Evaluating the Forecasts of Risk Models

Jeremy Berkowitz
Federal Reserve Board

May 20, 1998
This Draft: March 16, 1999

Address correspondence to:

Jeremy Berkowitz
Trading Risk Analysis
Mail Stop 91
Federal Reserve Board
20th and C Streets, N.W.
Washington, D.C.
jberkowitz@frb.gov

Abstract: The forecast evaluation literature has traditionally focused on methods for assessing point-forecasts. However, in the context of risk models, interest centers on more than just a single point of the forecast distribution. For example, value-at-risk (VaR) models which are currently in extremely wide use form *interval forecasts*. Many other important financial calculations also involve estimates not summarized by a point-forecast. Although some techniques are currently available for assessing interval and density forecasts, none are suitable for sample sizes typically available. This paper suggests an new approach to evaluating such forecasts. It requires evaluation of the entire forecast distribution, rather than a value-at-risk quantity. The information content of forecast distributions combined with ex post loss realizations is enough to construct a powerful test even with sample sizes as small as 100.

Acknowledgements: I gratefully acknowledge helpful input from Peter Christoffersen, Michael Gordy, Matt Pritsker and Jim O'Brien. Any remaining errors and inaccuracies are mine. The opinions expressed do not necessarily represent those of the Federal Reserve Board or its staff.

1. Introduction

The forecast evaluation literature has traditionally focused on methods for assessing point-forecasts. However, in the context of risk models interest centers on more than just a single point of the forecast distribution. For example, the widely used value-at-risk (VaR) approach to quantifying portfolio risk delivers *interval forecasts*.[1] VaR is used to assess corporate risk exposures and has received the official imprimatur of central banks and other regulatory authorities. In 1997, a Market Risk Amendment to the Basle Accord permitted banks to use VaR estimates for setting bank capital requirements related to trading activity.

Many other important financial calculations involve estimates not summarized by a point-forecast. For example, the Standard Portfolio Analysis of Risk (SPAN) system, first implemented by the Chicago Mercantile Exchange in 1988, has become a very widely used approach to calculating margin requirements for customers and clearinghouse members.[2] SPAN is essentially a combination of stress tests performed on each underlying instrument in the relevant portfolio.

In practice, risk managers and financial institutions are primarily concerned with two types of model failure. The most familiar model inadequacy is that predicted values-at-risk understate (or overstate) the true amount of risk, given a particular confidence level. In other words, the model provides an inaccurate unconditional interval forecast. A second concern is that, even if the model delivers the correct *average* VaR, it may not do so at every point in time given available information. For example, in markets with persistent volatility VaR forecasts should be larger than average when volatility is above its long-run average and vice-versa.

In response, Chatfield (1993) and Christoffersen (1998) have proposed methods for evaluating interval forecasts. Unfortunately, these approaches are ill-suited to sample sizes typically available. Indeed, there is an inherent difficulty in assessing the performance of VaR

---

[1]See, Jorion (1997) for a recent survey and Risk Publications' VAR: Understanding and Applying Value-at-Risk for a compendium of research papers.

[2]In addition to the CME, the Chicago Board of Trade, the London Futures and Options exchange and at least a dozen other exchanges currently use SPAN. See Kupiec (1993) for a detailed description.

models, as stressed by Kupiec (1996).[3] Kupiec argues that in order to verify the accuracy of VaR models very large data sets are required. Intuitively, this is because users typically set the VaR level deep in the tail of the distribution (the Accord stipulates a 99% level). Violations are thus expected to occur only once every 100 days.

Data limitations are likely to be all the more relevant to credit risk models. Implementation of VaR-like credit models is an area of active research and has recently caught the attention of regulators. In 1998, the Federal Reserve System Task Force on Internal Credit Risk Models summarized the advantages and shortcomings of such an approach.[4]

Yet, at most banking institutions, internal credit risk models have not been in use long enough to generate long historical performance data. Moreover, assets in the banking book are typically not valued at nearly the frequency of trading book instruments (which are marked-to-market daily). Finally, credit risk models are generally designed for much longer horizons than trading risk models. The Federal Reserve's report indicates that risk horizons are typically one-year. As a result, there are very few *non-overlapping* observations with which to evaluate model performance.

If risk models, of any kind, are to be evaluated with such small samples, a more general perspective on model performance is required. The current practice of restricting attention to a single quantile, the VaR, must be broadened. In particular, risk models can be accurately tested by examining the sequence of *forecasted distributions* implied by the model. Kupiec (1996) is quite right that testing VaR models requires unrealistically large samples. The solution is not to abandon evaluation, but to require more from the model than just a VaR.[5]

---

[3]Lopez (1997) extends Kupiec's critique to some more recent backtesting methods.

[4]Credit Risk Models at Major U.S. Banking Institutions: Current State of the Art and Implications for Assessments of Capital Adequacy (1998), available on the world-wide web at http://www.bog.frb.fed.us/boarddocs/creditrisk/

[5]For regulatory purposes, this approach would extend reporting requirements of banks from one point on the forecast distribution to many points. However, this would not be an onerous extension of current regulations. VaR estimates are generally obtained through simulation or variance-covariance methods (assuming normality), so that many points of the distribution are already available internally.

In this way, the time series information contained in realized profits and losses is augmented by the cross-sectional information in ex ante forecasted *distributions*. The additional information content is then readily converted into testable propositions.

Some tools for evaluating distribution (or density) forecasts are described in Crnkovic and Drachman (1997) and Diebold, Gunther and Tay (1998). Unfortunately, none of these methods provide formal tests of forecast adequacy which can be applied in practice. The Crnkovic and Drachman test requires at least 1000 observations before it becomes reliable. The Diebold et al. approach is primarily qualitative (graphical) in nature and therefore not appropriate for testing.

The approach developed in the present paper extends the results of Rosenblatt (1952). Rosenblatt's transformation is used in both Crnkovic and Drachman (1996) and Diebold, Gunther and Tay (1998). Since the tests associated with the Rosenblatt-transform are inadequate in realistic sample sizes, a modified transformation and new testing framework are introduced. The proposed testing framework is then compared to existing methods in a set of simulations.

The remainder of the paper is organized as follows. Section 2 briefly discusses the existing approaches to assessing risk models. Section 3 presents a new framework for testing model output. Section 4 discusses extensions of the basic framework to finding the source of model failures. Section 5 reports the results of Monte Carlo experiments. Section 6 concludes.

2. Existing Approaches

It is useful to begin by defining the value-at-risk approach to quantifying portfolio risk. Let y_t be the change in the value of the portfolio of interest at time t. For a given horizon T and confidence level, α, the VaR is a threshold such that losses greater than VaR occur with probability 1-α. For example, a 99% two-week VaR is the quantity, $\bar{y}$, such that $pr(y_t<\bar{y})=.01$. More formally, let the probability density of y_t be $f(y_t)$, and the associated distribution function $F(y_t)=\int_{-\infty}^{y_t} f(u)du$. We can write the α level VaR as a percentile by using the inverse distribution function,

$$\bar{y}_\alpha = F^{-1}(\alpha). \tag{1}$$

Statistical evaluation of the model has traditionally proceeded by checking the number of observed violations -- the number of times losses exceed the VaR $y_t < \bar{y}$. What makes this such a daunting task is that datasets spanning a year or two yield at most a handful of violations. For example, the "backtesting" requirements set by the Basle Supervisory Committee (BSC) are based on the number of times the observed losses violate the reported bound, $\bar{y}$, in one year (250 trading days). In the 1996 report, the BSC directed that 4 observed violations be considered within normal tolerance, whereas 5 violations warranted penalties.[6] Not surprisingly, Lucas (1998) finds that such an ad hoc approach is not likely to be an optimal monitoring system.

Rather than confining attention to a very small number of violations, it is possible to transform ex ante forecast distributions into a series of independent and identically distributed (iid) random variables. Specifically, Rosenblatt (1952) defines the transformation,

$$x_t = \int_{-\infty}^{y_t} \hat{f}(u)du = \hat{F}(y_t) \tag{2}$$

where y_t is the *ex post* portfolio profit/loss realization and $\hat{f}(\cdot)$ is the *ex ante* forecasted loss density. Rosenblatt shows that x_t is iid and distributed uniformly on (0,1). Therefore, if banks are required to regularly *report forecast distributions*, $\hat{F}(\cdot)$, regulators can use the Rosenblatt-transformation and then test for violations of independence and/or of uniformity. Moreover, this result holds regardless of the underlying distribution of the portfolio returns, y_t, and even if the forecast model $\hat{F}(\cdot)$ changes over time.

