Validating Backtests of Risk Measures

John Cotter and Yan Ping Zhong[*]

Centre for Financial Markets

University College Dublin

September 1, 2007

[*] John Cotter is at the Centre for Financial Markets, School of Business, University College Dublin, Carysfort Avenue, Blackrock, Co. Dublin, Ireland; email: john.cotter@ucd.ie. Yan Ping Zhong is the Centre for Financial Markets, School of Business, University College Dublin, Carysfort Avenue, Blackrock, Co. Dublin, Ireland; email: yan.zhong@ucd.ie. Cotter's contribution to the study has been supported by a University College Dublin School of Business research grant.

Validating Backtests of Risk Measures

Abstract

Financial risk model evaluation or backtesting is a key part of the internal model's approach to market risk management as laid out by the Basle Committee on Banking Supervision (2004). However there are a number of backtests that may be applied and there is little guidance as to the most appropriate method. The goal of this paper is to analyze the ability of various evaluation methodologies to gauge the accuracy of risk models. We compare evaluation effectiveness using the standard binomial approach, together with the interval forecast backtesting, the density forecast backtesting and the probability forecast backtesting. Our comparison is completed for three risk measures: Value-at-Risk (VaR), Expected Shortfall (ES) and Spectral Risk measure (SRM). We pay special attention to applications related to ES and SRM as backtesting of these models have not been explored in any detail thus far. Based on the Monte Carlo simulations and the empirical study, a number of interesting results emerge. Firstly within hypothesis-based tests, including the binomial backtesting, the interval forecast backtesting and the density forecasts backtesting, the overall dominance of density forecast backtesting is confirmed. In particular, the backtesting for SRM and ES is more effective than for VaR in identifying an incorrect model from alternative models in a small sample setting. Secondly, we propose a loss function for SRM where the probability forecast backtesting is capable of identifying accurate models from alternative models. Thirdly, in all of the backtesting methods examined, the choice of the distribution specification is a more important factor in determining the evaluation performance than the choice of the volatility specification.

1. Introduction

Inspired by the large increase in trading activity and several catastrophic market risk events[1], there has been a surge in the efforts of financial market participants devoted to risk management. A key part of the internal model's approach to market risk management is financial risk model evaluation or backtesting, as laid out by the Basle Committee on Banking Supervision (2004). A large body of research has emerged in the search for better approaches to evaluate risk model adequacy, such as Kupiec (1995), Christoffersen (1998), Lopez (1999a, b), and Berkowitz (2001). These studies have concentrated on developing new tests with alternative procedures.

In the literature on risk model evaluation, the standard approach to backtesting is the binomial method, and most studies have concentrated on how to achieve the optimal Value-at-Risk (VaR) model. Very few papers focus on backtesting procedures, and those that do concentrate on examining the adequacy of VaR measures (Campbell, 2005). It is not clear, however, whether general results on risk model validation are specific to the narrow evaluation method that have been widely applied in the literature (e.g., the binomial test), and whether these results hold consistently for a broad range of risk measures.

The purpose of this study is to assess the evaluation effectiveness of a range of backtesting methods to gauge the accuracy of risk models – that is, to determine whether the model chosen is accurate and performing consistently with assumptions on which the model is based. This paper extends recent research in this area by evaluating the binomial backtesting method that has been proposed by the Basle Accord, and compares and contrasts several alternative approaches, namely the interval forecast backtesting, the density forecast backtesting, and the probability forecast backtesting, with application to two alternative risk measures –ES and SRM – as well as VaR.

[1] such as the stock market crash of October 1987, crisis in Asian market of July-October 1997, the September 1998 LTCM debacle and the bursting of the high technology Dot-Com bubble of 2000-2002 with 30% losses of equity values.

The key contribution of this paper incorporates two aspects. Firstly, even though ES and SRM are coherent (and hence subadditive as well) and so satisfy many of properties we would desire a priori from a 'respectable' risk measure[2] , they are not in Basel II due to expected difficulties concerning backtesting (see Yamai and Yoshiba, 2002). This paper provides alternative evaluation procedures on backtesting ES and SRM. It extends evaluation methods that suggested by Crnkovic and Drachman (1996), Diebold *et al* (1998), Berkowitz (2001) and Lopez (1999a), and make modifications to match characteristics of ES and SRM. Secondly, this paper examines and compares evaluation effectiveness for various backtesting methods. In particular, we have determined the probability with which tests reject the specified null hypothesis when in fact it is incorrect with a range of sample sizes. The economic importance of not being able to detect an inaccurate risk model or an under-reporting institution's loss potentially become much more pronounced as the cumulative probability estimate being verified becomes smaller. As noticed by Kupiec (1995), it does not appear possible for a bank or its supervisor to verify reliably the accuracy of an institution's internal model loss exposure estimates using standard statistical technique.

The literature on backtesting evaluation methods are large and varied, and cover a range of situations. The main concerns addressed in the literature are whether VaR models are adequate and performing consistently. For example, the binomial backtesting suggested by Kupiec (1995) that is the currently standard approach in the Basle Accord, attempts to determine whether the observed frequency of violations is consistent with the frequency of expected violations according to the VaR model and chosen coverage rates. However, as noticed by Kupiec (1995), the power of detecting incorrect models is very low for this test. One problem with such test is that it ignores the dependency in violations, and exclusively focuses on the unconditional coverage property. Chatfield (1993) and Christoffersen (1998) have proposed methods for testing two fundamental hypotheses

[2] Loosely speaking, let X and Y represent any two portfolio's P/Ls (or future values, or the portfolios themselves) over a given forecast horizon, and let $\rho(.)$ be a measure of risk. The risk measure $\rho(.)$ is subadditive if it satisfies $\rho(X+Y) \le \rho(X) + \rho(Y)$. Subadditive is the most important criterion we would expect a 'respectable' risk measure to satisfy. It can be demonstrated that VaR is not subadditive unless we impose the empirically implausible requirement that returns are elliptically distributed. Given the importance of subadditivity, the VaR's non-subadditivity makes it very difficult to regard the VaR as a 'respectable' measure of risk.

concerning the process of VaR violations for a given coverage rate: the hypothesis of unconditional coverage and the hypothesis of independence [3]. A more recent independence test that has been suggested by Christoffersen and Pelletier(2004) uses the insight that if VaR violations are completely independent from each other, then the amount of time that elapses between VaR violations should be independent of the amount of time that has elapsed since the last violation. In this sense, the time between VaR violations should not exhibit any kind of 'duration dependence'. Unfortunately, as noticed by Christoffersen (1998) and Berkowitz (2001), these approaches are ill-suited to sample sizes typically available, such as 250 observations.

Crnkovic and Drachman (1997), Diebold, Gunther and Tay (1998) as well as Berkowitz (2001) have suggested the density forecast backtesting based on multiple VaR levels. The basic idea is that the unconditional coverage and independence properties of an accurate VaR measure should hold for any level of coverage rates. Berkowitz (2001) find that this evaluation method is capable of identifying incorrect risk models in a very small sample setting (e.g., such as 100 observations). Lopez (1999a, b) proposes an alternative evaluation method – the probability forecast backtesting, based on regulatory loss functions. He shows that the loss function for VaR is able of differentiating accurate model from alternative models.

There are two general approaches in risk forecasting – either conditional on current market conditions or on the unconditional market environment. Both approaches have advantages and disadvantages. Thus, the choice of methodology is situation dependent. For example, a pension fund manager has an average time horizon that is quite different from that of an options trader. Furthermore, financial returns data have at least two stylized facts: fat tails and volatility dependence. It is well known that volatility clustering is absent from a monthly return series, however the fat tail property does not fade. For longer time horizons, an unconditional model is appropriate for the calculation

[3] This approach is sometimes called Event Probability Forecast Approach (Clements and Taylor, 2003).

of large loss forecasts[4]. In many situations where the investment horizon is short, conditional volatility models may be preferable for risk forecasting. This paper therefore carries out two settings in the Monte Carlo simulation – unconditional and conditional setting, and supports the findings via an empirical study.

