

Testing Density Forecasts, With Applications to Risk Management

Jeremy BERKOWITZ

Graduate School of Management, University of California, Irvine, CA 92697-3125
(jberkowitz@gsm.uci.edu)

The forecast evaluation literature has traditionally focused on methods of assessing point forecasts. However, in the context of many models of financial risk, interest centers on more than just a single point of the forecast distribution. For example, value-at-risk models that 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, existing methods tend to display low power in sample sizes typically available. This article suggests a new approach to evaluating such forecasts. It requires evaluation of the entire forecast distribution, rather than a scalar or interval. The information content of forecast distributions combined with ex post realizations is enough to construct a powerful test even with sample sizes as small as 100.

KEY WORDS: Densities; Evaluation; Forecasting; Risk Management.

Although the forecast evaluation literature has traditionally focused on point forecasts, many models in economics and finance produce forecasts that cannot be easily summarized by a point forecast. For example, the widely used value-at-risk (VaR) approach to quantify portfolio risk delivers *interval forecasts* (see Jorion 1997 for a recent survey). These models are used to assess corporate risk exposures and have 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 single point on the forecast density. 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 calculate margin requirements for customers and clearing-house members. SPAN is essentially a combination of stress tests performed on each underlying instrument in the relevant portfolio (see Artzner, Delbaen, Eber, and Heath 1999).

To evaluate the performance of such models and their forecasts, regulators and financial institutions must be able to compare the forecasts to subsequent outcomes. The most familiar such exercise is verifying that the model accurately delivers a given interval forecast. Early approaches to this problem were handled as if the upper end of the interval was a point forecast—the number of times the interval was exceeded is compared to the expected number.

However, Christoffersen (1998) emphasized that, being interval forecasts, there is more information in interval forecasts than a point. To see why, note that even if the model delivers the correct *average* coverage, it may not do so at every point in time. For example, in markets with persistent volatility, forecasts should be larger than average when volatility is above its long-run average and vice versa. Christoffersen proposed methods for evaluating interval forecasts. These methods go part of the way toward addressing the critique of Kupiec (1995), who argued that very large datasets are required to verify the accuracy of such models. 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. Simply counting the number of violations per year obviously uses very little of the data. Christoffersen suggested making greater use of the violations by noting that violations should occur 1% of the time and should not be bunched together—violations should be *conditionally* unpredictable. Nevertheless, interval evaluation methods remain quite data intensive since they only make use of whether or not a violation occurs (e.g., Lopez 1999).

If risk models are to be evaluated with small samples, a more general perspective on model performance is required. In particular, models can be more accurately tested by examining many percentiles implied by the model. To take this argument to its logical extreme, we might evaluate the entire forecast density. Forecasts at *every* percentile would then be compared to realized data. In this way, the time series information contained in realized profits and losses is augmented by the cross-sectional information in ex ante forecast *densities*. The additional information content is readily converted into testable propositions. Density forecast evaluation, however, is affected by the density's interior. Since the interior characterizes small day-to-day disturbances, it may be of substantially less concern to financial institutions, managers, and regulators than the tail behavior.

For this reason, attention has recently shifted to “excess” measures that account for the expected magnitude of large losses. In particular, Artzner et al. (1999) suggested the tail expected loss (EL) measure, $E(y_i | y_i < \bar{y}_i)$. These authors showed that this measure of risk fulfills a set of intuitively sensible axioms such as monotonicity and subadditivity while quantiles do not. Surprisingly, Artzner et al. also showed that scenario analyses such as SPAN can be cast as equivalent to an expected loss measure. Basak and Shapiro (1998) found, in the context of a stochastic equilibrium model, that risk

management based on EL leads to lower losses than interval-based management.

From a statistical point of view, it is clear that EL measures contain more information than intervals. An exceedence implies that a violation occurred and also conveys information regarding the size of the violation. Forecast evaluation, which is based on extracting information from observed outcomes, should be more practical with exceedence measures. Moreover, EL measures allow risk managers to focus on tail events. This article is the first that I am aware of to present methods tailored to evaluate such tail forecasts.

The approach developed in this article builds on Crnkovic and Drachman (1996) and Diebold, Gunther, and Tay (1998). Diebold et al. focused on qualitative, graphical analyses rather than testing. Because their interest centered on developing tools for diagnosing *how* models fail, they did not pursue formal testing. I formalize (and extend) the Diebold et al. approach because in some applications highly specific testing guidelines are necessary. In the regulatory context, formal testing allows for uniformity across agencies, countries, and time. The Basle Accord, for this reason, specifies both a "back-testing" (forecast evaluation) procedure and a rejection region (more than four violations of a 99% quantile per year are said to indicate a bad model). At the same time, no financial model is ever literally true, and testing procedures should therefore be useful in providing guidance as to why a model is rejected. I would like to have tests that are formal yet flexible enough to handle a variety of different hypotheses regarding a model. I therefore suggest a variety of extensions to the basic testing framework.

The importance of this research area is likely to increase because credit-risk models are an area of active research that has recently caught the attention of regulators—for example, see the discussion in the Federal Reserve System's *Credit Risk Models at Major U.S. Banking Institutions: Current State of the Art and Implications for Assessments of Capital Adequacy* (1998). 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), and credit-risk models are generally designed for much longer horizons than trading-risk models. These developments accentuate the need for backtesting techniques that are suited to small samples (see Lopez and Saidenberg 2000).

The remainder of the article is organized as follows. Section 1 briefly discusses the existing approaches to assess risk models. Section 2 presents a new framework for testing model output. Section 3 reports the results of Monte Carlo experiments. Section 4 concludes.

1. EXISTING APPROACHES

I am interested in a stochastic process, y_t , which is being forecasted at time $t - 1$. Let the probability density of y_t be $f(y_t)$ and the associated distribution function $F(y_t) = \int_{-\infty}^{y_t} f(u)du$. Interval forecasts such as VaR models

are based on the inverse distribution function,

$$\bar{y}_t = F^{-1}(\alpha). \quad (1)$$

For example, a 99% two-week VaR is the quantity \bar{y} such that $\text{pr}(y_t < \bar{y}) = .01$.