A wide variety of tests would then be available both for independence and for uniformity. Crnkovic and Drachman (1997) suggest using the Kuiper statistic for uniformity which belongs to the family of statistics considered in Durlauf (1991) and Berkowitz (1998). Unfortunately, nothing appears to be gained from their approach since it requires sample sizes on the order of 1000 (as the authors themselves note). It is easy to see why this is so. The Kuiper and related

---

[6]Basle Committee on Banking Supervision (1996), "Supervisory Framework for the use of 'backtesting' in conjuction with the internal models approach to market risk capital requirements," available at http://www.bis.org

statistics are based on the distance between the observed density of x_t and the theoretical density (a straight line). The distance between two functions, $f(x)$ and $g(x)$, of course requires a large number of points. Moreover, since the Kuiper statistic is $\max_{x_1,x_2}|f(x)-g(x)|$, distance is indexed by a maximum. The maximum is a statistic not well suited to small samples.

Alternatively, Diebold, Gunther and Tay (1998) suggest the CUSUM and CUSUM-squared statistics as well as some qualitative, graphical assessment approaches. The CUSUM test is an application of the Central Limit Theorem: normalized sums of independent U(0,1) are asymptotically normal, N(.5,1/12). It is therefore straightforward to construct confidence intervals at any desired confidence level. Unfortunately, it is easy to show that the CUSUM test will have no power against some plausible alternatives. That is, the probability of rejecting the null will be zero even when the null is false.

Consider, for example, a series of portfolio returns, y_t, which are generated from any symmetric distribution. Suppose that the forecaster's model matches the true mean of y_t and is symmetric but can be in any other respects wrong. In particular, consider a risk model which fails to capture fat-tails in the data.[7] Then we have the following result.

Proposition 1 The CUSUM test has asymptotically zero power to reject the model for any fixed confidence level.

Proof

See appendix.

Intuitively, this proposition arises because the CUSUM test checks whether the Rosenblatt-transformed data is centered at .5 on the interval (0,1). Even if the model 'misses' fat-tails, the transformed data will indeed be centered on .5 (although not iid).

It is important to emphasize that such examples are not merely mathematical oddities. Financial firms and regulators are in fact very concerned with the possibility that their risk models do not adequately account for fat-tails. This failing will become evident in the results of the Monte Carlo experiments.

---

[7]A common example of such failures is if GARCH effects are present in the data but are not explicitly modeled.

Diebold, Gunther and Tay (1998) also advocate some graphical approaches to forecast evaluation. Since the transformed data should be uniformly distributed, histograms should be close to flat. Diebold, et al. demonstrate that histograms of transformed forecast data can reveal useful information about model failures. Specifically, if the model fails to capture fat-tails, the histogram of the transformed data will have peaks near zero and one.

A final approach to validating forecast intervals (and therefore VaR models) is that of Christoffersen (1998). Christoffersen notes that risk-model violations should not only occur α% of the time, but should also be uncorrelated across time.[8] Combining these properties, the variable defined as

$$\begin{aligned} I_t &= 1 \ \ \textit{if violation occurs} \\ &= 0 \ \ \textit{if no violation occurs} \end{aligned}$$

should be a Bernoulli sequence with parameter α (if the model is correct). Since violations occur so rarely (by design), testing to see whether violations form a Bernoulli requires at least several hundred observations.[9] The key problem is that Bernoulli variables take on only 2 values (0 and 1) and takes on the value 1 very rarely. We will make use of the full distribution of outcomes -- the goal is to extract as much information as possible from the available sample.

3. The likelihood-ratio testing framework

This section proposes a modification of the Rosenblatt-transformation. It will deliver as many iid N(0,1) variables *as there are observations.* As a result, it will be possible to construct powerful likelihood-ratio tests. Let $\Phi^{-1}(\cdot)$ be the inverse of the standard normal distribution function. Then we have the following proposition for any sequence of forecasts, regardless of the underlying distribution of portfolio profit and loss.

---

[8]In fact, Christoffersen notes that violations should be uncorrelated with any variable available at the time forecasts are made.

[9]This is the same weakness pointed out by Kupiec (1995) in the context of a different set of statistical tests.

Proposition 2. The statement, $x_t = \int_{-\infty}^{y_t} \hat{f}(u)du \sim \text{ iid } U(0,1)$, implies that

$z_t = \Phi^{-1}\left(\int_{-\infty}^{y_t} \hat{f}(u)du\right) \sim \text{ iid } N(0,1)$,

Proof.

See appendix.

Proposition 2 suggests a simple twist on the Rosenblatt-transformation. We can transform the observed portfolio returns to create a series, $z_t = \Phi^{-1}(\hat{F}(y_t))$, that should be iid standard normals. What makes it so useful is that, under the null, the data follows a normal distribution, rather than uniform. This allows us to bring to bear the powerful tools associated with the Gaussian likelihood. In particular, likelihood-ratio tests can now be implemented.

3a. The basic testing framework

Suppose we have generated the sequence $z_t = \Phi^{-1}(\hat{F}(y_t))$ for a given model. Since z_t should be independent across observations and standard normal, a wide variety of tests can be constructed. In particular, the null can be tested against a first-order autoregressive alternative with mean and variance possibly different than (0,1).[10] We can write,

$$z_t - \mu = \rho(z_{t-1} - \mu) + \varepsilon_t, \tag{3}$$

so that the null hypothesis described in proposition 2 is that, $\mu = 0$, $\rho = 0$, and $\text{var}(\varepsilon_t) = 1$. The exact log-likelihood function associated with equation (3) is well known and is reproduced here for convenience:

$$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log[\sigma^2/(1-\rho^2)] - \frac{(z_1 - \mu/(1-\rho))^2}{2\sigma^2/(1-\rho^2)} - \frac{T-1}{2}\log(2\pi) - \frac{T-1}{2}\log(\sigma^2) - \sum_{t=2}^{T}\left(\frac{(z_t - \mu - \rho z_{t-1})^2}{2\sigma^2}\right),$$

where σ^2 is the variance of ε_t. For brevity, we write the likelihood as a function only of the

---

[10]In the discrete-valued framework of Christoffersen (1998), the alternative hypothesis is a first-order two-state Markov chain.

unknown parameters of the model, $L(\mu,\sigma^2,\rho)$.

A likelihood-ratio test of independence across observations can be formulated as,

$$LR_{ind} = -2\left(L(\hat{\mu},\hat{\sigma}^2,0) - L(\hat{\mu},\hat{\sigma}^2,\hat{\rho})\right) \tag{4}$$

where the hats denote estimated values. This test statistic is a measure of the degree to which the data support a nonzero persistence parameter. Under the null hypothesis, the test statistic is distributed $\chi^2(1)$, chi-square with 1 degree of freedom, so that inference can be conducted in the usual way.

Of course, the null hypothesis is not just that the observations are independent but that they have mean and variance equal to (0,1). In order to jointly test these hypotheses, define the combined statistic as,

$$LR = -2\left(L(0,1,0) - L(\hat{\mu},\hat{\sigma}^2,\hat{\rho})\right). \tag{5}$$

Under the null hypothesis, the test statistic is distributed $\chi^2(3)$. Since the LR test explicitly accounts for the mean, variance and autocorrelation of the transformed data, it should have power against very general alternatives.

3b. The expected value of large losses

A type of model failure of particular interest to financial institutions and regulators is the possibility that the expected loss associated with a violation is not modelled correctly. Merely knowing that large losses will occur 1% of the time is not likely to satisfy realistic risk management needs. It is also necessary to gauge the size of such losses should they occur. Indeed, it is entirely possible for a model to correctly deliver the VaR, so that $\hat{f}(y_t<\bar{y})=\alpha$, while the expected loss due to violation, $\hat{E}(y_t|y_t<\bar{y})$, is completely inaccurate.

Existing VaR tests cannot capture failures of this nature. This is the case because traditional backtesting procedures (including those of Christoffersen (1998)) parse all losses into violations or non-violations. They are silent on the size of the violations.

The methods of this paper, however, will have power against such alternatives. Indeed, the likelihood-ratio framework described above can be extended to test virtually any departure of model from data.

