatility as being fairly constant over the long run. In technical terms, the question concerns whether *conditional* volatility estimates, which place more weight on recent financial market data, outperform *unconditional* volatility estimates, which are based on information from a fairly long historical observation period.

The answer, as the paper makes clear, is that *it depends*. Specifically, it depends on the horizon—or holding period—being examined. The results in the paper indicate that for holding periods of about ten days or more, there is little evidence that volatility is forecastable and, therefore, that more complex estimation techniques are warranted. For shorter horizons, in contrast, the paper concludes that volatility dynamics play an important role in our understanding of financial market behavior.

The basic message of the paper—that the appropriate estimation technique depends on the holding period used in the value-at-risk estimate—implies that there is no simple response to the question, What is the best way to construct value-at-risk models? The answer will clearly vary with the value-at-risk estimates' purpose.

As valuable as the contribution of the Christoffersen et al. paper is, there are some extensions that would link the work even more closely to the real world issues that supervisors and risk managers are likely to face. In particular, the analysis is based on examinations of the behavior of individual financial time series, such as equity price indices, exchange rates, and U.S. Treasury bond returns. Essentially, the analysis considers each individual financial variable as a very simple portfolio consisting of just one instrument. An interesting extension would be to see how or whether the conclusions of the analysis would change if more complex portfolios were considered. That is, would the conclusions be altered if the volatility of portfolios of multiple instruments were considered?

The results already suggest that the ability to forecast volatility is somewhat dependent on the financial variable in question—for instance, Treasury bond returns appear to have forecastable volatility for holding periods as long as twenty days, compared with about ten days for some of the other variables tested. It would be interesting, then, to build on this observation by constructing portfolios comprised of a mixture of instruments that more closely mirror the portfolio compositions that financial institutions are likely to have in practice. Such an experiment presumes, of course, that the risk manager is interested in knowing whether the volatility of the *portfolio* can be forecast, as opposed to the volatility of individual financial variables. In practice, risk managers and supervisors may be interested in knowing the answer to both questions.

LOPEZ

Finally, the paper by my colleague Jose Lopez addresses another important area in the world of value at risk: model validation. The paper explores the question, How can we assess the accuracy and performance of a value-at-risk model? To answer this question, it is first necessary to define what we mean by "accuracy." As the paper points out, there are several potential definitions. First, by accuracy, we could mean, how well does the model measure a particular percentile of the profit-and-loss distribution? This is the definition that has been incorporated into the market risk capital requirements through the so-called backtesting process. As the paper points out, approaches to assessing model accuracy along this dimension have received considerable attention from both practitioners and researchers, and the properties of the associated statistical tests have been explored in several studies.

However, the main contribution of the Lopez paper is its suggestion that alternative approaches to evaluating the performance of value-at-risk models are possible. For instance, another potential approach involves specifying a characteristic of value-at-risk models that a risk manager or a supervisor may be particularly concerned about—say, the model's ability to forecast the size of very large losses—and designing a method of evaluating the model's performance according to this criterion. Such approaches are not formal hypothesis tests, but instead involve specifying what is known as a "loss function," which captures the particular concerns of a risk manager, supervisor, or other interested party. In essence, a loss function is a shorthand method of calculating a numerical score for the performance of a value-at-risk model.

The results in the Lopez paper indicate that this loss function approach can be a useful complement to more traditional hypothesis-testing approaches. I will not go over the detail of his analysis, but the loss function approach appears to be able to provide additional information that could allow observers to separate accurate and inaccurate value-at-risk models. The important conclusion here is *not* that the loss function approach is superior to more traditional hypothesis-testing methods or that it should be used in place of these methods. Instead, the appropriate conclusion, which is spelled out in the paper, is that the loss function approach is a potentially useful supplement to these more formal statistical methods.

A further implication of the analysis is that the assessment of model performance can vary depending on who is doing the assessing and what issues or characteristics are of particular concern to the assessor. Each interested party could assess model performance using a different loss function, and the judgments made by these different parties could vary accordingly.

Before moving on to my concluding remarks, I would like to discuss briefly the material in the last section of the Lopez paper. This last section proposes a method for implementing the loss function approach under somewhat more realistic conditions than those assumed in the first section of the paper. Specifically, the last section proposes a method for calibrating the loss function in the entirely realistic case in which the "true" underlying distribution of profits and losses is unknown. Using a simulation technique, the paper demonstrates how such an approach could be used in practice, and offers some illustrations of the type of information about model accuracy that the approach could provide.

The material in this last section is a promising beginning, but before the actual usefulness of this application of the loss function approach can be assessed, it seems necessary to go beyond the relatively stylized simulation framework presented in the paper. The ideal case would be to use actual profit-and-loss data from a real financial institution's portfolio to rerun the experiments presented in the paper. Admittedly, such data are unlikely to be readily available outside financial institutions, which makes such testing difficult. However, the issue of whether the proposed loss function approach actually provides useful additional information about model performance is probably best assessed using real examples of the type of portfolio data that would be encountered if the method was actually implemented.

Concluding Remarks

In making a few brief concluding remarks about the lessons that can be drawn from these three papers, I would like to point out two themes that I see running through the papers' results. First, as discussed above, the papers highlight the point that in the world of value-at-risk modeling, there is no single correct way of doing things. The papers illustrate that the "right approach" often depends on the question that is being asked and the circumstances influencing the concerns of the questioner. The most important contribution of these papers is their helping us to understand what the "right answer" might be in certain situations, whether that situation is the presence of a fat-tailed distribution or different holding period horizons. Furthermore, the papers illustrate that in some situations, multiple approaches may be required to get a full picture of the behavior of a given portfolio or the performance of a particular model. In both senses, the three papers in this session have helped to provide concrete guidance on how to make such choices as circumstances vary.

The second theme that I see emerging from these papers is a little less direct than the issues I have just discussed. In my view, the papers reinforce the point that value-at-risk modeling—indeed probably most types of risk modeling—is a dynamic process, with important innovations and insights occurring along the way. It has been several years since I myself first started working on value-at-risk issues, as part of the original team that developed the internal models approach to market risk capital charges. Even at that stage, many financial institutions had already devoted considerable time and resources—over periods spanning several years—to the development of the models they were using for internal risk management. Despite this long history, these papers clearly indicate that serious thinking about value at risk is still very much a live issue, with innovations and new insights continuing to come about.

For that reason, no value-at-risk model can probably ever be considered complete or final; it is always a matter of keeping an eye on the most recent developments and incorporating them where appropriate. This is probably a pretty obvious observation to those of you who are involved in risk modeling on a hands-on basis. Nonetheless, it is an important observation to keep in mind as new studies emerge illustrating new shortcomings of old approaches and new approaches to old problems. These studies—such as the three presented here today—do not reflect the failure of past modeling efforts, but instead demonstrate the importance of independent academic research into the practical questions facing risk managers, supervisors, and others interested in risk modeling.