One approach to validating forecast intervals is that of Christoffersen (1998). Christoffersen noted that the interval should be exceeded or violated $\alpha\%$ of the time, and such violations should also be uncorrelated across time. Combining these properties, the variable defined as

$$I_t = 1 \quad \text{if violation occurs} \\ = 0 \quad \text{if no violation occurs}$$

should be an iid 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. The key problem is that Bernoulli variables take on only two values (0 and 1) and take on the value 1 very rarely. Density evaluation methods make use of the full distribution of outcomes and thus extract a greater amount of information from the available data.

Rather than confining attention to rare violations, it is possible to transform all the realizations into a series of independent and identically distributed (iid) random variables. Specifically, Rosenblatt (1952) defined the transformation,

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

where y_t is the ex post portfolio profit/loss realization and $\hat{f}(\cdot)$ is the ex ante forecasted loss density. Rosenblatt showed 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 this probability integral 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 (1996) suggested using the Kuiper statistic that belongs to the family of statistics considered by Durlauf (1991) and Berkowitz (in press) for uniformity. Unfortunately, their approach requires sample sizes on the order of 1,000—as the authors themselves pointed out. It is easy to see why this is so. The Kuiper and related 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 $\max_x |f(x) - g(x)|$, distance is indexed by a maximum that can encounter problems in small samples. This is verified in a set of Monte Carlo simulations reported in Section 3.

Diebold et al. (1998) advocated a variety of graphical approaches to forecast evaluation. Since the transformed data should be uniformly distributed, histograms should be close to flat. Diebold et al. demonstrated 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 0 and 1.

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.

2. THE LIKELIHOOD RATIO TESTING FRAMEWORK

This section introduces an extension of the Rosenblatt transformation, which delivers, under the null hypothesis, iid $N(0,1)$ variates. This allows for estimation of the Gaussian likelihood and construction of likelihood-based test statistics that are convenient and flexible and possess good finite-sample properties.

As has been seen, it is difficult to test for uniformity with small data samples. The tests that the statistical profession has settled on are nonparametric and exploit the fact that the uniform density is a flat line. Aside from the Kuiper statistic, Diebold et al. (1998) noted that iid $U(0,1)$ behavior can also be tested via the Kolmogorov–Smirnov (KS) test or the Cramer–von Mises test. Unfortunately, nonparametric tests are notoriously data intensive—Crnkovic and Drachman's (1996) research suggests a need for at least 1,000 observations.

It is also difficult to devise parametric tests when the null hypothesis is a $U(0,1)$ random variable. Nesting the iid $U(0,1)$ model for x_t within a more general model would require letting the support depend on the unknown parameter. As is well known, in this case the likelihood ratio (LR) and other statistics do not have their usual asymptotic properties because of the discontinuity of the objective function.

I instead advocate a simple transformation to normality. First, the transformation is computationally trivial. Once the series has been transformed it is straightforward to calculate the Gaussian likelihood and construct LR, Lagrange Multiplier (LM), or Wald statistics and, for some classes of model failure, the LR test is uniformly most powerful (UMP). That is, the LR procedure has higher power than any other test for a fixed confidence level for every value of the unknown parameters. The circumstances under which this holds will be discussed in Section 2. Lastly, even when it cannot be shown to be uniformly most powerful, the LR test often has desirable statistical properties and good finite-sample behavior (e.g., Hogg and Craig 1965, pp. 258–265).

Another attractive feature of the likelihood-based testing framework is that the researcher has wide latitude in deciding which and how many restrictions to test. With small samples, however, we are likely to want a tightly parameterized test. Lastly, as will be discussed, the LR statistic can sometimes be decomposed into economically meaningful constituent LR statistics.

Although one could test the mean, variance, skewness, and so forth of iid $U(0,1)$ data, such procedures are not typically as well behaved in available sample sizes. Another interesting possibility would be to use a robustified Wald or LM test based on the quasi maximum likelihood (ML) estimates as described

by Bollerslev and Wooldridge (1992). I do not pursue that approach here. Instead, I focus on the LR test and leave investigation of LM and Wald statistics to future research.

Let $\Phi^{-1}(\cdot)$ be the inverse of the standard normal distribution function. Then I have the following result for any sequence of forecasts, regardless of the underlying distribution of portfolio returns.

Proposition 1. If the series $x_t = \int_{-\infty}^{y_t} f(u)du$ is distributed as an iid $U(0,1)$, then

$$z_t = \Phi^{-1} \left[\int_{-\infty}^{y_t} f(u)du \right] \text{ is an iid } N(0,1).$$

Proposition 1 is a well-known transformation, which is used for simulating random variates. It suggests a simple extension of the Rosenblatt transformation in which I transform the observed portfolio returns to create a series, $z_t = \Phi^{-1}(\hat{F}(y_t))$, that should be iid standard normal. What makes it so useful is that, under the null, the data follow a normal distribution. This allows me to bring to bear the convenient tools associated with the Gaussian likelihood.

In addition, I can, in some respects, treat the transformed data as if it has the same interpretation as the original untransformed data. The following proposition formalizes this notion.

Proposition 2. Let $h(z_t)$ be the density of z_t and let $\phi(z_t)$ be the standard normal. Then $\log[f(y_t)/\hat{f}(y_t)] = \log[h(z_t)/\phi(z_t)]$.

Proof. The Φ^{-1} -transformed data can be written as a compound function, $z_t = \Phi^{-1}(\hat{F}(y_t))$, where \hat{F} is the model forecast and Φ^{-1} is the inverse normal distribution. Using the Jacobian of transformation, the distribution of z_t is given by $\phi(\cdot)(f(\cdot)/\hat{f}(\cdot))$. Taking logs and rearranging, I have the claimed result.