To see how, it is useful to proceed in two steps. First, we show that if the true data process has fat-tails relative to the model, then the Φ^{-1}-transformed data will also be fat-tailed. Once this is established, straightforward likelihood-ratio tests for fat-tailed alternatives can be applied to the transformed data. We begin by defining fat-tailedness of the left tail.

Definition. The distribution $f(y_t)$ is fat-tailed relative to a model, $\hat{f}(y_t)$, if $f(y_t) > \hat{f}(y_t)$ for all $y_t < k$ where k is some constant.

This definition formalizes the intuitive idea that, beyond some point in the tail, a fat-tailed distribution has relatively more mass. Armed with this definition, the next step is to establish that fat tails in the profit/loss data will show up in the transformed data.

Proposition 3. If the observed data, y_t, is fat-tailed relative to a model, then the Φ^{-1}-transformed data will be fat-tailed relative to the standard normal.

Proof.

Recall that the Φ^{-1}-transformed data can be written as a compound function, $z_t = \Phi^{-1}\left(\hat{F}(y_t)\right)$, where $\hat{F}$ is the model forecast and Φ^{-1} is the inverse normal distribution. The distribution of z_t is:

$$\left|\frac{\Phi^{/}(\cdot)}{\hat{F}^{/}(\cdot)}\right| f(\cdot) \tag{6}$$

where $f(\cdot)$ is the true density of the data, y_t. Rearranging equation (6) gives, $\varphi(\cdot)\ \frac{f(\cdot)}{\hat{f}(\cdot)}$, which makes clear that if the model is correct, so that $\hat{f}(\cdot)=f(\cdot)$, then the transformed data is normally distributed. It also indicates that all model failures affect the Φ^{-1}-transformed data in a straightforward way: the normal density is distorted by the error term, $\frac{f(\cdot)}{\hat{f}(\cdot)}$. If the model fails to capture fat-tailed data, then (by definition) $\hat{f}(y_t)<f(y_t)$ for large values of y_t. This implies that $\varphi(\cdot)\ \frac{f(\cdot)}{\hat{f}(\cdot)}$ has fatter tails than a normal. *QED*.

It now remains to formulate a test of the tail and apply it to the Φ^{-1}-transformed data. We will again do this within an LR testing framework. The idea is to focus exclusively on the behavior of the data within the tail. Any observations which *do not* fall in the tail, as defined by less than the VaR, will be truncated. Therefore a test statistic can be based on the likelihood of a truncated normal. The log-likelihood in this case (see, for example, Johnson and Kotz (1970)) may be written,

$$L = -\frac{T}{2}\left(\ln 2\pi + \ln\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{y_i<VaR}(y_i-\mu)^2 + \sum_{y_i<VaR}\ln\Phi\left(\frac{VaR-\mu}{\sigma}\right). \tag{7}$$

This expression contains only observations falling in the tail but they are treated as continuous variables. The first two terms in equation (7) represent the usual Gaussian likelihood of losses. The third term is a normalization factor arising from the truncation below VaR.

Equation (7) establishes a middle-ground between traditional VaR evaluation and the full LR model described in section 3a. VaR models have traditionally been assessed by splitting the data into violations and non-violations. The number of violations are then compared to the expected number of violations -- completely ignoring the *magnitude* of the violations. The truncated normal likelihood, on the other hand, uses the actual continuously valued observations for the violations. Tests based on equation (7) should therefore be more powerful than traditional approaches while still allowing users to ignore model failures which may not be of interest -- failures which take place entirely in the *interior* of the distribution.

To construct an LR test, note that the null hypothesis again requires that $\mu = 0$, $\rho = 0$, and $\sigma^2 = 1$. Therefore we can evaluate a restricted likelihood, $L(0,0,1)$ and compare it to an unrestricted likelihood, $L(\hat{\mu},\hat{\rho},\hat{\sigma}^2)$. As before, the test statistic is based on the difference between the constrained and unconstrained values of the likelihood,

$$LR = -2\left(L(0,1,0) - L(\hat{\mu},\hat{\sigma}^2,\hat{\rho})\right). \tag{8}$$

Under the null hypothesis, the test statistic is distributed $\chi^2(3)$. This forms an LR test that the mean and variance of the violations equal those implied by the model.

4. Finding the Source of Model Failures

In practice, risk models necessarily require a variety of potentially restrictive modelling assumptions. Generally the first step is to choose a set of underlying factors, such as interest rates and exchange rates. These factors are assumed to drive changes in the prices of assets in the portfolio of interest. Distributional assumptions are then imposed on the factors in order to generate Monte Carlo VaR forecasts or density forecasts.

The next step is to posit a pricing model for assets in the portfolio that may depend in complicated ways on the factors. For example, Black-Scholes may be used to relate changes in interest rates and equities to changes in held options.

In reality, each of these modelling assumptions is done for expediency and does not represent a belief that the model is literally correct. Therefore, simply rejecting the "correctness" of a particular model is unsatisfying. Rather, it is important to distinguish among sources of model failure. Is it the pricing model or the distributional assumptions for the factors which are inaccurate?

The LR testing framework can be used to diagnose the source of model rejections. The key is to create pseudo-data which is not conditioned on any distributional assumptions. This is accomplished by using historical simulation in conjunction with any arbitrary pricing model. By comparing historical simulations to Monte Carlo data (both generated under the same pricing model), it is possible to test the distributional assumptions.

To formalize the procedure, begin by writing the true portfolio price process as $y_t = p(x_t)$, where x is some set of factors with distribution, $f(\cdot)$. We denote the density of portfolio prices $h_{p(f)}(y_t)$ to emphasize that it depends on the (true) distributions of the factors.

The risk model to be tested is $\hat{p}(\hat{f}(\cdot))$, where $\hat{p}_t$ is a pricing model and $\hat{f}(\cdot)$ is an assumed distribution. Both the pricing model and the distribution may not be correct. The full model $\hat{p}(\hat{f}(\cdot))$ implies an entire forecast distribution, $h_{\hat{p}(\hat{f})}(y_t)$. With a series of realized profit/loss data, y_t, the full model can be tested using the Φ^{-1}-transform as described in section 3.

Suppose we do so and the LR test rejects the full model. We now would like to test whether the distributional assumptions are responsible for the rejection. Consider the following procedure:

1. Generate a series of pseudo-data by historical simulation from the pricing model, $y^*_{t-i}=\hat{p}(x_{t-i})$, i=1,...,T. The actual historical realizations of the factor prices are plugged into the pricing model instead of making any distributional assumptions.

2. Transform the pseudo-data using the density forecasts of the full risk model, $z_t=\Phi^{-1}\left(\int_{-\infty}^{y_t^*} h_{\hat{p}(\hat{f}(\cdot))}(u)du\right)$. This transform differs from the test of section 3 only by using historical simulations instead of actual realized portfolio price changes. If the distributional assumptions are correct, the z_t should be distributed iid standard normal.

3. Apply the LR test (equation 5) to the transformed data, z_t. We refer to this test as LR_{dist}.

It is easy to show that when the LR_{dist} test rejects the null, the distributional assumptions are statistically false *whether or not the pricing model is accurate*. The following proposition formalizes this statement.

Proposition 4. If $\hat{f}(\cdot)=f(\cdot)$, then the transformed data, $\Phi^{-1}\left(\int_{-\infty}^{y_t^*} h_{\hat{p}(\hat{f}(\cdot))}(u)du\right)$ will be iid standard normal, regardless of whether the pricing model is correct.

Proof.
Using the Jacobian of transformation as in the proof of proposition 3, the distribution of z_t is $|\varphi(\cdot)h^{-1}_{\hat{p}(\hat{f})}(\cdot)|\ g(y^*)$, where $g(y^*)$ is the distribution of y_t^*. By construction, $y_t^*=\hat{p}(f)$. So in the event that $\hat{f}(\cdot)=f(\cdot)$, $\varphi(\cdot)h^{-1}_{\hat{p}(\hat{f})}(\cdot)=\varphi(\cdot)g^{-1}(y^*)$. Therefore, the distribution of z_t collapses to $\varphi(\cdot)$, the standard normal. *QED*.

Of course, if we have rejected the full model but the LR_{dist} does not reject, we conclude that the pricing model is the problem. One important caveat to this approach should be noted. Rejection of both the full model and the distributional assumptions tells us nothing *directly* about the pricing model. The pricing model may or may not be valid. To test it, it is first necessary to have distributional assumptions which match the data. Once that is done and the LR_{dist} no longer rejects, the full model LR test can be re-calculated to see if the pricing model is adequate.[11] The suggested testing sequence is shown schematically in Figure 1.