Furthermore, the impact of distributional assumptions and dynamic volatility estimations on the validity of the backtest methodologies is studied in the paper. Deciding on which distribution and/or volatility specification to use for a particular asset is a common task for finance practitioners and risk professionals. For instance, in spite of the massive literature on volatility forecasting, a clear consensus on which model to use has not yet been reached. As argued in Poon and Granger (2003), most of the volatility forecasting studies do not produce very conclusive results because only a subset of alternative models are compared, with a potential bias towards the method developed by the authors. It is further claimed that lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task.

Based on simulation and empirical results, our findings can be summarized as follows. Firstly, within hypothesis-based tests - the binomial backtesting and the interval forecast backtesting, and the density forecasts backtesting, the overall dominance of the density forecast backtesting is confirmed. The binomial backtesting and the interval forecast backtesting cannot detect a model failure with a small sample size, such as 250 observations, as recommended by the Basel II. Therefore, the likelihood of misclassifying incorrect risk models as correct is lower for the density forecast backtesting than for frequency-based backtests. Furthermore, the backtesting on SRM and ES is more effective than that of VaR in identifying an incorrect model from alternative models in a small sample setting.

[4] However, even if the time horizon is shorter, financial institutions often prefer unconditional risk forecast methods to avoid undesirable frequent changes in risk limits for traders and portfolio managers. (Danielsson, 2000)

Secondly, we propose a loss function for SRM where the probability forecast backtesting is capable of identifying accurate models from alternative models - that is, the quadratic probability score for the true model is lower than that of alternative models, even with sample sizes as small as 125 observations.

Thirdly, in all backtesting methods examined, the choice of distribution specifications is a more important factor in determining the evaluation performance than the choice of volatility specifications. That is, all four methods are capable of differentiating between the true model and alternative models with the same variance dynamics but different distributional assumptions.

The remainder of the paper is organized as follows. Section 2 reviews risk measures to be examined. Section 3 describes various evaluation methodologies and extensions on ES and SRM. Section 4 reports and discusses simulation results within conditional and unconditional settings. Section 5 presents empirical results on the three most heavily traded futures contracts. Section 6 provides a summary and conclusion.

2. Measures of Risk

2.1. Value-at-Risk (VaR)

Since its introduction in the 1996 amendment to the Basel Accord (see Basel Committee on Banking Supervision (1996a) and Basel Committee on Banking Supervision (1996b)), VaR has become the standard risk measurement among regulatory and financial institutions due to the ease with which it can be computed and implemented (see Jorion, 1997; Dowd, 1998; Duffie and Pan, 1997). It is an estimator of maximum loss over a target horizon for a predefined probability level.

Assume a distribution function $F_t(y) \equiv \Pr(y_t \le y | I_{t-1})$ represents different distributions[5], VaR is the $p = 1 - \alpha$ quantile of *P/L* distribution *F*, where α is the coverage rate, such as a 95% and a 99%.

$$VaR_\alpha = q_p = F^{-1}(p) \tag{1}$$

It is worth noting that a VaR places all its weight on a single quantile that corresponds to a chosen coverage rate, and it places no weight on any other.

2.2. Coherent Risk Measure

In recent research papers, VaR has been heavily criticized as a risk measure on the grounds that it does not satisfy the properties of coherence and, most particularly, that it is not subadditive (Artzner *et al.,* 1997, 1999; Acerbi, 2004 and Tasche, 2002). Subadditivity implies that a portfolio risk should reflect diversification benefits. The failure of VaR to be subadditive can lead to undesirable outcomes, especially in a portfolio context[6].

A risk measure $\rho(\cdot)$ is coherent if it satisfies the following conditions: (for details see Artzner *et al*., 1997, 1999)

$$\begin{array}{ll} \rho(mX) = m\rho(X) & \text{(Homogeneity)} \\ \rho(X) \ge \rho(Y), \quad \text{if } X \le Y & \text{(Monotonicity)} \\ \rho(X+Y) \le \rho(X) + \rho(Y) & \text{(Subadditivity)} \\ \rho(X+a) = \rho(X) - a & \text{(Translation invariant)} \end{array}$$

2.2.1. Expected Shortfall (ES)

ES is the average of worst $(1-\alpha)$ losses. Unlike VaR, ES is coherent and takes account of magnitude of losses exceeding VaR. As stated by Szegö (2002), the severity of a loss

[5] The distribution *F* can be estimated either unconditionally or conditionally.

[6] One of problems with VaR is non-subadditive is that if regulators use non-subadditive risk measures to set capital requirements, then a financial firm might be tempted to break itself up to reduce its regulatory capital requirements, because the sum of capital requirements of smaller units would be less than the capital requirement of the firm as a whole. See Dowd (2004).

is critical in risk management because a single catastrophic loss could put a firm into bankruptcy.

In the case of a continuous loss function, the ES is given by

$$ES_{\alpha} = E\{y_t \big| y_t > q_p(y_t)\}$$

$$ES_{\alpha} = (1-\alpha)^{-1}\int_{\alpha}^{1} q_p dp \qquad (2)$$

Using an ES measure implies taking an average of quantiles in which tail quantiles have an equal weight and non-tail quantiles have a zero weight.

2.2.2. Spectral Risk Measure (SRM)

Recently Acerbi (2002, 2004) proposed SRM as a risk measure that directly relates to the user's risk spectrum or risk-aversion function. 'Well-behaved' spectral risk measures are a subset of the family of coherent risk measures, and therefore have the attraction of coherency. SRM_{ϕ} is defined as a weighted average of quantiles of a loss distribution. If p is a probability level and q_p is the p-quantile of a loss distribution, so q_p is the loss such that the probability of a loss, less than or equal to it is p. Then the spectral risk measure is

$$SRM_{\phi} = \int_{0}^{1} \phi(p) q_p dp \qquad (3)$$

Where the weighting function, $\phi(p)$ is also known as the risk spectrum or risk-aversion function as long as it satisfies the following properties: a) non-negative; b) normalization; 3) increasingness (for detail, see Cotter and Dowd, 2006)[7].

3. Evaluation Methodologies

[7] The reasonable example is an exponential risk-aversion function: $\phi(p) = \dfrac{e^{-(1-p)/\gamma}}{\gamma(1-e^{-1/\gamma})}$ where $\gamma \in (0, \infty)$ (see Acerbi, 2004). This function satisfies the conditions required of a SRM, and is also attractive because it is a simple function that depends on a single parameter, the value of which reflects the risk aversion of the user.

Currently regulators have three available hypothesis-testing methods for evaluating the accuracy of VaR risk models[8]: the binomial backtesting, the interval forecast backtesting and the density forecast backtesting. A key issue with any hypothesis-based tests is its power – that is, its ability to reject the risk model when it is incorrect. The main objective of this paper is to evaluate the performance of various backtesting methods to gauge the accuracy of risk models. Therefore, we first undertake a brief study of properties of these evaluation methods.

3.1. Frequency-Based Tests

The principle of frequency-based backtests is that the hits sequences (violations) should satisfy two properties: the unconditional coverage property and the independence property (Christoffersen, 1998). If a risk model is adequate, it will generate 'correct' frequency of violations, and violations are independent of each other. Evidence of violations clustering would suggest that the model is misspecified, even if the model correctly predicts the unconditional coverage.

3.1.1. Binomial Backtesting Method

Currently the standard approach to evaluation is the basic frequency (or binomial) test proposed by Kupiec (1995). It examines whether the observed frequency of violations (e.g., the frequency of tail losses that exceed VaR) is consistent with the frequency of tail losses predicted by the model. In particular, under the null hypothesis that the model is 'good' (or consistent with the data), the number of violations follows a binomial distribution.

The hit sequence of $VaR_t(p)$ violations is defined as

$$I_t = \{ \begin{matrix} 1, & \text{\textit{if} } y_t > VaR_t(p) \\ 0, & \textit{else} \end{matrix}$$

[8] The choice of risk measurement method by regulators is based on the tools available to verify model quality (Kupiec binomial backtesting currently standard approach embodies in Market Risk Amendment, 1996). This is the main reason that the VaR approach is often preferred to coherent risk measures (Kerkhof and Melenberg, 2002).