Proposition 2 establishes that inaccuracies in the density forecast will be preserved in the transformed data. If, for example, $f > \hat{f}$ over some range, it will also be the case that $h(z_t) > \phi(z_t)$ in the corresponding region of a standard normal.

Neither the Rosenblatt transformation nor the further transformation to normality impose any distributional assumptions on the underlying data. Rather, correct density forecasts imply normality of the transformed variables. In principle, one would like to test not only the moments of and serial correlation in z_t but also for nonnormality.

The LR test, however, only has power to detect nonnormality through the first two moments of the distribution. If the first two conditional moments are correctly specified, then the likelihood function is maximized at their true values (see Bollerslev and Wooldridge 1992). One solution is to subject the standardized z_t series to a nonparametric normality test if a given LR test fails to reject the null. Alternatively, it might be possible to test normality simultaneously with other restrictions imposed by the null if the LR framework were abandoned in favor of LM tests. I do not pursue that strategy in this article.

2.1 The Basic Testing Framework

Suppose we generated the sequence $z_t = \Phi^{-1}(\widehat{F}(y_t))$ 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, for example, be tested against a first-order autoregressive alternative with mean and variance possibly different from (0,1). I can write

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

so that the null hypothesis described in Proposition 1 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, I write the likelihood as a function only of the unknown parameters of the model, $L(\mu, \sigma^2, \rho)$.

An LR test of independence across observations can be formulated as

$$\text{LR}_{\text{ind}} = -2(L(\hat{\mu}, \hat{\sigma}^2, 0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})), \quad (4)$$

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-squared with 1 df, 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). To jointly test these hypotheses, define the combined statistic as

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

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.

The alternative hypothesis can easily be generalized to higher-order dependence. I might, for example, consider the second-order autoregression,

$$z_t - \mu = \rho_1(z_{t-1} - \mu) + \rho_2(z_{t-2} - \mu) + \varepsilon_t. \quad (6)$$

The likelihood function associated with $\{\mu, \rho_1, \rho_2, \sigma^2\}$ is easily written and minimized using textbook methods.

Similarly, I can augment the set of alternatives to include nonlinear dependence,

$$z_t = \mu + \rho_1 z_{t-1} + \dots + \rho_n z_{t-n} \\ + \gamma_1 z_{t-1}^2 + \dots + \gamma_m z_{t-m}^2 + \varepsilon_t. \quad (7)$$

The null hypothesis of iid $N(0,1)$ now implies $(n+m+2)$ restrictions. Significant dynamics (persistence) in the powers of z_t are evidence of nonlinear dependence. If the sample size permits, I can extend (7) in the obvious way to include lags of z_t^3 and/or higher powers of z_t . As more restrictions are included in the LR test, it becomes increasingly close to a nonparametric test. A nonparametric test will, in principle, reject the null in the presence of *any* departure from iid normality in large samples. The cost, of course, is that nonparametric tests tend to be quite data intensive. The LR approach is attractive in that I can fine-tune the number of restrictions, or moments to be matched, to fit the relevant situation. Equation (6) could also be tested using a simple joint F test of the estimated parameters.

Lastly, Equation (6) could be augmented to include exogenous variables. Statistical significance of such variables might indicate missing factors that should be included in the underlying forecast model. For example, one might add lagged squared returns to the right side of Equation (6) as a proxy for lagged volatility as advocated by Christoffersen and Diebold (2000).

Optimality of LR Tests. When can I assert that LR tests are in some sense optimal? Recall that my maintained null hypothesis is that $z_t \sim N(0,1)$. It is known that against one-sided alternatives, such as $H_a: z_t \sim N(\mu, \sigma^2)$ $\mu < 0$, $\sigma^2 > 1$, the LR test is indeed UMP (Hogg and Craig 1965, p. 253). The one-sided region $\mu < 0$ represents left-skewed losses, while $\sigma^2 > 1$ indicates excess volatility in z_t . A UMP test against the less specific alternative $H_a: \mu \neq 0, \sigma^2 \neq 1$, does not exist. Nevertheless, to the extent that excess volatility in z_t reflects *unmodeled* volatility in the underlying portfolio returns, the one-sided alternative is likely the model failure of interest.

Similar results extend to one-sided hypothesis tests of mean, variance, and autocorrelation. However, in the case of the autocorrelation it is not obvious that either $\rho > 0$ or $\rho < 0$ necessarily corresponds to a "natural" alternative. Lastly, note that, even when it cannot be shown to be uniformly most powerful, the LR test often has desirable statistical properties.

Parameter Uncertainty. Following previous authors such as Diebold and Mariano (1995), Christoffersen (1998), and Diebold et al. (1998), I view the forecasts as primitives—my approach does not depend on the method used to produce the forecasts. The drawback, as stressed by West (1996), is that forecasts are sometimes produced by models estimated with small samples and are thus subject to parameter uncertainty. Nevertheless, in my context the cost of abstracting from parameter uncertainty may not be severe.

First, in practice many density forecasts are not based on estimated models. For example, the large-scale market risk models at many U.S. banking institutions combine calibrated parameters, estimated parameters, and ad hoc modifications that reflect the judgment of management. Another leading example, cited by Diebold et al. (1998), is the density forecasts of inflation of the Survey of Professional Forecasters.

Lastly, existing work suggests that parameter uncertainty is of second-order importance when compared to other sources of inaccurate forecasts such as model misspecification (e.g., Chatfield 1993). Diebold et al. (1998) found that the effects of parameter estimation uncertainty are inconsequential in simulation studies geared toward sample sizes relevant in finance.

2.2 Focusing on Large Losses

A type of model failure of particular interest to financial institutions and regulators is that in which the forecasted magnitude of large losses is inaccurate. Basak and Shapiro (1998) and Artzner (1999), for example, emphasized the risk measure expected loss (EL) given that a violation occurs $\hat{E}(y_i | y_i < \bar{y})$. Indeed, in many cases risk managers are *exclusively* interested in an accurate description of large losses or tail behavior. They do not want to reject a model that forecasts EL well on the basis of a failure to match the small day-to-day moves that characterize the interior of the forecast distribution.