5. Simulation Experiments

In this section, the traditional coverage test investigated by Kupiec (1995), the Christoffersen (1998) test and the proposed LR test are compared in a set of Monte Carlo experiments. The risk models will be modeled after techniques that are commonly used by risk managers at large financial institutions for constructing VaR forecasts. At the same time, the data generating processes are necessarily kept fairly simple to allow for a computationally tractable simulation study.

The first data generating process is consistent with Black-Scholes (1973) option pricing. Sequences of equity prices are drawn from a lognormal diffusion model with a drift of 12% and diffusion coefficient of 10%. The risk free rate is set to 7% and the process is simulated over a 6 month horizon. I consider sample sizes 50, 100, 150, 200 and 250 which may be viewed as corresponding to different observation frequencies since the horizon is kept constant. The initial value of the stock is $40. For each observation, I calculate the value of a call option with strike price $44 using the Black-Scholes formula which is correct in this environment.

It is now possible to consider the performance of some common risk models in forecasting one-step ahead changes in portfolio value. Table 1 presents rejection rates of the true model and two approximate models using various backtesting techniques. In all cases, the desired confidence level of the test is fixed at .95.

The top panel of the Table, labeled size, reports the Monte Carlo rejection rates when the model coincides with the true model, Black-Scholes. The first two columns present rejection rates for the conventional unconditional coverage test. In the column labeled α=.95, the underlying risk model is a 95% VaR, while in the second column it is a 99% VaR. Rejection

[11] If the model fails to include all of the relevant factors, it will show up as a rejection of the pricing model. It will not show up in a test of the distributional assumptions.

rates in the first two columns are uniformly smaller than .05, indicating that the traditional VaR test is under-sized in small samples. This is perhaps not surprising -- with so few violations, very large samples are required to generate rejections.

Columns 3 and 4 reports coverage rates of the Christoffersen (1998) Bernoulli test. At a desired conditional coverage rate of 95%, the Christoffersen test is approximately correctly sized, although not for the 99% model.

Column 5 displays the coverage of the LR test developed in the present paper. In this environment, the LR test appears to reject a bit too frequently.

Lastly, in the two rightmost columns, the rejections rates of the LR_{tail} tests are shown for cutoff point .95 and cutoff point .99. These test statistics display approximately correct coverage.

The two lower panels show rejection rates when the model is wrong, and are therefore labeled "power". The middle panel indicates that the risk model is a Delta approximation to Black-Scholes. That is, the change in portfolio value, Δv_t, is approximated by a first order Taylor series expansion of Black-Scholes. Specifically, $\Delta v_t \approx \delta\varepsilon_t + \theta_t$ where ε_t is the innovation in the underlying stock, where δ is $\frac{\partial v_t}{\partial \varepsilon_t}|_{\varepsilon=0}$ and θ_t is $\frac{\partial v_t}{\partial t}|_{\varepsilon=0}$ (see, for example, Pritsker (1997)). To generate a risk forecast, ε_t are randomly drawn from a lognormal many times. Each of these shocks is then converted into a value, $\delta\varepsilon_t + \theta_t$, from which a distribution may be tabulated.

The lower panel show rejection rates for a second order Taylor series model. This differs from the first only by the addition of the second derivative with respect to ε_t.

With a 95% VaR, the traditional test --unconditional coverage -- and the Christoffersen (1998) Bernoulli test for conditional coverage begin to show some rejections as the sample size increases beyond about 200. However, for many realistic situations such as credit risk model forecasts, available samples will be much closer to 100. In this range, there is not even a 5% chance of rejecting a false model. The situation is of course even worse for a 99% VaR. On the other hand, the LR test would detect the fat-tails 88% of the time even with only 100 observations. Perhaps surprisingly, the LR_{tail} statistics do about as well as the full LR test.

Table 2 reports coverage rates for the LR_{dist} test statistic of distribution assumptions described in Section 4. Because all the models assume lognormality, the coverage rates should be 5% in all three panels. Nevertheless, only the Christoffersen (1998) Bernoulli test (for a 95%

VaR) and the LR tests attain correct coverage. As with the basic framework, the other tests tend to under-reject the null.

5b. Stochastic Volatility

A well documented source of fat-tailedness in financial returns is stochastic volatility -- autocorrelation in the conditional variance. Moreover, as recent turbulence in financial markets has underlined, risk models are particularly vulnerable at times of high volatility. Closed-form options pricing formulas valid in this context have recently become available (e.g., Heston (1993) and Bates (1996). Yet, they are rarely used in practice largely because of the high computational cost of evaluating the integrals that appear in such formulae.[12]

This presents a natural framework for evaluating backtesting techniques. Can models which ignore the time-variation in volatility be rejected? I generate data from Heston's (1993) mean-reverting stochastic volatility process,

$$dS(t) = \mu S dt + \sqrt{\sigma_t S}\ dz_1(t)$$
$$d\sigma(t) = (\alpha - \beta\sigma_t)dt + \eta\sqrt{\sigma_t}\ dz_2(t)$$

Following Heston, I set the volatility of volatility, η, to 10% and the long-run average of volatility is also 10%. All other parameters are left as before.

Given this data, I then generate risk forecasts for a call option with the Black-Scholes model, the delta method and the delta-gamma approximations. In addition, I consider two ad hoc models that feature modifications designed to capture the stochastic volatility. I take the delta approximation, $\delta\varepsilon_t + \theta_t$, but instead of drawing ε_t from a lognormal, it is drawn from the (correct) stochastic volatility model. This is akin to the widespread practice of plugging in time-varying estimates of volatility into Black-Scholes. The second model is an analogous modification of the delta-gamma approximation.

The results are shown in Table 3. The top panel again shows the probability of rejecting

[12]Although a single call option can typically be priced with the Heston (1993) model in a about one second, it is prohibitively slow for portfolios with thousands of options.

the model given that it is true, with a desired coverage of 5%. These coverage rates are very similar to those obtained in the Black-Scholes world of Table 1.

The next panel, labeled Black-Scholes, shows the rates at which Black-Scholes risk forecasts are rejected. All methods show increasing power as sample sizes increase. However, only 95% VaR models can be reasonably backtested by either traditional or Bernoulli tests - 99% VaR models reject 6% or 7% even with 250 observations. The LR statistic rejects the model about 46% of the time in this sample size. In this case, the LR_{tail} tests perform noticeably worse than the full LR statistic. The loss of information that results from confining attention to the tails yields lower power.

The next two panels indicate that even accounting for stochastic volatility, the delta method is rejected far more often than Black-Scholes. An unconditional test of VaR coverage rejects 20% of the time with T=250, while the LR test rejects over 90% of the time.

Interestingly, with a second order approximation (delta-gamma models), the performance of the LR tests deteriorates noticeably -- the model better fits the data -- yet the unconditional and Christoffersen (1998) tests do not.

Table 4 shows the LR_{dist} test statistics for the distributional assumptions. In this case, the models which assume lognormal innovations should be rejected while those which assume stochastic volatility should not -- for all pricing models. Several interesting results emerge. First, power against lognormality is higher when the pricing model is Black-Scholes than when the pricing model is a linear or quadratic approximation. Second, the power of backtesting techniques to reject distributional assumptions is substantially lower than the corresponding power to reject the full model. Although this is to be expected, it is perhaps surprising that lognormality is rejected at 22% rather than 94% with a delta approximation. It is possible that these results reflect that relative robustness of linear models.

Tables 5 and 6 explore the Monte Carlo evidence corresponding to the flowchart shown in Figure 1. That is, I mimic the *sequential* testing of full model and then distributional assumptions using the basic LR test and the LR_{dist} test. For example, the top panel of Table 5 shows the probability of rejecting lognormality, given that the Black-Scholes model has already been rejected by the basic framework. Again, the true data has stochastic volatility. Although both the model and the distributional assumptions are false, the sequential procedure correctly rejects the lognormal in 60% of trials

In addition, Table 5 indicates that sequential testing is correctly sized with a sample of 250 observations -- panels 2 and 4 report rejections of the stochastic volatility assumption.

Table 6 reports the results of three-step sequential tests. For example, the top panel reports the probability of rejecting stochastic volatility, given that Black-Scholes has been rejected and lognormality has been rejected. The results appear somewhat erratic, perhaps indicating that 1000 Monte Carlo trials is not sufficient to examine three-step testing.