Given *m P/L* observations and a predicted frequency of violations equal to p , the probability of n violations follows:

$$\Pr(n \mid m, p) = \binom{m}{n} p^{n} (1-p)^{m-n} \tag{4}$$

This test has a simple intuition, is easy to apply and only requires knowledge of *m*, *n* and *p*. However, it lacks power (e.g., the ability to identify incorrect risk models) except with a large sample sizes. Kupiec (1995) shows that with sample size of 250 observations and 99% coverage rate, the odds of detecting financial institutions that systematically under reporting their VaR are about 65% - only slightly better than a coin flip.

This is because the binomial backtesting loses potentially valuable information. Firstly since it focuses exclusively on the frequency of violations over the sample period, this test throws away information about the temporal pattern of violations, thus it ignores the independent property in violations. Secondly the binomial backtesting discards information on the sizes of violations predicted by risk forecasting models. This has the unfortunate implication that an 'incorrect' risk model will pass a frequency test if it generates an acceptably accurate frequency of violations, even if its forecasts losses, larger than VaR, are very poor.

3.1.2. Interval Forecast Backtesting Method

One way to test that predicted violations are iid is the interval forecast backtesting suggested by Christoffersen (1998). His idea is to test the prediction of correct unconditional coverage (e.g., the model generates the 'correct' frequency of violations) and independence (e.g., the predicted violations are independent of each other) separately. In the presence of the time-dependent heteroskedasticity often found in financial time series, the issue of independence of violations is of particular important. Berkowitz and O'Brien (2002) have reported on the performance of actual VaR forecasts from six large U.S. commercial banks. Even though banks tend to be conservative – they have fewer than expected violations – violations are large and appear to be clustered in time and across banks.[9] From the perspective of a regulator worried about systemic default, rejecting a particular bank's risk model due to the clustering of violations is particularly important, if violations also happen to be correlated across banks.

If *n* is the number of violations in the sample and *m* is the number of observations, then the observed frequency of violations is n/m. Given that the predicted probability of violations is *p*, the unconditional coverage test can be expressed in terms of a likelihood ratio (*LR*) test. The test statistic

$$LR_{uc} = -2\ln\left[(1-p)^{m-n} p^{n}\right] + 2\ln\left[(1-n/m)^{m-n}(n/m)^{n}\right] \tag{5}$$

is distributed as a $\chi^2(1)$, a chi-squared with one degree of freedom.

Turning to the independence prediction, Christoffersen (1998) considers a two-state (e.g., correct forecast/wrong forecast) Markov chain, as a likelihood ratio test of the null hypothesis that successive observations are statistically independent, against the alternative hypothesis that observations are from a first-order Markov chain.

Assume a binary first-order Markov chain, $\{I_t\}$ with transition probability matrix

$$\Pi = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

where $\pi_{ij} = \Pr\left(I_t = j \middle| I_{t-1} = i\right)$[10], i and j refer to states of violations / non-violations.

Under the hypothesis of independence, the test statistic

$$LR_{ind} = -2\ln\left[(1-\hat{\pi}_2)^{n_{00}+n_{10}}\,\hat{\pi}_2^{n_{01}+n_{11}}\right] + 2\ln\left[(1-\hat{\pi}_{01})^{n_{00}}\,\hat{\pi}_{01}{}^{n_{01}}\,(1-\hat{\pi}_{11})^{n_{10}}\,\hat{\pi}_{11}{}^{n_{11}}\right] \tag{6}$$

is distributed as a $\chi^2(1)$, where n_{ij} is the number of observations with state i followed by j.

[9] The majority of violations appear to take place during the August 1998 Russia default and ensuing Long-Term Capital Management (LTCM) debacle.

[10] The Maximum likelihood estimate probability is π_{ij}: $\hat{\pi}_{01} = \dfrac{n_{01}}{n_{00}+n_{01}}$, $\hat{\pi}_{11} = \dfrac{n_{11}}{n_{10}+n_{11}}$, $\hat{\pi}_2 = \dfrac{n_{01}+n_{11}}{n_{00}+n_{10}+n_{01}+n_{11}}$

It follows that under the combined hypothesis of correct coverage and independence – the hypothesis of correct conditional coverage, the test statistic

$$LR_{cc} = LR_{uc} + LR_{ind} \tag{7}$$

is distributed as a $\chi^2(2)$, a chi-squared with two degrees of freedom. If a risk model is adequate, then violations should be Bernoulli variables.

The advantage of this approach is to identify the source of model failures, while at the same time it enables us to test both coverage and independence properties. However, this approach is not without its faults. The interval forecast backtesting remains quite data-intensive, since it only takes two values (0 and 1) to establish of whether or not a violation occurs. Berkowitz (2001) reports that with a 95% VaR, the test shows some rejections as a sample size increases to 500 observations. A part of reason for this low rejection power is that the interval forecast backtesting also discards the useful information of magnitude of losses.

3.2. Density Forecast Backtesting Method

In general, there is no need to restrict attention to a single VaR level. The unconditional coverage and independence properties of an accurate VaR measure should hold for any level of α. As suggested by Crnkovic and Drachman (1996) and Diebold, Gunther and Tay (1998), realized values of variables whose density is being forecast should be mapped to their probability integral transform or forecast cumulative density values (or Rosenblatt transformation). If y_t is the day-*t P/L* value, and this observation is associated with a forecasted cumulative density function $\hat{F}_t(\cdot)$, which in principle might change from one day to the next, then the transformed observations is the value of $\hat{F}_t(\cdot)$ evaluated at y_t:

$$\hat{U}_t = \int_{-\infty}^{y_t} \hat{f}_t(u)du = \hat{F}_t(y_t) \tag{8}$$

where $\hat{f}_t(\cdot)$ is the probability density function. $\hat{U}_t$ is the forecast probability of observing an outcome no greater than that actually realized. If $\hat{F}_t(\cdot)$ is correct, then $\hat{U}$ has a uniform *U*[0,1] distribution. If a sequence of density forecasts is correctly conditionally calibrated then, analogously to the no-autocorrelation requirement (or independency property) discussed above, the corresponding *U*-sequences is *i.i.d.* *U*[0,1]. Diebold, Gunther and Tay (1998) present histograms of U for visual assessment of unconditional uniformity, and various autocorrelation tests. Diebold, Tay and Wallis (1999) use the chi-squared goodness-of-fit test, also the Kolmogorov-Smirnov test on the sample distribution function of U. The series of density forecasts they evaluate - the U.S. Survey of Professional Forecasters' (SPF) inflation forecasts.

The application of the Rosenblatt transformation paves the way to apply distribution equality tests to assess model adequacy. In particular, under the null hypothesis that the model is adequate, we would expect the lowest 5% of transformed observations to fall in the region between 0 and 0.05, the next lowest 5% of observations to fall between 0.05 and 0.1, and so on. So under the null hypothesis that a risk model is adequate, the Rosenblatt transformed data are predicted to be distributed as standard uniform (e.g., iid *Uniform (0,1)*).

Unfortunately, testing the iid uniform distribution hypothesis is cumbersome because bounded support may cause technical difficulties. Berkowitz (2001) suggests transforming an iid uniform $\hat{U}_t$ to an iid standard normal variable, $\hat{Z}_t$ using the inverse cumulative distribution function, Φ^{-1}

$$\hat{Z}_t = \Phi^{-1}\left(\hat{U}_t\right) = \Phi^{-1}\left(\int_{-\infty}^{y_t} \hat{f}_t(u)du\right) = \Phi^{-1}\left(\hat{F}_t(y_t)\right) \tag{9}$$

The Berkowitz transformation converts a uniform series into a standard normal series, and therefore a risk model adequacy can be examined by means of tests for standard normality. Assume for the time being that there is an iid prediction, in which case the full null prediction is that $\hat{Z}_t$ is iid *N(0,1)*. Berkowitz suggests that we can test this by nesting the null hypothesis within a first-order autoregressive process with a possible different mean and variance. If we write this process as

$$Z_t - \mu = \rho\left(Z_{t-1} - \mu\right) + \varepsilon_t \tag{10}$$

then the null hypothesis predicts that $\mu = 0, \rho = 0$ and σ^2 , the variance of ε_t should equal 1[11].