If this is the case, the basic LR framework may not be appropriate. The problem is that the LR test will asymptotically detect any departure in the first two conditional moments of the data.

In this section, I suggest tests that allow the user to intentionally ignore model failures that are limited to the interior of the distribution. In particular, I propose LR tests based on a censored likelihood. Loosely speaking, the shape of the forecasted tail of the density is compared to the observed tail. A rejection based on the tail density is taken as a proxy for rejection of the mean of the tail, $\hat{E}(y_i | y_i < \bar{y})$. One could back-test the tail with a point forecast-style test—just as VaR was traditionally tested against the expected number of violations. For example, I could compare the forecast EL to the realized EL. But, again, this does not make full use of the information in the tail.

Consider a tail that is defined by the user. Any observations that do not fall in the tail will be intentionally truncated. Let the desired cutoff point $VaR = \Phi^{-1}(\alpha)$. For example, I might choose $VaR = -1.64$ to focus on the 5% lower tail. Then define the new variable of interest as

$$z_i^* = \begin{cases} VaR & \text{if } z_i \geq VaR \\ z_i & \text{if } z_i < VaR. \end{cases}$$

The log-likelihood function for joint estimation of μ and σ is

$$L(\mu, \sigma | z^*) = \sum_{z_i^* < VaR} \ln \frac{1}{\sigma} \phi\left(\frac{z_i^* - \mu}{\sigma}\right) + \sum_{z_i^* = VaR} \ln \left(1 - \Phi\left(\frac{VaR - \mu}{\sigma}\right)\right) \quad (8)$$

$$= \sum_{z_i^* < VaR} \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (z_i^* - \mu)^2\right) + \sum_{z_i^* = VaR} \ln \left(1 - \Phi\left(\frac{VaR - \mu}{\sigma}\right)\right). \quad (9)$$

Therefore, a test statistic can be based on the likelihood of a censored normal. This expression contains only observations falling in the tail, but they are treated as continuous variables. The first two terms in Equation (9) represent the usual Gaussian likelihood of losses. The third term is a normalization factor arising from the truncation.

Equation (9) establishes a middle ground between traditional interval forecast evaluation and the full LR model described in Section 3.1. Tests based on Equation (9) should be more powerful than traditional approaches while still allowing users to ignore model failures that may not be of

interest—failures that take place entirely in the interior of the distribution.

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

$$LR_{tail} = -2(L(0, 1) - L(\hat{\mu}, \hat{\sigma}^2)). \quad (10)$$

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

Of course, the LR test in (10) will have power to detect any mismatch in the first two moments of the tail. In particular, the LR_{tail} statistic will asymptotically reject if the tails have excessively small losses relative to the forecast, not just if they are too large.

To construct a test of expected losses, $E(z_i | z_i < VaR)$, that are greater than those of an iid $N(0, 1)$, one might instead devise a one-sided test of the null. For example, consider $H_a: z_i \sim N(\mu, \sigma^2) \mu < 0, \sigma^2 > 1$. If either $\mu < 0$ or $\sigma^2 > 1$, the density places greater probability mass in the tail region than does an iid $N(0, 1)$.

3. SIMULATION EXPERIMENTS

In this section, the traditional coverage test investigated by Kupiec (1995), the Christoffersen (1998) test, the Kuiper statistic of Crnkovic and Drachman (1996) and the related KS statistic, and the proposed LR test are compared in a set of Monte Carlo experiments. It should be kept in mind that the tests of Kupiec (1995) and Christoffersen (1998) are interval-forecast not density-forecast tests. As such, they cannot be expected to display the same power as density forecasts. Nevertheless, they are commonly used in practice and provide a useful benchmark. Similarly, the risk models are chosen to mimic techniques that are commonly used by risk managers at large financial institutions for constructing interval 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 asset prices, S_t , are drawn from a lognormal diffusion model,

$$dS_t = \mu S_t dt + \sigma S_t dz_t,$$

where the drift, μ , is set to 12% and the diffusion coefficient, σ , to 10%. The constant risk-free rate is set to 7%, and the process is simulated over a six-month horizon. I consider sample sizes 50, 100, 150, 250, and 500, which may be viewed as corresponding to different observation frequencies since the time-to-expiration is kept constant. The initial value of the stock is \$40. For each observation, I calculate the value of a call option, C_t , written on S_t with strike price \$44. Assuming no restrictions on short-selling, no transaction costs or taxes, and no arbitrage, the Black-Scholes formula correctly prices the call option in this environment.

Table 1. Alternative Backtesting Techniques (size and power: data generated under lognormal distribution)