Since Tables 1-6 fixed the confidence level of backtesting techniques at .95, it is of interest to explore whether the results are sensitive to the confidence level chosen by the risk manager. Figure 2 illustrates the tradeoff between confidence level and power for the Black-Scholes alternative in a sample size is 100. The horizontal axis is the confidence level of the test, ranging from .01 to .20. For both the violation-based approaches (the unconditional coverage test and the Bernoulli test) and the LR test, the benefit to lowering confidence levels from 95% to 80% is an increase nearly *twofold* in power.

Figures 3 and 4 display power curves for the gamma approximation and delta-gamma approximations, respectively. The plots appear qualitatively similar to Figure 1, although the values of the vertical axis are higher. Figure 4, in particular, indicates that even in sample sizes of 100, the power of the LR test can be boosted from 60% to 70% by reducing the confidence level from .95 to .9.

6. Conclusion

In recent years, there has been increasing concern among researchers, practitioners and regulators over how to evaluate risk models. Several authors have commented that only by having thousands of observations can value-at-risk models be assessed. In this paper, we follow Crnkovic and Drachman (1997) and Diebold, Gunther and Tay (1997) in emphasizing that small sample problems are exacerbated by looking only at violations of the value-at-risk. Evaluation of the entire forecast distribution, on the other hand, allows the user to extract much more information from model and data.

A new technique and set of statistical tests are suggested for comparing models to data. Through the use of a simple transformation, the forecast distribution is combined with ex post realizations to produce testable hypotheses. The testing framework is flexible and intuitive. Moreover, in a set of Monte Carlo experiments, the LR testing approach appears to deliver extremely good power properties. The probability of rejecting plausible alternatives is not only higher than existing methods, but approaches 80 to 90% in sample sizes likely to be available in realistic situations.

References

Berkowitz, J. (1998), "Generalized Spectral Estimation," Board of Governors, *Finance and Economics Discussion Series*, 1996-37.

Bates, D. S. (1996), "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options," *The Review of Financial Studies*, 9, 69-107.

Black, F. and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637-654.

Chatfield, C. (1993), "Calculating Interval Forecasts," *Journal of Business and Economics* Statistics, 11, 121-135.

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, forthcoming.

Crnkovic, C. and Drachman, J. (1997), "Quality Control," in *VAR: Understanding and Applying Value-at-Risk*. London: Risk Publications.

Courant, R. (1961). *Differential and Integral Calculus, Volume I*. New York: Interscience Publishers.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998),"Evaluating Density Forecasts," *International Economic Review*, forthcoming.

Diebold, F. X., and Lopez, J. A. (1996) "Forecast Evaluation and Combination," in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics*. Amsterdam: North-Holland.

Durlauf, S. N. (1991), "Spectral Based Testing of the Martingale Hypothesis," *Journal of Econometrics*, 50, 355-376.

Heston, S. L. (1993), "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options," *The Review of Financial Studies*, 6, 327-343.

Johnson, N. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions-1*. New York: Wiley.

Jorion, P. (1997). *Value-at-Risk: the New Benchmark for Controlling Market Risk.* Chicago: Irwin Publishing.

Kupiec, P. H. (1993), "The Performance of S&P500 Futures Product Margins under the SPAN Margining System, Board of Governors, *Finance and Economics Discussion Series*, 1993-27.

Kupiec, P. H. (1995), "Techniques for Verifying the Accuracy of Risk Measurment Models," *Journal of Derivatives,* winter, p. 73-84.

Lopez, J. A (1997), "Regulatory Evaluation of Value-at-Risk Models," Federal Reserve Bank of New York, Research Paper #9710.

Lucas, A. (1998), "Testing Backtesting: an Evaluation of the Basle Guidelines for Backtesting Internal Risk Management Models of Banks," manuscript, Free University Amsterdam.

Pritsker, M. (1997), "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time," *Journal of Financial Services Research*, 12, 201-242.

Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, 23, 470472.

Appendix

Proposition 1

Before proceeding to prove proposition 1, we will first need the following lemma.

Lemma

The ratio of two symmetric functions centered around the same mode is itself symmetric.

Proof

Define two functions, $f(\cdot)$ and $p(\cdot)$, both be symmetric about the a mode (and mean) Ex_t. Writing deviations of x_t from its mean as y, we have $f(-y)=f(y)$ and $p(-y)=p(y)$, so that $f(-y)/p(-y)=f(y)/p(y)$ which is the desired result. *QED.*

Proof of proposition 1

Consider a series of portfolio returns, y_t, which are generated from any symmetric distribution. Suppose that the forecaster's model matches the true mean of y_t and is symmetric but can be in any other respects wrong. In particular, take a risk model which fails to capture fat-tails in the data. Then the Rosenblatt-transformed data is $x_t=\int_{-\infty}^{y_t} p(u)du=P(y_t)$, where $p(\cdot)$ is the forecast density. The distribution of x_t is given by the usual Jacobian of transformation

$$g(x_t)=\left|\frac{\partial P^{-1}(x_t)}{\partial x_t}\right| f(P^{-1}(x_t)) \tag{9}$$

where $f(\cdot)$ is the density of y_t. But $\left|\frac{\partial P^{-1}(x_t)}{\partial x_t}\right| = \left|1/\frac{\partial P(x_t)}{\partial x_t}\right| =1/\frac{\partial}{\partial y_t}\int_{-\infty}^{y_t} p(u)du$, which is identically $1/p(\cdot)$ by the fundamental theorem of integral calculus (see, for example, Courant (1961), p. 111). As a result, we can simplify equation (9) to

$$g(x_t)=\frac{1}{p(\cdot)}\ f(\cdot) \tag{10}$$

From the lemma, we have that $g(x_t)$ is symmetric. Finally, note that range of $g(x_t)$ is the open set (0,1). Since the function is symmetric, it must be the case that its median and mean are exactly .5. The null of the CUSUM test is that $E\,x_t=.5$, so that it will have no power against this alternative.

Therefore, even if the model 'misses' fat-tails or important dynamics, the Rosenblatt-transformation will generate data centered at .5 on the interval (0,1) (although not iid). *QED.*

Proof of Proposition 2

First note that $g(x_t) = 1$ by definition of the uniform-(0,1). The distribution of $z_t = \Phi^{-1}(x_t)$ is therefore given by,

$$\left| \frac{\partial \Phi(z_t)}{\partial z_t} \right|.$$

Since $\Phi(\cdot)$ is a distribution it cannot take negative values. This gives,

$$\frac{\partial \Phi(z_t)}{\partial z_t} = \frac{\partial}{\partial z_t}\left(\int_{-\infty}^{z_t} \varphi(u)du \right).$$

Straightforward application of the fundamental theorem of integral calculus (see, for example, Courant (1961, p.111)), implies that

$$\frac{\partial}{\partial z_t}\left(\int_{-\infty}^{z_t} \varphi(u)du \right) = \varphi(z_t),$$

the standard normal density. *QED.*

Table 1
Alternative Backtesting Techniques
Size and Power: Data Generated under Black-Scholes

	Condit'l Coverage α=.95	Uncondit'l Coverage α=.99	Bernoulli α=.95	Bernoulli α=.99	LR	LR_{tail} α=.95	LR_{tail} α=.99
Size							
T= 50	0.018	0.004	0.032	0.004	0.068	0.068	0.106
T=100	0.014	0.002	0.044	0.012	0.066	0.040	0.060
T=150	0.008	0.000	0.030	0.010	0.110	0.050	0.048
T=200	0.026	0.000	0.058	0.006	0.094	0.044	0.040
T=250	0.022	0.000	0.052	0.006	0.096	0.040	0.036
Power: Delta Monte Carlo (lognormal)							
T= 50	0.000	0.000	0.000	0.000	0.836	0.856	0.848
T=100	0.000	0.000	0.008	0.000	0.884	0.900	0.890
T=150	0.006	0.000	0.026	0.006	0.908	0.904	0.916
T=200	0.022	0.000	0.056	0.000	0.918	0.926	0.924
T=250	0.042	0.000	0.058	0.000	0.924	0.910	0.924
Power: Delta-Gamma Monte Carlo (lognormal)							
T= 50	0.000	0.000	0.004	0.000	0.490	0.634	0.722
T=100	0.000	0.000	0.046	0.004	0.488	0.590	0.654
T=150	0.000	0.000	0.036	0.006	0.510	0.546	0.596
T=200	0.010	0.000	0.064	0.002	0.550	0.566	0.602
T=250	0.020	0.000	0.066	0.000	0.544	0.546	0.604

Notes: The Table compares the Monte Carlo performance of alternative techniques for validating forecast models over 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying geometric Brownian motion. Size indicates that the forecast model is Black-Scholes and the null hypothesis is therefore true. The panels labeled "power" display rejection rates for approximate forecast models. For the unconditional and Bernoulli VaR (interval forecast) and LR_{tail} procedures, the underlying VaR has a 95% or 99% confidence level. For the backtesting procedures, the desired confidence level is 95% (desired size is 5%).