The exact log-likelihood function associated with Equation (10) is known to be:

$$-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\left[\frac{\sigma^2}{\left(1-\rho^2\right)}\right]-\frac{[Z_1-\mu/(1-\rho)]^2}{2\sigma^2/\left(1-\rho^2\right)}-\frac{m-1}{2}\log(2\pi)-\frac{m-1}{2}\log(\sigma^2)-\sum_{t=2}^{m}\left(\frac{(Z_t-\mu-\rho Z_{t-1})^2}{2\sigma^2}\right)$$

The likelihood ratio test for the null hypothesis is then:

$$LR_{full} = -2\left(L(0,1,0) - L\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right)\right) \tag{11}$$

where $\hat{\mu}, \hat{\sigma}^2$ and $\hat{\rho}$ are maximum likelihood estimates of parameters concerned. The LR statistic is distributed under the null hypothesis as a $\chi^2(3)$, a chi-squared with three degrees of freedom. Since the LR test explicitly accounts for mean, variance, and autocorrelation of the transformed data, it should have power against very general alternatives.

However, this testing procedure has a weakness: as noticed by Dowd (2004), it focuses on whether the first two moments of distribution are compatible with standard normality, but it has little power in the face of departures from the standard normality that manifest themselves in the higher moments of distribution. We therefore include the Jarque-Bera (JB) statistic as a supplement to the LR test in the paper.

The JB test was proposed by Jarque and Bera (1980). This test is based on the difference between skewness and kurtosis of the data set $\{z_1, z_2, ..., z_n\}$ and those of the assumed normal distribution.

The null hypothesis and the alternative for the JB test are:

$$H_0 : z_i^{'}s \sim iidN\ \left(\mu, \sigma^2\right);$$
$$H_A : not\ \ H_0$$

The JB test statistic is:

[11] In this paper, we only consider the one-step-ahead forecast and $\varepsilon \sim iid\ N(0,1)$

$$JB_{full} = n\left(\frac{\alpha_3^2}{6} + \frac{(\alpha_4 - 3)^2}{24}\right) \tag{12}$$

where

$$\alpha_3 \equiv \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^3}{s^3}$$

$$\alpha_4 \equiv \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^4}{s^4}$$

$$s^2 \equiv n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

Here, $\bar{z}$ is the sample mean, and s^2, α_3 and α_4 are the second, third and fourth sample moments about the mean, respectively. The JB statistic has an asymptotic distribution, which is $\chi^2(2)$ under the null hypothesis. This test is known to have very good power properties in testing for normality.

3.2.1. Evaluation methods of ES

A type of model failure of particular interest to financial institutions and regulators is the inaccuracy of forecasted magnitude of large losses. Basak and Shapiro (2001) and Artzner et al.,(1999), for example, emphasize ES given that a violation occurs $\hat{E}(y_t | y_t > VaR_t(p))$. In order to formally test for misspecifications in the tail of the density forecast, Berkowitz (2001) suggested the LR test based on a censored likelihood that allow the user to intentionally ignore model failures that are limited to the interior of the distribution. Let the desired cutoff point, $VaR = \Phi^{-1}(p)$, we define the new variable of interest as

$$z_t^* = \begin{cases} VaR & if\ z_t \geq VaR \\ z_t & if\ z_t < VaR \end{cases} \tag{13}$$

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

$$\begin{aligned} L(\mu, \sigma | z^*) &= \sum_{z^* < VaR} \ln \frac{1}{\sigma}\phi\left(\frac{z_t - \mu}{\sigma}\right) + \sum_{z_t^* = VaR} \ln\left(1 - \Phi\left(\frac{VaR - \mu}{\sigma}\right)\right) \\ &= \sum_{z^* < VaR}\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma}(z_t^* - \mu)^2\right) + \sum_{z_t^* = VaR} \ln\left(1 - \Phi\left(\frac{VaR - \mu}{\sigma}\right)\right) \end{aligned} \tag{14}$$

The likelihood ratio test for the null hypothesis is then:

$$LR_{tail} = -2\left(L(0,1) - L\left(\hat{\mu}, \hat{\sigma}^2\right)\right) \qquad (15)$$

Under the null hypothesis, the test statistic is distributed $\chi^2(2)$. This test has 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 has excessively small / large losses relative to the forecast. We compare the shape of the forecasted tail of the density to the observed tail. A rejection based on the tail density taken as a proxy for rejection of the mean of the tail, or ES.

3.2.2. Evaluation methods of SRM

This paper further adds to the literature on evaluation methods by backtesting on SRM. Consider the left tail of the forecasted distribution, a SRM is calculated as a weighted average of quantiles of a loss distribution. A key issue to evaluation risk models is the accuracy of forecasting loss distributions, since a weighting function (or risk spectral) in SRM calculation reflects a user's attitude toward the risk, it should not affect the forecast ability of a risk model. Christoffersen and Pelletier (2004) suggested that if we want to test that $\tilde{U}_t$ observations over the interval $[0, p]$ (in our case, $p = 50\%$) largest losses are themselves uniform, we can construct a rescaled $\tilde{U}_t$ variable as

$$\tilde{U}_t^* = \begin{cases} \tilde{U}_t / p & \tilde{U}_t \in [0, p] \\ Else\, not\, define. \end{cases} \qquad (16)$$

Then we can test the null hypothesis that the risk model provides the correct tail distribution as $\tilde{U}_t^*$ is iid *U(0,1)* or equivalently, $\tilde{Z}_t^* = \Phi^{-1}\left(\tilde{U}_t^*\right)$ is iid *N(0,1)*. The previous LR_{full} and JB_{full} test framework (Eq. 11 and 12) can be applied. The shape of the forecasted tail of the density is compared to the observed tail. A rejection based on the tail density taken as a proxy for rejection of SRM[12].

[12] This framework can be applied on the ES. However, in this paper, we do not pursue this approach because parameters estimation based on a small sample size are unrealizable. Furthermore, as indicated by Lawford (2005), the JB statistic is broke-down for the sample size less then 4.

While the large-sample distribution of the LR test and the JB statistic we have discussed above are well known, they may not lead to reliable inference in realistic risk management settings[13]. The nominal sample sizes can be reasonably large, say two to four years of daily data, but the scarcity of violations of, for example, the 1% VaR renders the effective sample size small. In order to make rejection power comparable across statistics with different sample sizes, we estimate the Monte Carlo critical value that gives rise to 0.05 under the null. The finite sample critical value is given in Table 1.

[Insert Table 1 here]

The other problem with such a method is parameter uncertainty. If the parameter is estimated, it is possible that transformed observations departs from iid *N(0,1)*, even when the density forecast model coincides with the true density. As stressed by Bawa *et al.*, (1979), the predictive distribution of an asset return that is obtained by integrating the conditional distribution over the parameter space is different from the predictive distribution that is obtained when the parameters are treated as known. West (1996) indicates that forecasts are produced by model estimates with small samples size are subject to parameters uncertainty problem. However, in the risk management practice, financial institutions are very much after the total risk measure due to the forecast errors, a distinction between "model risk" and "estimation risk" is not a practical concern. Moreover, existing work suggests that parameter uncertainty is of second-order importance when compared to other source of inaccurate forecasts such as model misspecification (for detail, see Chatfield, 1993). Diebold, Gunther and Tay (1998) find that the effects of parameter estimation uncertainty are inconsequential in simulation studies geared towards sample sizes relevant in finance[14].