| | Uncondit'l coverage $\alpha = .95$ | Uncondit'l coverage $\alpha = .99$ | Bernoulli $\alpha = .95$ | Bernoulli $\alpha = .99$ | Kuiper statistic | KS statistic | LR statistic | LR_{tail} $\alpha = .95$ | LR_{tail} $\alpha = .99$ |
|--|--|--|-----------------------------|-----------------------------|---------------------|-----------------|-----------------|-------------------------------|-------------------------------|
| Size | | | | | | | | | |
| $T = 50$ | .017 | .001 | .011 | .004 | .011 | .040 | .075 | .070 | .113 |
| $T = 100$ | .016 | .001 | .013 | .007 | .024 | .049 | .085 | .036 | .053 |
| $T = 150$ | .011 | .001 | .018 | .004 | .026 | .048 | .080 | .040 | .051 |
| $T = 250$ | .017 | .003 | .015 | .009 | .033 | .053 | .094 | .049 | .054 |
| $T = 500$ | .035 | .006 | .023 | .007 | .036 | .058 | .104 | .053 | .052 |
| Power: Delta Monte Carlo (lognormal) | | | | | | | | | |
| $T = 50$ | .001 | .002 | .004 | .001 | .166 | .429 | .720 | .767 | .777 |
| $T = 100$ | .007 | .001 | .016 | .000 | .298 | .585 | .795 | .806 | .805 |
| $T = 150$ | .023 | .002 | .024 | .001 | .381 | .665 | .820 | .818 | .825 |
| $T = 250$ | .093 | .004 | .081 | .003 | .463 | .723 | .837 | .836 | .838 |
| $T = 500$ | .283 | .018 | .264 | .018 | .568 | .781 | .858 | .833 | .839 |
| Power: Delta-Gamma Monte Carlo (lognormal) | | | | | | | | | |
| $T = 50$ | .001 | .002 | .011 | .001 | .343 | .310 | .330 | .547 | .643 |
| $T = 100$ | .007 | .004 | .025 | .002 | .423 | .470 | .355 | .433 | .521 |
| $T = 150$ | .038 | .004 | .043 | .003 | .461 | .538 | .365 | .417 | .496 |
| $T = 250$ | .097 | .014 | .072 | .013 | .487 | .589 | .370 | .370 | .429 |
| $T = 500$ | .200 | .049 | .164 | .049 | .561 | .682 | .430 | .374 | .409 |

NOTE: The table compares the Monte Carlo performance of alternative techniques for validating forecast models over 2,000 simulations. In each simulation, the portfolio of interest is composed of a call option 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 size-adjusted 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 size is .05. KS denotes the Kolmogorov-Smirnov test statistic.

It is now possible to consider the performance of some common risk models in forecasting one-step-ahead changes in portfolio value, ΔC_{t+1} . Table 1 presents rejection rates of the true model and two approximate models using various forecast evaluation techniques. In all cases, the desired confidence level of the test is fixed at .05.

The top panel of the table, labeled "Size," reports the Monte Carlo rejection rates when the model coincides with the true model, Black-Scholes. Although options can be priced exactly, the one-step-ahead forecast density of an option is not known analytically and is therefore tabulated by simulating 2,000 one-step-ahead option prices. The first two columns present rejection rates for the conventional unconditional coverage test. In the column labeled " $\alpha = .95$," the underlying risk model is a 95% VaR, while in the second column it is a 99% VaR. Rejection rates in the first two columns are uniformly smaller than .05, indicating that the traditional VaR test is undersized in small samples. This is perhaps not surprising—with so few violations, very large samples are required to generate rejections. When $T = 500$, rejection rates reach about .03, which suggests that several hundred observations are sufficient for approximately correct size.

Columns 3 and 4 report coverage rates of the Christoffersen (1998) Bernoulli test. The size properties of these tests are quite similar to those of the unconditional test. However, the Christoffersen test is somewhat more data intensive—it requires information on the dynamics of violations not just the number of violations. This difference probably explains the slightly lower size of the Christoffersen test. Columns 5 and 6 display the Kuiper and the KS statistics. Both statistics are approximately correctly sized in samples of 250 or larger.

Column 7 displays the coverage of the LR test developed in this article. In this environment, the LR test appears to reject a bit too frequently. A likely reason for this is the approximation error associated with simulating the forecast density of a Black-Scholes option price. Theoretically, the Black-Scholes model will always be rejected in an infinitely large sample (assuming a fixed number of simulations to approximate the density). This approximation error is unfortunately unavoidable because no closed-form solution is presently available for the density.

In the two rightmost columns, the rejection rates of the LR_{tail} tests are shown for tail levels of 95% and 99%. These test statistics display approximately correct size. I expect the LR_{tail} test to reject less frequently than the basic LR test—the LR_{tail} only uses a subset of the observations (those in the tail). This fact, combined with the numerical approximation problem mentioned previously, appears to cancel out, resulting in good size properties.

The two lower panels show rejection rates when the model is wrong and are therefore labeled "Power." To make the rejection rates comparable across statistics, the estimated rates are size adjusted: For each statistic, I estimate the Monte Carlo critical value that gives rise to .05 size under the null. Power is then calculated using this critical value.

The middle panel reports the results when the model is a delta approximation to Black-Scholes. That is, the change in portfolio value, ΔC_t , is approximated by a first-order Taylor series expansion of Black-Scholes. Specifically, $\Delta C_t \approx \delta \epsilon_t + \theta_t$, where ϵ_t is the innovation in the underlying stock, δ is $\partial C_t / \partial \epsilon_t|_{\epsilon=0}$, and θ_t is $\partial C_t / \partial t|_{\epsilon=0}$ (e.g., see Pritsker 1997). To generate a forecast, the ϵ_t are randomly drawn from a log-normal many times. Each of these shocks is then converted

Table 2. Alternative Backtesting Techniques (size and power: data generated under stochastic volatility process)