Table 2
Alternative Backtesting of Distributional Assumptions
Size: Data Generated under Black-Scholes

	Uncondit'l Coverage α=.95	Uncondit'l Coverage α=.99	Bernoulli α=.95	Bernoulli α=.99	LR	LR_{tail} α=.95	LR_{tail} α=.99
Lrdist Black-Scholes							
T=50	0.022	0.004	0.036	0.004	0.046	0.066	0.104
T=100	0.016	0.004	0.046	0.014	0.044	0.026	0.050
T=150	0.016	0.002	0.040	0.028	0.068	0.032	0.040
T=200	0.030	0.000	0.070	0.026	0.056	0.030	0.030
T=250	0.036	0.004	0.072	0.026	0.060	0.024	0.022
LRdist Delta Monte Carlo (lognormal)							
T= 50	0.018	0.004	0.032	0.004	0.044	0.062	0.114
T=100	0.016	0.002	0.046	0.012	0.038	0.024	0.050
T=150	0.008	0.000	0.030	0.010	0.062	0.030	0.040
T=200	0.026	0.000	0.060	0.006	0.058	0.034	0.044
T=250	0.024	0.000	0.054	0.006	0.048	0.022	0.024
LRdist Delta-Gamma Monte Carlo (lognormal)							
T= 50	0.022	0.000	0.040	0.004	0.040	0.064	0.106
T=100	0.024	0.002	0.064	0.014	0.038	0.034	0.068
T=150	0.010	0.000	0.026	0.010	0.068	0.036	0.050
T=200	0.026	0.002	0.060	0.008	0.056	0.034	0.046
T=250	0.026	0.000	0.050	0.004	0.048	0.018	0.024

Notes: The Table compares the Monte Carlo performance of alternative techniques for validating forecast models over 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying geometric Brownian motion. Size indicates that the forecast model is Black-Scholes and the null hypothesis is therefore true. The panels labeled "power" display rejection rates for approximate forecast models. For the unconditional and Bernoulli VaR (interval forecast) and LR_{tail} procedures, the underlying VaR has a 95% or 99% confidence level. For the backtesting procedures, the desired confidence level is 95% (desired size is 5%).

Table 3: Alternative Backtesting Techniques
Size and Power: Data Generated under Stochastic Volatility Process

	Condit'l Coverage α=.95	Uncondit'l Coverage α=.99	Bernoulli α=.95	Bernoulli α=.99	LR	LR_{tail} α=.95	LR_{tail} α=.99
Size							
T= 50	0.014	0.002	0.034	0.002	0.078	0.072	0.118
T=100	0.006	0.000	0.044	0.012	0.068	0.050	0.062
T=150	0.006	0.000	0.036	0.008	0.114	0.066	0.072
T=200	0.018	0.002	0.054	0.004	0.090	0.046	0.042
T=250	0.032	0.002	0.062	0.012	0.108	0.060	0.048
Power: Black-Scholes							
T= 50	0.080	0.024	0.088	0.032	0.180	0.128	0.130
T=100	0.156	0.024	0.204	0.036	0.284	0.202	0.148
T=150	0.192	0.030	0.236	0.058	0.420	0.330	0.270
T=200	0.240	0.050	0.272	0.072	0.446	0.358	0.334
T=250	0.264	0.056	0.298	0.070	0.462	0.384	0.358
Power: Modified delta Monte Carlo (sotch. vol. distribution)							
T= 50	0.046	0.054	0.070	0.056	0.848	0.842	0.812
T=100	0.104	0.108	0.188	0.156	0.898	0.886	0.880
T=150	0.126	0.180	0.240	0.208	0.932	0.922	0.918
T=200	0.184	0.206	0.304	0.242	0.924	0.926	0.922
T=250	0.200	0.200	0.306	0.234	0.940	0.914	0.904
Power: delta Monte Carlo (lognormal)							
T= 50	0.050	0.066	0.080	0.070	0.828	0.828	0.824
T=100	0.120	0.126	0.204	0.164	0.892	0.878	0.866
T=150	0.160	0.192	0.276	0.228	0.928	0.924	0.914
T=200	0.204	0.234	0.312	0.264	0.918	0.922	0.920
T=250	0.246	0.224	0.344	0.258	0.946	0.940	0.936
Power: Modified delta-gamma Monte Carlo (stoch.vol. distribution)							
T= 50	0.122	0.182	0.156	0.188	0.456	0.472	0.486
T=100	0.204	0.230	0.270	0.270	0.602	0.558	0.552
T=150	0.238	0.292	0.318	0.334	0.670	0.610	0.602
T=200	0.266	0.322	0.346	0.338	0.722	0.672	0.656
T=250	0.288	0.318	0.380	0.356	0.760	0.684	0.676
Power: Delta-gamma Monte Carlo (lognormal)							
T= 50	0.122	0.180	0.152	0.188	0.462	0.470	0.494
T=100	0.202	0.240	0.274	0.276	0.614	0.594	0.604
T=150	0.244	0.310	0.330	0.354	0.712	0.646	0.632
T=200	0.290	0.334	0.372	0.354	0.742	0.680	0.676
T=250	0.320	0.342	0.420	0.376	0.772	0.722	0.714

Notes: The Table compares the Monte Carlo performance of alternative techniques for validating forecast models over 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying stochastic volatility process taken from Heston (1993) which provides closed-form option prices. The panels labeled "power" display rejection rates for approximate forecast models. For the unconditional and Bernoulli VaR (interval forecast) and LR_{tail} procedures, the underlying VaR has a 95% or 99% confidence level. For the backtesting procedures, the desired confidence level is 95% (desired size is 5%).

Table 4
Alternative Backtesting of Distributional Assumptions
Size and Power: Data Generated under Stochastic Volatility Process

	Condit'l Coverage α=.95	Uncondit'l Coverage α=.99	Bernoulli α=.99	Bernoulli α=.99	LR	LR_{tail} α=.95	LR_{tail} α=.99
Power: Black-Scholes							
T= 50	0.146	0.136	0.168	0.140	0.212	0.160	0.194
T=100	0.142	0.146	0.234	0.214	0.246	0.176	0.164
T=150	0.118	0.148	0.198	0.216	0.282	0.202	0.214
T=200	0.152	0.186	0.268	0.256	0.330	0.250	0.236
T=250	0.162	0.168	0.298	0.268	0.346	0.236	0.214
Size: Modified delta Monte Carlo							
T= 50	0.018	0.004	0.032	0.004	0.044	0.062	0.114
T=100	0.016	0.002	0.046	0.012	0.038	0.024	0.050
T=150	0.008	0.000	0.030	0.010	0.062	0.030	0.040
T=200	0.026	0.000	0.060	0.006	0.058	0.034	0.044
T=250	0.024	0.000	0.054	0.006	0.048	0.022	0.024
Power: Delta Monte Carlo (lognormal)							
T= 50	0.030	0.006	0.038	0.006	0.072	0.088	0.172
T=100	0.024	0.002	0.072	0.012	0.120	0.110	0.140
T=150	0.020	0.004	0.044	0.018	0.152	0.122	0.148
T=200	0.040	0.008	0.080	0.018	0.192	0.168	0.198
T=250	0.048	0.002	0.100	0.014	0.222	0.178	0.180
Size: Modified delta-Gamma Monte Carlo							
T= 50	0.018	0.000	0.030	0.004	0.046	0.068	0.100
T=100	0.022	0.004	0.066	0.014	0.046	0.032	0.058
T=150	0.014	0.000	0.026	0.006	0.068	0.038	0.052
T=200	0.026	0.000	0.060	0.006	0.060	0.034	0.042
T=250	0.032	0.000	0.054	0.006	0.042	0.016	0.024
Power: Delta-gamma Monte Carlo (lognormal)							
T= 50	0.020	0.000	0.024	0.004	0.074	0.096	0.158
T=100	0.022	0.002	0.072	0.014	0.114	0.110	0.134
T=150	0.018	0.004	0.040	0.010	0.148	0.122	0.148
T=200	0.040	0.004	0.082	0.020	0.182	0.158	0.188
T=250	0.052	0.000	0.096	0.014	0.200	0.162	0.176

Notes: The Table compares the Monte Carlo performance of alternative techniques for validating forecast models over 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying geometric Brownian motion. Size indicates that the forecast model is Black-Scholes and the null hypothesis is therefore true. The panels labeled "power" display rejection rates for approximate forecast models. For the unconditional and Bernoulli VaR (interval forecast) and LR_{tail} procedures, the underlying VaR has a 95% or 99% confidence level. For the backtesting procedures, the desired confidence level is 95% (desired size is 5%).