[13] It is well known that the likelihood ratio tests rely on asymptotic theory and it is only valid if the sample sizes are reasonably large and well balanced across populations. For small, sparse, skewed, or heavily tied data, the asymptotic theory may not be valid. See Agresti and Yang (1987) for some empirical results, and Read and Cressie (1988) for a more theoretical discussion. For general discussion and references to earlier literature see Stuart, Ord and Arnold (1999, Ch. 25).

[14] More detail on how to deal with parameter uncertainty problem, see Bao et al. (2004), Duan (2003), and West (1996).

3.3. Probability Forecast Backtesting Method

It is often the case that financial institutions or regulators are not only interested in how individual models perform, but also in how different models compare to each other. In practice, it is rarely the case that we can find an optimal model. All the models proposed by different researchers can be possibly misspecified and the true distribution is in fact too complicated to be represented by a simple mathematical function (Sawa, 1978). Our task is then to investigate which model can approximate the true data generate process most closely. This can be done using the probability forecast backtesting method that give each model a score in terms of some loss function. The loss scores can then be used to rank models – the lower score, the better model.

This "loss function" approach can be a useful supplement to these more formal statistical methods and provides a way to define the institution's criteria of an "accurate" model. For example, we can design a loss function in which the modeler can weight the penalties to assign to violations given their frequency, magnitude, or time dependencies, and compare them with expected tail loss numbers. The main benefit of this type of analysis is that it provides a measure of relative performance that can be used for "backtesting" different models (see Dowd, 2002). This method is not a hypothesis-based test; instead, the forecasted distribution transforms what might happen in the future into probability forecasts (Lopez, 1999a). That is, the accuracy of a risk model is gauged by how well the probability forecasts from the model minimize a loss function that represents the user's interests.

This evaluation approach has a number of attractions: because they are not statistical tests, forecast evaluation approaches do not suffer from the low power of standard tests, such as frequency-based tests. This makes them useful for backtesting with small data sets. Furthermore, this approach allows us to tailor a loss function to take account of particular concerns. For example, a risk manager might be more concerned about higher losses than lower losses, and therefore wish to given higher losses a greater weight in his/her loss function. However, this evaluation method cannot be use directly to identify a risk model as "acceptably accurate" or "inaccurate" in an absolute sense.

The ranking process has three key components: loss function, benchmark, and proper score function. The crucial component in evaluating forecast accuracy is a loss function[15], which represents the 'cost' associated with various pairs of forecasts and realizations. A benchmark gives us an idea of the score we could expect from a 'good' model. A score function[16] provides summary measures for the evaluation of probability forecasts, by assigning a numerical score based on the forecast and on the event or value that materializes. In terms of evaluation, scoring rules measure the quality of the probability forecasts and rank competing forecast procedures.

The most common score is Brier's quadratic probability score (QPS)[17] , which is defined as

$$QPS = \frac{2}{n}\sum_{t=1}^{n}\left(L_t - B\right)^2$$

where L_t is the loss function, B is the benchmark and n is the sample size. It is an analogous score to the mean squared error (MSQ) for probability forecasts and thus is a quadratic loss function. Because it is quadratic, QPS penalizes deviations of actual losses from their expected value. It gives greater weight to very high losses than to smaller losses, which makes intuitive sense.

3.3.1. A Loss Function for VaR

The most straightforward is a binary loss function proposed by Lopez (1999a):

$$L_t = \begin{cases} 1 & \text{if} \quad y_t > VaR_t \\ 0 & \quad\quad y_t \le VaR_t \end{cases} \tag{17}$$

This loss function is exclusively concerned with the frequency of tail losses, thus it ignores the magnitude of tail losses. The benchmark is p - the expected value of $E(L_t)$.

3.3.2. A Loss Function for ES

[15] See Carmona (2005) for a comprehensive review in this area.

[16] Gneiting and Raftery (2005) have given a comprehensive review and developed a theory of proper scoring rules.

[17] Selten (1998) gave an axiomatic characterization. Quadratic probability score (QPS) has a negative orientation – that is, smaller values indicate a more accurate forecast.

In the literature, there are a number of loss functions closely relating to ES. We consider one loss function in particular, suggested by Dowd (2004). This loss function takes the form of tail loss itself, if violation occurs, and 0 otherwise.

$$L_t = \begin{cases} y_t & \text{if} \quad y_t > VaR_t \\ 0 & \quad\quad\; y_t \le VaR_t \end{cases} \tag{18}$$

The expected value of the tail loss is ES, so we can choose ES as benchmark.

3.3.3. A Loss Function for SRM

We state the loss function for SRM, in an analogous way to Bao et al. (2004). In their paper, authors define the loss function as the distance between the candidate density forecast model and the true model. The forecasted density is transformed using Berkowitz normal transformation as we discussed in section 3.2. If a risk forecasting model is adequate, the transformed variables should distributed as iid *N(0,1).* Thus, the benchmark is the standard normal density. The score function used in their paper is Kullback-Leibler Information Criterion, or logarithm score function.

We specify the loss function as the distance between the transformed candidate's forecasted density and the standard normal density. The QPS is calculated as the different between each quantile of two densities. As indicated by Gneiting and Raftery (2005), specifying a predictive cumulative distribution function is equivalent to specifying all predictive quantiles. If a one-step-ahead density forecast is correctly specified and hence optimal, the transformed observations should be distributed as standard normal, as it dominates all other density forecasts for any loss function (Granger and Pesaran, 2000a, b; Diebold et al., 1998). The transformation of SRM is produced in section 3.2.1. Thus, the loss function is defined as $L_t = \tilde{Z}_t^* = \Phi^{-1}\left(\tilde{U}_t^*\right)$, quantiles of the transformed candidate's forecasted distribution.

$$L_t = \begin{cases} \tilde{Z}_t^* \\ Else\,not\,define. \end{cases} \tag{19}$$

The benchmark is corresponding quantiles of the standard normal distribution, or z_t. Thus the QPS is

$$QPS = \frac{2}{n}\sum_{t=1}^{n}\left(L_t - z_t\right)^2 \qquad (20)$$

We penalize departures from the standard normal distribution for overestimate or underestimate risk measures. Still, we do not consider a weighting scheme of SRM in this loss function. Indeed, if we attach weights to quantile differences, it would not make a difference to the model's ranking.

4. Simulation Experiments

The goal of experiments is to analyze the performance of various evaluation methodologies when they are applied to three risk measures. Using this information, we hope to reduce chances of model misclassification. For three hypothesis-based backtests, we focus on the power of statistic tests (e.g., the ability to identify the incorrect model). We consider two coverage probabilities (e.g., $\alpha = 99\%$ and 95%) and based on these coverage rates to calculate a VaR and an ES, and illustrate evaluation results of the unconditional coverage test (or "*UC*") for the binomial backtesting, and the independent test (or *"IND"*) and the conditional coverage test (or "*CC*") for the interval forecast backtesting. We specify the null hypothesis (e.g., $\mu = 0; \sigma = 1; \rho = 0$ on SRM and $\mu = 0; \sigma = 1$on ES) for the *LR* test and the null hypothesis ($skewness = 0; kurtosis = 3$) for the *JB* statistic for the density forecasts backtesting.

With respect to the probability forecast backtesting, its ability to classify risk models (e.g., accurate versus inaccurate) is gauged by how frequently QPS values for the true data generating process (or "DGP") are lower than that of alternative models. Three types of loss function are therefore examined in this paper. The risk measures will be modeled with techniques that are commonly used by researchers for constructing risk measures (Duffie and Pan, 1997; Dowd, 1998). At the same time, the data generating process is necessarily kept simple to allow for a computationally tractable simulation study.

For a typology of various models of $\{y_t\}_{t=1}^{T}$, let it follows the stochastic process

$$y_t = \sigma_t \eta_t \tag{21}$$

where σ_t can be estimated either conditional or unconditional, $\eta_t \equiv \varepsilon_t / \sigma_t$ and η_t has distribution f_t . As can be seen, a density forecast model based on (21) can be decomposed into two parts: specification of σ_t^2 and specification of the distribution of $\{\eta_t\}_{t=1}^T$.