| | Uncondit'l coverage $\alpha = .95$ | Uncondit'l coverage $\alpha = .99$ | Bernoulli $\alpha = .95$ | Bernoulli $\alpha = .99$ | Kuiper statistic | KS statistic | LR statistic | LR_{tail} $\alpha = .95$ | LR_{tail} $\alpha = .99$ |
|--|--|--|-----------------------------|-----------------------------|---------------------|-----------------|-----------------|-------------------------------|-------------------------------|
| Size | | | | | | | | | |
| T = 50 | .016 | .001 | .012 | .004 | .016 | .038 | .071 | .064 | .106 |
| T = 100 | .020 | .003 | .016 | .009 | .023 | .046 | .092 | .047 | .064 |
| T = 150 | .014 | .002 | .021 | .006 | .025 | .046 | .079 | .047 | .056 |
| T = 250 | .027 | .003 | .022 | .005 | .029 | .045 | .104 | .054 | .060 |
| T = 500 | .037 | .005 | .020 | .007 | .040 | .069 | .129 | .060 | .057 |
| Power: Black-Scholes | | | | | | | | | |
| T = 50 | .157 | .033 | .097 | .028 | .086 | .064 | .133 | .094 | .096 |
| T = 100 | .191 | .056 | .173 | .053 | .184 | .104 | .233 | .169 | .152 |
| T = 150 | .267 | .063 | .210 | .063 | .255 | .113 | .293 | .220 | .201 |
| T = 250 | .365 | .108 | .308 | .108 | .352 | .164 | .394 | .328 | .303 |
| T = 500 | .475 | .182 | .464 | .182 | .544 | .300 | .574 | .526 | .506 |
| Power: Modified Delta Monte Carlo (stoch. vol. distribution) | | | | | | | | | |
| T = 50 | .081 | .091 | .090 | .080 | .321 | .464 | .762 | .759 | .739 |
| T = 100 | .149 | .183 | .211 | .170 | .496 | .588 | .833 | .817 | .804 |
| T = 150 | .201 | .232 | .244 | .223 | .595 | .693 | .853 | .841 | .833 |
| T = 250 | .308 | .282 | .334 | .278 | .644 | .740 | .887 | .866 | .860 |
| T = 500 | .444 | .369 | .459 | .369 | .749 | .822 | .909 | .883 | .877 |
| Power: Delta Monte Carlo (lognormal) | | | | | | | | | |
| T = 50 | .093 | .104 | .103 | .092 | .341 | .458 | .749 | .753 | .739 |
| T = 100 | .165 | .202 | .226 | .191 | .538 | .613 | .829 | .821 | .812 |
| T = 150 | .234 | .248 | .274 | .241 | .640 | .716 | .854 | .848 | .840 |
| T = 250 | .330 | .311 | .361 | .309 | .697 | .763 | .880 | .864 | .867 |
| T = 500 | .498 | .409 | .512 | .408 | .805 | .854 | .924 | .902 | .901 |
| Power: Modified Delta-Gamma Monte Carlo (stoch. vol. distribution) | | | | | | | | | |
| T = 50 | .165 | .192 | .164 | .185 | .417 | .379 | .392 | .402 | .436 |
| T = 100 | .239 | .306 | .285 | .303 | .563 | .524 | .505 | .456 | .468 |
| T = 150 | .309 | .349 | .326 | .347 | .642 | .607 | .574 | .529 | .530 |
| T = 250 | .389 | .388 | .391 | .388 | .695 | .669 | .647 | .587 | .584 |
| T = 500 | .485 | .481 | .485 | .481 | .776 | .768 | .733 | .683 | .674 |
| Power: Delta-Gamma Monte Carlo (lognormal) | | | | | | | | | |
| T = 50 | .170 | .194 | .169 | .188 | .452 | .370 | .392 | .419 | .454 |
| T = 100 | .259 | .313 | .306 | .309 | .619 | .529 | .522 | .487 | .504 |
| T = 150 | .338 | .365 | .351 | .363 | .703 | .648 | .600 | .549 | .560 |
| T = 250 | .420 | .411 | .430 | .411 | .753 | .701 | .687 | .642 | .644 |
| T = 500 | .544 | .517 | .547 | .517 | .837 | .805 | .781 | .741 | .740 |

NOTE: The table compares the Monte Carlo performance of alternative techniques for validating forecast models over 2,000 simulations. In each simulation, the portfolio of interest is composed of a call option on an underlying stochastic volatility process taken from Heston (1993) that provides closed-form option prices. The panels labeled "power" display size-adjusted 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 size is .05.

into a value, $\delta\epsilon_t + \theta_t$, from which a distribution may be tabulated. The lower panel shows 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 . Thus, the main difference between the middle and lower panels is that only the delta-gamma model can accommodate nonlinearity in 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 to 500. 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 .05 probability of rejecting a false model. The situation is, of course, even worse for a 99% VaR. The Kuiper statistic of Crnkovic and Drachman (1996) has power in the .20 to .40

range. The KS statistic shows substantially higher power in the .60 to .70 range. On the other hand, the LR test would detect the fat tails with probability .80 even with only 100 observations. Perhaps surprisingly, the LR_{tail} statistics do about as well as the full LR test.

Stochastic Volatility

A well-documented source of fat-tailedness in financial returns is stochastic volatility—autocorrelation in the conditional variance. Moreover, as illustrated by the financial market turbulence following the Russian default in 1998, 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; Bates 1996). Yet, they are rarely used in practice largely because

of the high computational cost of evaluating the integrals that appear in such formulas.

This presents a natural framework for evaluating backtesting techniques. Can models that 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 ε_t being drawn from a lognormal, it is drawn from the (correct) stochastic volatility model. This is akin to the widespread practice of plugging 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 2. The top panel again shows the probability of rejecting the model, given that it is true, with nominal size .05. These coverage rates are very similar to those obtained under a lognormal diffusion shown in 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 at a rate of .18 even with 500 observations.

Interestingly, the Kuiper statistic displays power nearly as high as that of the KS statistic in this context. The LR statistic, however, outperforms both the Kuiper and the KS statistics in terms of rejection rates—even on a size-adjusted basis. The LR_{tail} tests perform slightly worse than the full LR statistic. Lower power results from the loss of information that comes from confining attention to events in the tail. The next two panels indicate that even accounting for stochastic volatility, the delta method is rejected far more often than Black-Scholes. At $T = 500$ the LR test rejects the delta model modified for stochastic volatility with probability .91, as compared to a rate of only .57 for the Black-Scholes model.

These comparisons shed some interesting light on the trade-offs between modeling time-varying volatility and modeling nonlinearity. Recall that the true option price is now sensitive to stochastic volatility and is nonlinear. In terms of approximating models, the delta method is linear in the underlying but adjusts for time-varying volatility, while the Black-Scholes model is highly nonlinear but assumes constant volatility. The rejection rates in Table 2 suggest that, at least in this context, linearity is a far worse approximation in terms of forecast accuracy. Further support for the importance of nonlinearity can be gleaned by comparing the delta approximation to the second-order approximation (delta-gamma models).