Table 5
Testing the Distributional Assumption
Stochastic Volatility Process

Pr(reject lognormal | rejected Black-Scholes)

T= 50	0.644
T=100	0.613
T=150	0.571
T=200	0.610
T=250	0.593

Pr(reject stoch. vol. | rejected lognormal delta Monte Carlo)

T= 50	0.048
T=100	0.043
T=150	0.066
T=200	0.063
T=250	0.051

Pr(reject lognormality | rejected modified delta Monte Carlo)

T= 50	0.080
T=100	0.134
T=150	0.152
T=200	0.201
T=250	0.223

Pr(reject stoch. vol. | rejected lognormal delta-gamma Monte Carlo)

T= 50	0.082
T=100	0.068
T=150	0.095
T=200	0.073
T=250	0.052

Pr(reject lognormal | rejected modified delta-gamma Monte Carlo)

T= 50	0.132
T=100	0.176
T=150	0.191
T=200	0.233
T=250	0.224

Notes: The Table compares the Monte Carlo rejection rates of the LRdist test described in the text and displayed schematically in Figure 1. Rejection rates are calculated from 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying stochastic volatility process for which the Heston (1993) model provides an analytic option price. ‘Modified’ indicates that the model assumes a stochastic volatility process rather than the lognormal. The nominal confidence level is fixed at 95%

Table 6
Testing a Second Distributional Assumption
Stochastic Volatility Process

Pr(reject stoch. vol | rejected Black-Scholes. rejected lognormality)

T= 50	0.069
T=100	0.046
T=150	0.125
T=200	0.059
T=250	0.116

Pr(reject lognormality | rejected modified delta Monte Carlo, rejected SV distribution)

T= 50	0.150
T=100	0.000
T=150	0.258
T=200	0.172
T=250	0.250

Pr(reject stoch. vol | rejected delta Monte Carlo, rejected lognormal)

T= 50	0.030
T=100	0.066
T=150	0.080
T=200	0.135
T=250	0.037

Pr(reject lognormality | rejected modified delta-gamma Monte Carlo, rejected SV distribution)

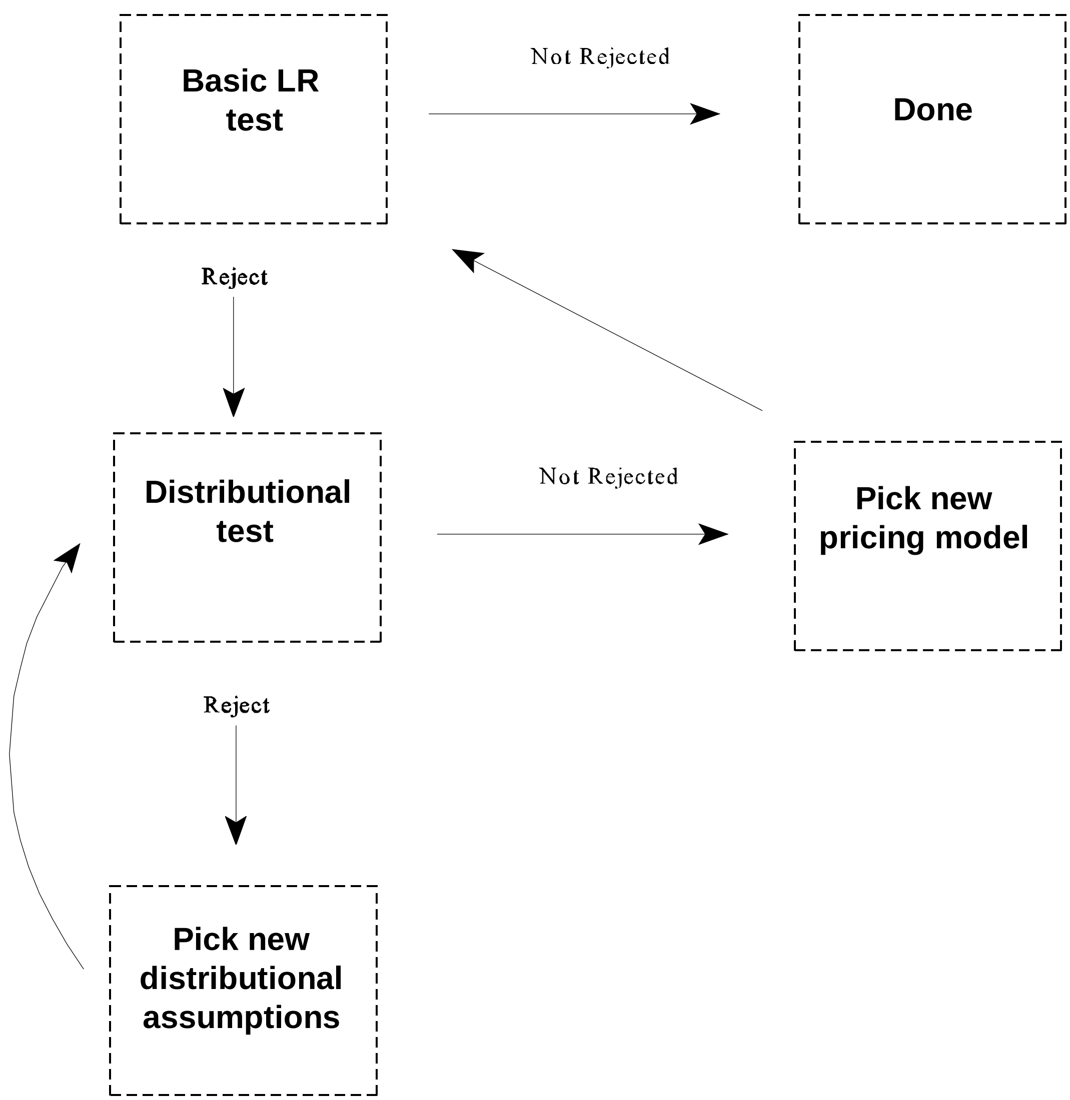
T= 50	0.000
T=100	0.091
T=150	0.177
T=200	0.250
T=250	0.286

Pr(reject stoch. vol | rejected delta-gamma Monte Carlo, rejected lognormal)

T= 50	0.030
T=100	0.018
T=150	0.054
T=200	0.046
T=250	0.053

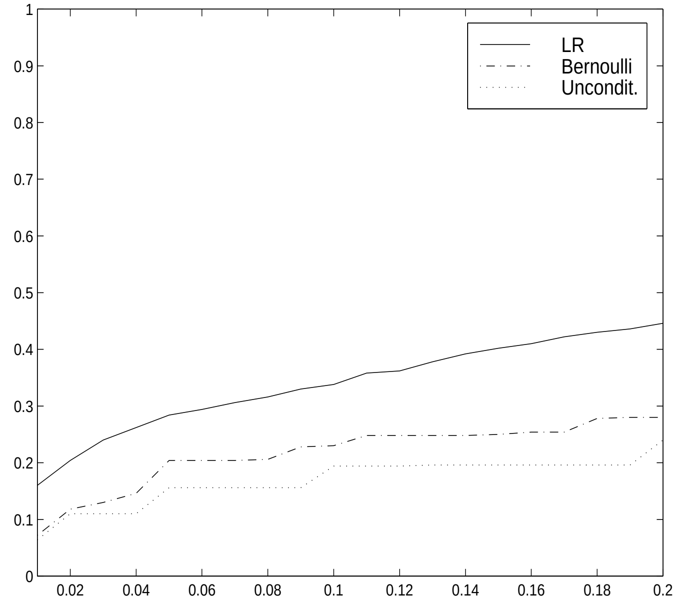
Notes: The Table compares the Monte Carlo rejection rates of the LRdist test described in the text and displayed schematically in Figure 1. Rejection rates are calculated from 1000 simulations. In each simulation, the portfolio of interest is comprised of three call options on an underlying stochastic volatility process for which the Heston (1993) model provides an analytic option price. ‘Modified’ indicates that the model assumes a stochastic volatility process rather than the lognormal. The nominal confidence level is fixed at 95%