The simulation exercise is conducted in two distinct settings – the unconditional and the conditional setting. In the first setting, the emphasis is on the shape of the $f(\cdot)$ distribution alone. To examine how well various backtesting methods perform under different distributional assumptions, experiments are carried out by setting $f(\cdot)$ to a standard normal distribution, a t-distribution with 4 and 6 degrees of freedom, a skew-t distribution, and a Generalized Pareto distribution (or "GPD"). We use the t(6) distribution to generate the dataset as the true DGP. The selection of alternative distributions reflects the shape variations to the true distribution. One special case is the GPD, which pays special attention in the tail area, and it provides superior tail estimation than for alternatives such as a normal distribution.

The second setting examines the performance of backtesting methods in the presence of variance dynamics in ε_t . Specifically, we assume that volatility dynamics are introduced by using the conditional heteroskedasticity of a GARCH model with t(6) innovations. Alternative models are a GARCH model with normal innovations, an exponential weighted moving average (or EWMA) process, and a homoskedastic model with t(6) innovations. Therefore, we analyze the evaluation performance affected by dynamic volatility estimations (e.g., GARCH vs. EWMA vs. homoskedastic volatility) as well as distribution assumptions (e.g., student's t vs. normal distribution).

A simple, symmetric GARCH (1, 1) model, where the daily variance evolves as

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{22}$$

We have chosen the parameter value $[\omega, \alpha, \beta] = [0.0000004, 0.0551, 0.9431]$ to mimic values typically obtained when pre-fitting the GARCH model to the S&P 500 index futures data.

An alternative model is a EWMA or RiskMetrics volatility model, where the daily variance evolves as

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1-\lambda)\varepsilon_{t-1}^2 \tag{23}$$

Following JP Morgan RiskMetrics, we fix $\lambda = 0.94$.

In all settings, simulation runs are structured similarly. For each run, the simulated y_t series is generated using the chosen data generating process (e.g., a t(6) model in the unconditional setting and a GARCH (1,1) model with t(6) innovations in the conditional setting). The chosen length of the in-sample series (after 1000 start-up observations) is 2000 observations, which roughly corresponds to eight years of daily observations. Alternative risk models are then used to generate one-step-ahead risk forecasts for the next 125, 250, 500, and 1000 observations[18] of y_t. The forecasts from various risk models are then evaluated using appropriate evaluation methodologies. The rejection rate is calculated based on the finite-sample critical value in Table 1 on 2000 simulations.

The simulation results are organized below with respect to the unconditional and the conditional settings - that is, results of four backtesting methods are presented for each DGP and alternative risk measures[19]. Three general points can be made regarding the results. Firstly, with respect to three hypothesis-based backtests, the power of frequency-based backtests (e.g., the binomial backtesting and the interval forecast backtesting) in rejecting incorrect null hypotheses is lower than that of the density forecast backtesting. Therefore, the chance of misclassifying inadequate risk model as adequate is high for the frequency-based backtests. In addition, the backtesting for ES and SRM is more effective

[18] Size of 125, 250, 500 and 1000 approximately corresponds to half year, one year, two years and four years data.

[19] We present simulation results on the left tail only. The simulation results on the right tail are available on request.

than for VaR in identifying an incorrect model from alternative models in a small sample setting.

Secondly, we propose the loss function for SRM where the probability forecast backtesting is persistently able to distinguish the correct risk model from alternative models in all of cases examined. That is, the QPS for the true model is lower than that of alternative models.

Thirdly, four evaluation methods are more sensitive to the chosen misspecifications of the distributional shape of f_t than to the chosen misspecifications of variance dynamics for all cases examined. That is, all four methods are capable of differentiating between the true model and alternative models with the same variance dynamics but different distributional assumptions.

4.1. Unconditional Simulation Experiment Results

As previously mentioned, an important issue in examining the performance of statistical evaluation methods is a finite-sample size of underlying test statistics. Table 2A reports the finite-sample rejection rates of risk measures for three hypothesis-based backtests examined in this paper. Table 2B presents the finite-sample QPS of risk measures using the probability forecast backtesting method. These finite-sample rejection rates are based on 2,000 simulations of sample sizes 125, 250, 500 and 1,000 and corresponding coverage rates – 99% and 95%. The desired confidence level of tests is 0.05.

[Insert Table 2A here]

Table 2A presents evaluation results of three hypothesis-based backtests on three risk measures. We compare the evaluation performance of the binomial backtesting, the interval forecast backtesting and the density forecast backtesting. The top panel of the table, labeled ‘size’, reports Monte Carlo rejection rates when the model coincides with the true model (e.g., the ‘t(6)’ model). The first two columns report rejection rates of a 99% VaR and a 95% VaR using the binomial test. ‘UC’ indicates the unconditional coverage test. Rejection rates in the first two columns are uniformly smaller than a 7%, which is approximately correct sized. This is perhaps not surprising – since the underlying process coincides with the true model. Columns 3 to 6 reports the rejection rates of a 99% VaR and a 95% VaR using the interval forecast backtesting. 'IND' indicates the independent test, and 'CC' indicates the conditional coverage test. The size properties of these tests are quite similar to those of the binomial test. However, as noticed by Berkowitz(2001), the rejection power of the interval forecast backtesting is a slightly lower than the binomial backtesting, due to the interval forecast backtesting requires information on the dynamics of violations not just the number of violations. Our results show that the rejection power of the interval forecast backtesting is not always lower than that of the binomial test. The last four columns present the rejection rates of a 99% ES, a 95% ES and a SRM in the density forecast backtesting. 'LR' indicates the likelihood ratio test and 'JB' indicates the Jarque and Bera test. These test statistics display approximately correct size, rejection rates for all three risk measures are around a 5%.

The lower panels show rejection rates when models are wrong, and are therefore labeled 'power' (e.g., the ability to identify wrong models). The panel labeled "Normal" reports the results when the model is estimated using a normal distribution. With a 95% VaR and a 99% VaR, the rejection rate of the binomial test and the interval forecast backtesting is below a 30%, even with 1000 observations. This should be expected -- with so few violations, very large samples are required to generate rejections. In addition, there is not even a 10% probability of rejecting a false model with a sample size of 250 observations as recommended by the Basel II. On the other hand, the density forecast backtesting would detect the model failure of a 70% times on a 99% ES and a 95% ES with 250 observation. The LR test shows that the rejection rate of a SRM with 250 observations is only a 21.8%. However, the rejection rate of the JB test reaches almost a 70%. Since the LR test can only detect a model failure on the first two moments of a distribution. We expect that the JB test have power to reject the incorrect model.

The panel labeled “Skew-t” reports the results when the model is estimated using a skew-t distribution. The rejection power of the binomial test and the interval forecast backtesting increase dramatically. With 250 observations, rejection rates reach a 40%, compare to the ‘Normal’ model of a 10%. These statistic results indicate that the frequency-based backtests is more sensitive to large variations on shape of the distribution. Still a 40% probability of rejecting a false model is low. However, with the density forecast backtesting and 250 observations, rejection rates of an ES (either a 99% or a 95%), and a SRM are over a 70%. Therefore, the probability of misclassifying inadequate risk model as adequate is lower for the density forecasts backtesting than for the frequency-based backtests.

We are also interested the impact of a small variation on the kurtosis of a distribution on the validity of the backtest methodologies, and report results on the panel labeled “t(4)”. Overall rejection rates are low – with the frequency-based test, a false model can be rejected with probability less than a 20%, and a 50% with the density forecast backtesting. We expect that the JB statistic would have power to capture the departure on the kurtosis. However, with rejection rates are less than 16% in all cases examined, the JB statistic is fail to detect this kind of a model failure.

The last panel labeled “GPD” reports the results when the model is estimated using a Generalized Pareto distribution. The size properties of all tests are very similar to that of the first panel (e.g., labeled “size”). This is because the GPD pays special attention to the tail and it allows for some extrapolation beyond the range of the data (Brooks et al., 2005). Therefore, these statistic results are consistent with prior empirical findings in Extreme Value Theory literature.