The power of the LR tests deteriorates noticeably—the model better fits the data.

Since Tables 1–2 fixed the confidence level of backtesting techniques at .05, it is of interest to explore whether the results are sensitive to the confidence level chosen by the risk manager. Figure 1 illustrates the trade-off between confidence level and power for the Black-Scholes alternative in a sample size of 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 .05 to .20 is an increase nearly *twofold* in power.

Figures 2 and 3 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 3, in particular,

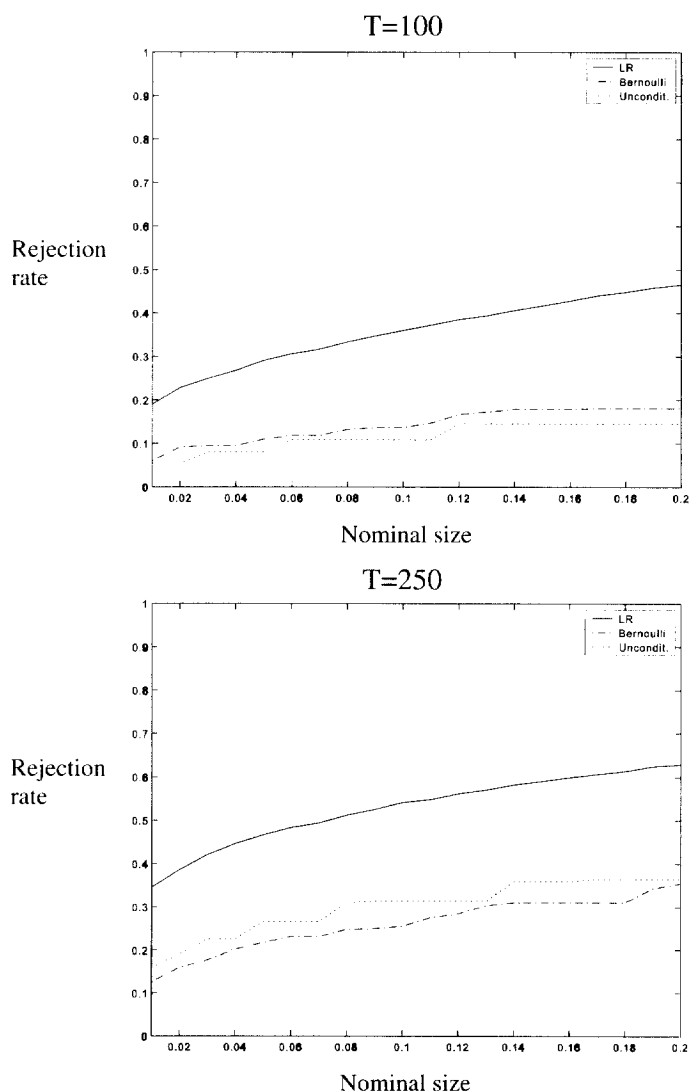


Figure 1. Power-Size Curves of Alternative Statistics in the Presence of Stochastic Volatility: Black-Scholes Approximation. Figure shows the fraction of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis of a Black-Scholes call option. The true distribution is that of a call in the presence of stochastic volatility. The sample sizes are set to 100, top graph, and 250, lower graph.

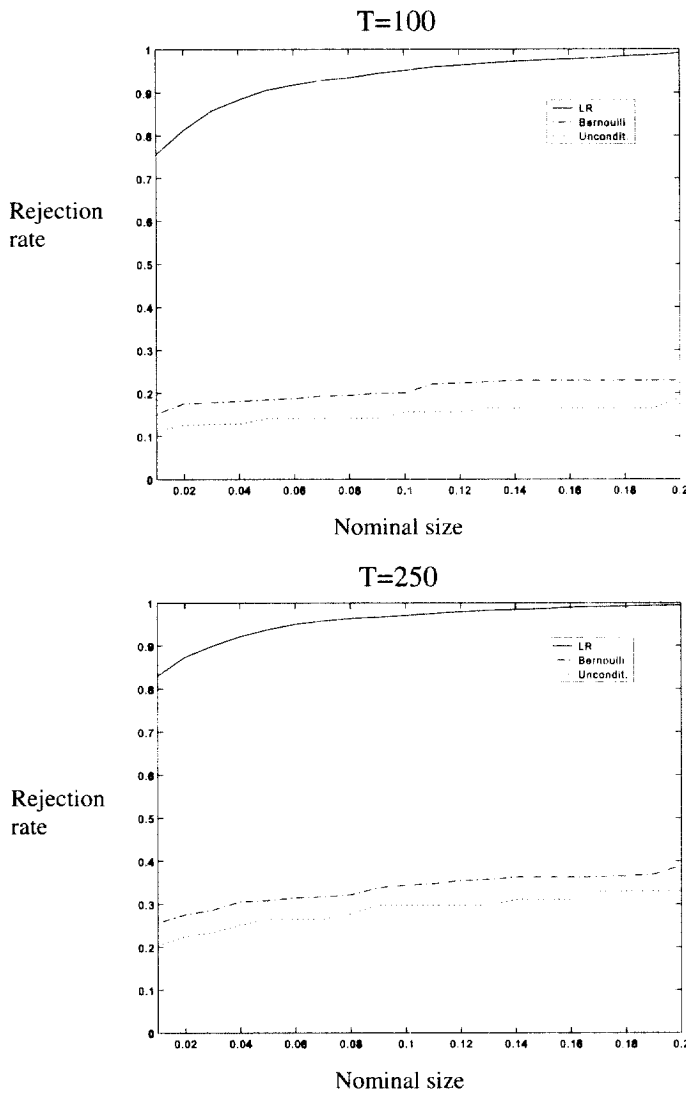


Figure 2. Power-Size Curves of Alternative Statistics in the Presence of Stochastic Volatility: Delta Approximation. Figure shows the fraction of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis of a linear approximation of Black-Scholes. The true distribution is that of a call in the presence of stochastic volatility. The sample sizes are set to 100, top graph, and 250, lower graph.