Figure 1. Sequential Testing Procedure

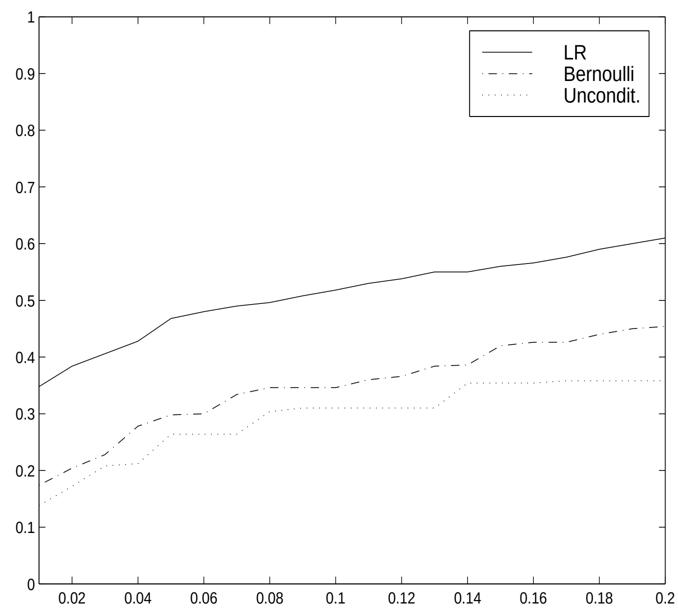


Notes to Figure: The flowchart indicates the sequence of tests suggested in the text for evaluating both a pricing model and a set of distributional assumptions. Arrows labeled “reject” indicate that the test statistic led to a rejection of the null hypothesis.

Figure 2
Alternative Backtesting Techniques
Power-Size Curves
Black-Scholes



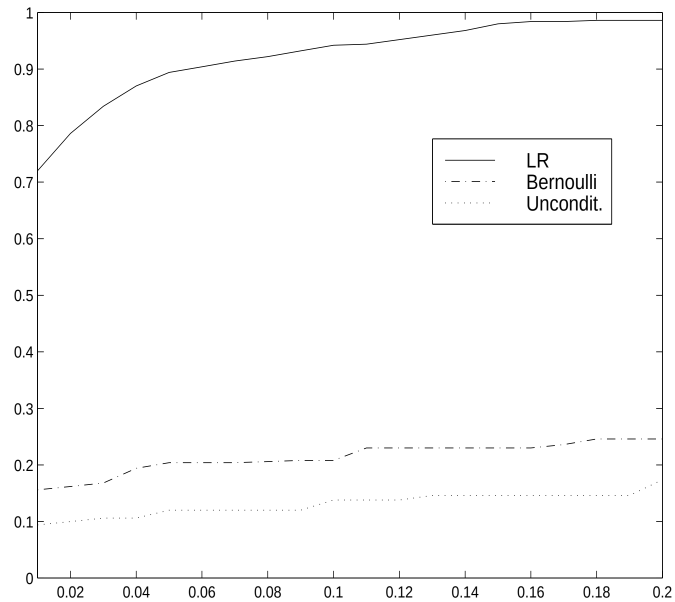
T=100



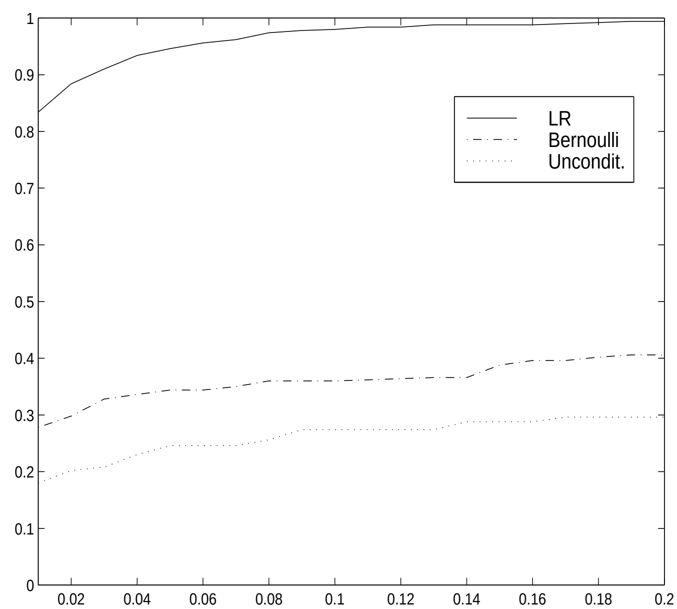
T=250

Notes: The percent of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis that the forecast distribution coincides with the true distribution. The true distribution is that of three call options on a geometric Brownian motion. The sample sizes are set to 100, top graph, and 250, lower graph.

Figure 3
Alternative Backtesting Techniques
Power-Size Curves
Delta Monte Carlo



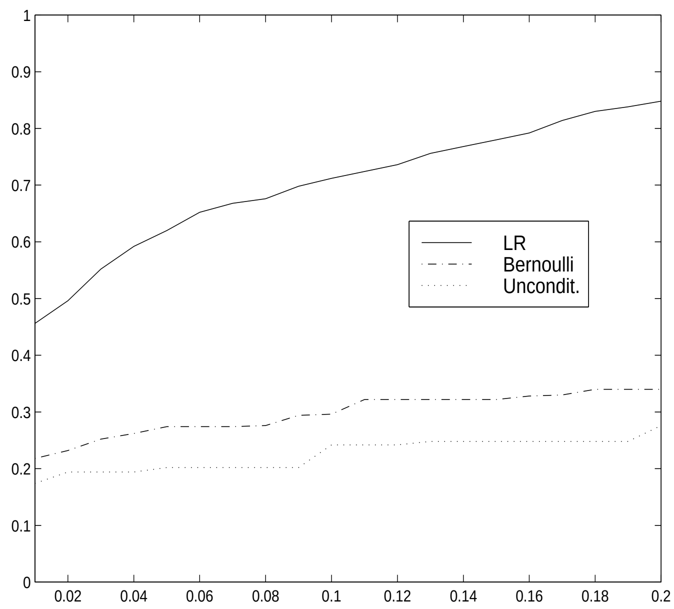
T=100



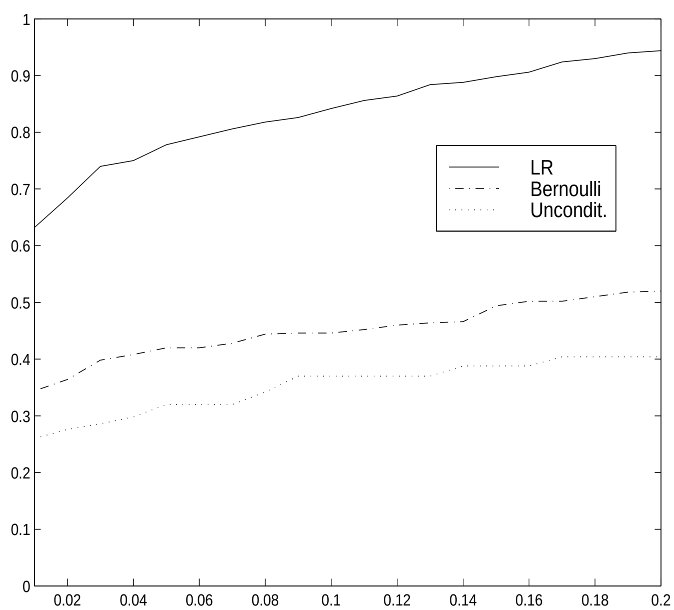
T=250

Notes: The percent of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis that the forecast distribution coincides with the true distribution. The true distribution is that of three call options on a geometric Brownian motion. The sample sizes are set to 100, top graph, and 250, lower graph.

Figure 4
Alternative Backtesting Techniques
Power-Size Curves
Delta-Gamma Monte Carlo



T=100



T=250

Notes: The percent of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis that the forecast distribution coincides with the true distribution. The true distribution is that of three call options on a geometric Brownian motion. The sample sizes are set to 100, top graph, and 250, lower graph.