[Insert Table 2B here]

Table 2B presents sets of comparative accuracy results of the probability forecast backtesting. We compare numerical scores on different DGPs with four different sample sizes, and each panel of the table represents one sample size. The first and second columns report the QPS of a 99% VaR and a 95% VaR. The QPS calculation is based on the loss function that suggested by Lopez (1999a), and labeled “Lopez”. Columns 3 and 4 present the QPS of a 99% ES and a 95% ES. The QPS calculation is based on the loss function that suggested by Dowd (2004), and labeled “Dowd”. The last column reports the QPS of a SRM. The QPS calculation is based on the loss function that we suggest in this paper, and labeled “SRMP”.

For all cases examined, the true model’s QPS (e.g., the ‘t(6)’ model) is lower than alternative risk models for each defined loss functions, except with a sample size less than 500 observation, the ‘Lopez” loss function cannot differentiate the correct model from alternative models – the QPS is lower for the “t(4)” model than for the true model. Furthermore, the ‘GPD’ model performs as well as the true model in majority cases examined. In particular, the ‘Skew-t’ model is clearly found to be inaccurate with respect to the true model – that is, the QPS is higher than that of rest models. Therefore, we propose the loss function for SRM where the probability forecast backtesting is persistently able to distinguish the correct risk model from alternative models, even with a sample size as small as 125 observations.

4.2. Conditional Simulation Experiment Results

Table 3A reports evaluation performance results of three hypothesis-based backtests when we take into account variance dynamics in the DGP. We illustrate four alternative volatility processes – the “GARCH-t(6)” model, the “Homoskedastic-t(6)” model, the “GARCH-normal” model, and the “EWMA-normal” (or “EWMA”) model.

[Insert Table 3A]

The top panel of the table, labeled ‘size’, reports Monte Carlo rejection rates when the model coincides with the true model (e.g., the ‘GARCH-t(6)’ model). In all cases examined, the overall rejection rates on risk measures – VaR, ES and SRM, are lower – less than a 10%, and approximately correct sized as we expected.

The lower panels show rejection rates when models are wrong, and are therefore labeled ‘power’ (e.g., the ability to identify wrong models). The panel labeled “Homoskedastic-t(6)” reports the results when the model is estimated using a long-run volatility with t(6) innovations. With a 99% VaR and a 95% VaR, rejection rates of the binomial test and the interval forecast backtesting are low – less than a 50% with 250 observations. Thus, the likelihood of misclassifying an inadequate model as adequate is high. In addition, rejection rates of the density forecast backtesting with a 99% ES and a 95% ES are also low – less than a 60% with 250 observations. However, the rejection rate is much high with SRM - over an 80% with 250 observations.

The panel labeled “GARCH-normal” reports the rejection rate when the model is estimated using a GARCH(1,1) model “GARCH-normal” with a normal distribution. Still, with 250 observations, rejection rates of the binomial test and the interval forecast backtesting are quite low - less than a 60%, though they are a slightly higher compare to the previous panel (e.g., less than a 50%). However, the rejection rate of the density forecast backtesting increases considerably. For instance, with 250 observations, rejection rates of a 99% ES and a 95% ES are over 70%, compare to the previous panel of under 60%. The rejection rate of SRM also increases from an 82.3% from the previous panel to a 97.5%.

The last panel labeled “EWMA” reports the rejection rate when the model is estimated using an exponential weighted moving average model with normal innovations. In all cases examined, rejection rates of the “EWMA” model are very similar to the “GARCH-normal” model. For instance, with 250 observation, rejection rates of the binomial test and the interval forecast backtesting are also around a 60% on VaR (either a 99% and a 95%), an 80% for ES (either a 99% and 95%), and over a 95% for SRM. These results suggest that backtests methods are more sensitive to the chosen misspecifications of distributional shapes than to the chosen misspecifications of variance dynamics. In addition, the evaluation performance on SRM is as good as on ES. Thus, the backtesting for SRM and ES is more effective than for VaR in identifying an incorrect model from alternative models.

[Insert Table 3B here]

Table 3B reports sets of QPS using the probability forecast backtesting when dynamic volatility is presented in the DGP. Evidently, the 'Lopez' loss function for VaR is unable to distinguish the correct model from alternative models - the QPS is lower for the 'Homoskedastic-t(6)' model than for the true model (e.g., the 'GARCH-t(6)' model) in a 97% of cases. As a sample size reaches 500 observations, the 'Dowd' loss function is capable of differentiating the correct model from alternative models. In contrast, the 'SRMP' loss function that we suggested in this paper is consistently able to identify the correct model from alternative risk models with a sample size as small as 125 observations. The 'GARCH – normal' model and the 'EWMA' model are clearly found to be inaccurate with respect to the true model.

These results also reveal that the probability forecast backtesting has more rejection power against alternative models with incorrect distribution assumptions, but has less power with respect to the true variance dynamics. For instance, with the 'Lopez' loss function, the QPS of the 'Homoskedastic - t(6)' model is lower than the true model (e.g., the 'GARCH-t(6)' model). Furthermore, the difference in QPS between the 'Homoskedastic - t(6)' model and the true model is very small with the 'Dowd' and the 'SRMP' loss functions.

These results shed interesting light on tradeoffs between modeling the distribution and modeling the time-varying volatility. The rejection rates in Tables 3A and 3B suggest that these tests have little power against alternative models characterized by a close approximation of the true variance dynamics, but have better power against incorrect distributional assumptions. This is consistent with the prior empirical findings by Lopez (1999a, b), Berkowitz (2001) and Bao et al., (2004).

5. Empirical Application

In this section, we illustrate previous evaluation methods by fitting three conditional volatility models – a GARCH(1,1) model with normal innovations, labeled a “GARCH-normal” model, a GARCH(1,1) model with t innovations, labeled a “GARCH-t” model, and an EWMA model with normal innovations, labeled “EWMA” model, to three most heavily traded index futures –the S&P500; the FTSE 100 and the Nikkei 225 for the period of 01/11/1998 to 31/10/2006. The data is obtained from DataStream and consists of 2087 daily close prices. Initially we estimate parameters over the period 01/11/1998 to 31/10/2004. The sample spans 6 years period and contains 1586 observations. This leaves an evaluation period of 500 observations covering two years of data. Having calculated the volatility forecasts based on parameters of this initial sample, the subsequent samples are produced by rolling forward one trading day, keeping the sample size constant at 1586 observations. Based on forecasted volatility estimates, we calculate a range of risk measures.

Figure 1 is a graphical representation of time series properties of daily logarithmic price changes in three index futures and squared daily logarithmic price changes from the period 01/11/1998 through 31/10/2006. The series of daily log changes is a mean zero process exhibiting periods of relative calm punctuated by periods of disturbed volatility. The squared daily logarithmic changes exhibit volatility clustering, *i.e.*, partial predictability of the conditional variance of this series. This implied that risk exposure is not identical at each point in time. For risk management purposes, a conditional heteroscedastic model is appropriate when evaluating the risk exposure conditional on the current volatility regime.

[Insert Figure 1 here]
[Insert Table 4 here]

Table 4 contains descriptive statistics of daily logarithmic return series of three index futures. The full set of returns follow usual stylized facts about futures data, namely, price movements do not belong to a normal distribution using a Kolmogorov – Smirnov test, there is negative skewness (the S&P 500 contract excepted), and also leptokurtosis present. Therefore, statistic determination of risk measures based on the normal distribution is inappropriate, and would lead to inadequate risk measures estimation. Leptokurtosis is demonstrated by a fat-tail characteristic, and this is most evident for three contracts. This finding may be due to the influence of extreme outliers for contracts analyzed.