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 .05 to .10.

4. CONCLUSION

In recent years, there has been increasing concern among researchers, practitioners, and regulators over how to evaluate models of financial risk. Several authors have commented that only by having thousands of observations can interval forecasts be assessed. In this article, I follow Crnkovic and Drachman (1996) and Diebold et al. (1998) in emphasizing that small-sample problems are exacerbated by looking only at intervals. Evaluation of the entire forecast distribution, on the other hand, allows the user to extract much more information from model and data.

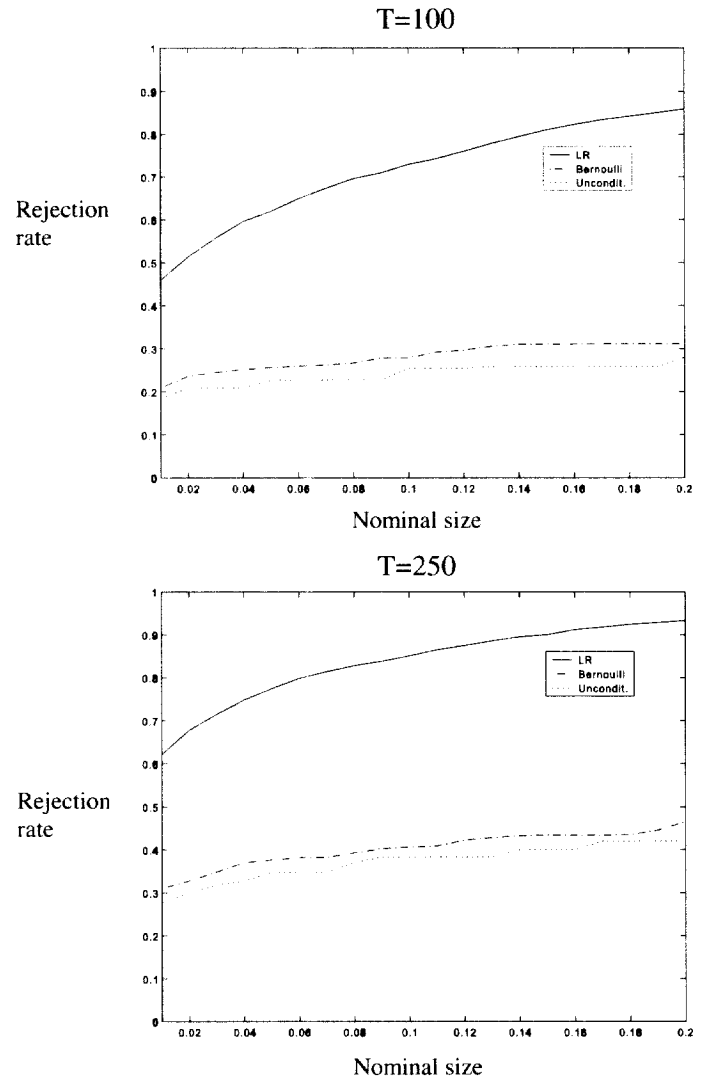


Figure 3. Power-Size Curves of Alternative Statistics in the Presence of Stochastic Volatility: Delta-Gamma Approximation. Figure shows the fraction of Monte Carlo simulations in which alternative backtesting techniques correctly rejected the null hypothesis of a delta-gamma approximation to Black-Scholes. The true distribution is that of a call in the presence of stochastic volatility. The sample sizes are set to 100, top graph, and 250, lower graph.

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.

ACKNOWLEDGMENTS

I gratefully acknowledge helpful input from Peter Christoffersen, Michael Gordy, Philippe Jorion, Matt Pritsker, and Jeffrey Wooldridge. Any remaining errors and inaccuracies are mine.

[Received January 2000. Revised January 2001.]

REFERENCES

- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999), "Coherent Measures of Risk," *Mathematical Finance*, 9, 203–228.
- Basak, S., and Shapiro, A. (1998), "Value-at-Risk Based Risk Management: Optimal Policies and Asset Prices," *Review of Financial Studies*, 14, 371–405.
- Bates, D. S. (1996), "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options," *Review of Financial Studies*, 9, 69–107.
- Berkowitz, J. (in press), "Generalized Spectral Estimation," *Journal of Econometrics*, 59.
- Black, F., and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637–654.
- Bollerslev, T., and Wooldridge, J. M. (1992), "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models With Time-varying Covariances," *Econometric Reviews*, 11, 143–172.
- Chatfield, C. (1993), "Calculating Interval Forecasts," *Journal of Business & Economic Statistics*, 11, 121–135.
- Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.
- Christoffersen, P., and Diebold, F. X. (2000), "How Relevant is Volatility Forecasting for Financial Risk Management," *Review of Economics and Statistics*, 82, 12–22.
- Crnkovic, C., and Drachman, J. (1996), "Quality Control," *Risk*, 9, 139–143.
- Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998), "Evaluating Density Forecasts," *International Economic Review*, 39, 863–883.
- Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.
- 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," *Review of Financial Studies*, 6, 327–343.
- Hogg, R. V., and Craig, A. T. (1965), *Mathematical Statistics*. New York: Macmillan.
- Jorion, P. (1997), *Value-at-Risk: The New Benchmark for Controlling Market Risk*. Chicago: Irwin.
- Kupiec, P. H. (1995), "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, winter, 73–84.
- Lopez, J. A. (1999), "Regulatory Evaluation of Value-at-Risk Models," *Journal of Risk*, 1, 37–64.
- Lopez, J. A., and Saidenberg, M. (2000), "Evaluating Credit Risk Models," *Journal of Banking and Finance*, 24, 151–165.
- 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," *The Annals of Mathematical Statistics*, 23, 470–472.
- West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.