[Insert Figure 2 here]

We show QQ plots of normal transform variables (e.g., Berkowitz normal transformation as we discussed in section 3.2) in Figure 2. QQ plots display empirical quantiles of the observed normal transform variables against theoretical quantiles from the normal distribution. If the distribution of normal transform is truly normal, then the QQ plot should be close to the 45-degree line. The QQ plots of Figure 2 show that for all contracts, three conditional processes fit poorly. There are too many extreme outliers in the tail of distributions, which is evidence that both tails of the normal density are too thin (except for the FTSE 100 index futures and the NIKKEI 225 index futures with the 'GARCH-t' model, the right tail of the normal density are too thick).

Table 5A reports the rejection power on three risk measures using the hypothesis-based backtests. Each panel represents a different index futures contract. The rejection power is represented by a p-value. In all cases, the desired confidence level of tests is 0.05 and based on the finite-sample critical value in Table 1.

[Insert Table 5A here]

There are a number of findings. Firstly, the frequency-based backtests can only reject the "EWMA" model. In contrast, the density forecast backtesting reject conditional models over 70% of the time. The high rejection power of the density forecast backtesting is consistent with the Monte Carlo simulation results in the previous section. Secondly, the backtesting method on SRM and ES is more effective than that of VaR in rejecting incorrect models. In particular, the density forecast backtesting on SRM reject all three conditional risk models. Thus, the evlaution performance of SRM is slightly better than ES. For instance, in the case of the ‘GARCH-normal’ model with the FTSE 100 index futures and the NIKKEI 225 index futures, no rejections are made by using ES. However, they are rejected by SRM. Thirdly, there is inconsistency results with our simulation study on distribution / or variance dynamics specifications. The rejection power is dissimilar between the ‘GARCH-normal’ model and ‘EWMA’ model. For instance, we cannot reject the ‘GARCH-normal’ model for the FTSE 100 and the Nikkei 225 index futures. However, the “EWMA” model is rejected for both index futures.

[Insert Table 5B here]

Table 5B reports sets of comparative accuracy results of QPS using the probability forecast backtesting with application to three index futures. With the ‘Lopez’ loss function, there is no consistency in rank among three models. Using the FTSE 100 and the Nikkei 225 index futures as examples, the “Lopez” loss function cannot differentiate the best model between the “GARCH-normal” model and the “GARCH-t” model. However, the ‘Down’ loss function and the ‘SRMP’ loss function are consistently ranking the ‘GARCH-t’ model as the best model. These results also confirm the previous simulation findings in the conditional setting. Clearly, the ‘Dowd’ loss function and the ‘SRMP’ loss function are capable of identifying correct model from alternative models.

6. Conclusion

We have presented evidence on the evaluation performance of four statistical tests on three risk measures – VaR, ES and SRM. We focus on two aspects of the risk model evaluation: the distribution specification and the volatility specification. While other papers have pursued evaluation effectiveness in a manner similar to ours, we believe ours to be the first to systematically investigate merits of various backtesting methods for ES and SRM.

Firstly, within hypothesis-based tests, including the binomial backtesting, the interval forecast backtesting and the density forecasts backtesting, the overall dominance of density forecast backtesting is confirmed. In particular, the backtesting for SRM and ES is more effective than for VaR in identifying an incorrect model from alternative models in a small sample setting. However, the binomial backtesting and the interval forecast backtesting cannot detect model failure with a small sample size, such as 250 observations as recommended by the Basel II. Therefore, the likelihood of misclassifying incorrect risk models as correct is lower for the density forecast backtesting than for frequency-based backtests.

Secondly, we propose a loss function for SRM where the probability forecast backtesting is capable of identifying accurate models from alternative models - that is, the quadratic probability score for the true model is lower than that of alternative models. In the conditional setting, the loss function for ES can only differentiate the accurate risk model when a sample size reaches or exceeds 500 observations, and the loss function for VaR cannot differentiate the correct risk model even with 1000 observations.

Thirdly, in all of the backtesting methods examined, the choice of the distribution specification is a more important factor in determining the evaluation performance than the choice of the volatility specification. That is, all four methods are capable of differentiating between the true model and alternative models with the same variance dynamics but different distributional assumptions.

References

Acerbi, C., 2004. Coherent representations of subjective risk-aversion. In: Szegö, G. (Ed.), Risk Measures for the 21st Century. John Wiley.

Acerbi, C., 2002. Spectral measures of risk: A coherent representation of subjective risk aversion. Journal of Banking and Finance 26, 1505-1518

Agresti A. and M. Yang, 1987. An empirical investigation of some effects of sparseness in contingency tables. Comm. Stat. 5, 9-21

Artzner, P., F. Delbaen, J.-M. Eber and D. Heath, 1999. Coherent measures of risk. Mathematical Finance 9 (3), 203-228

Artzner, P., F. Delbaen, J. Eber and D. Heath, 1997. Thinking coherently. Risk, No.10; Vol.11; Pg.68-71

Bao, Y., T.H. Lee and B. Saltoglu, 2004. A test of density forecast comparison with application to risk management. Mimeo, University of California, Riverside, and Marmora University, Istanbul.

Basak, S. and A. Shapiro, 2001. Value-at-Risk-based risk management: optimal policies and asset prices. Review of Financial Studies, 14, 371-405

Basel Committee on Banking Supervision, 2004. Overview of the Amendment to the Capital Accord to incorporate market risk, Basle II.

Basel Committee on Banking Supervision (1996a): Overview of the Amendment to the Capital Accord to Incorporate Market Risk. Bank for International Settlements.

Basel Committee on Banking Supervision (1996b): Amendment to the Capital Accord to Incorporate Market Risk. Bank for International Settlements.

Bawa, V.S., S.J. Brown and R.W. Klein, 1979. Estimation Risk and Optimal Portfolio Choice. North – Holland: New York.

Berkowitz, J., 2001. Testing Density Forecasts With Applications to Risk Management, Journal of Business and Economic Statistics, 19, 465-474.

Berkowitz, J. and J. O'Brien, 2002. How Accurate Are the Value-at-Risk Models at Commercial Banks, Journal of Finance, 57, 1093-1112.

Brooks, C., A.D. Clare, J.W. Dalle Molle and G. Persand, 2005. A comparison of extreme value theory approaches for determining value at risk. Journal of Emprical Finance 12, 339-352

Campbell, S.D., 2005. A Review of Backtesting and Backtesting Procedures. Finance and Economics Discussion Series. Divisions of Research & Statistics and Monetary Affairs. Federal Reserve Board, Washington, D.C.

Carmona, C.C., 2005. A review of forecasting theory using generalized loss functions. Working paper, University of California, San Diego.

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

Christoffersen, P. and D. Pelletier, 2004. Backtesting Value-at-Risk: A Duration-Based Approach. Journal of Financial Econometrics 2, 84-108

Christoffersen, P., 1998, Evaluating Interval Forecasts. International Economic Review, 39, 841-862.

Clement, M. P., and N. Taylor, 2002. Evaluating interval forecasts of high-frequency financial data. Working paper.

.

Cotter, J. and K. Dowd, 2006. Extreme Spectral risk measures: An application to futures clearinghouse margin requirements. Journal of Banking and Finance, Forthcoming.

Crnkovic C. & Drachman, J., 1997, “Quality Control”, in VaR: Understanding and Applying Value-at-Risk, London, Risk Publications.

Danielsson, J., 2000. (Un) Conditionality of Risk Forecasting. Mimeo, London School of Economics.

Diebold, F.X., T. Gunther and A. Tay, 1998. Evaluating Density Forecasts with Applications to Financial Risk Management, International Economic Review, 39, 863-883.

Diebold, F. X., A.S. Tay, and K.F. Wallis, 1999. Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In Engle, R. F., & White, H. (Eds.), Cointegration, causality, and forecasting: a festschrift in honour of Clive W. J. Granger. Oxford: Oxford University Press, 76–90.

Dowd, K., 2004. Measuring Market Risk. 2nd Edition, Wiley Finance, New York

Dowd, K., 2002. A Bootstrap Backtest, Risk, Vol.15 (10), 93-94.

Dowd, K., 1998. Beyond Value at Risk. Wiley, New York

Duan, J.C., 2003. A specification test for time series models by a normality transformation. Mimeo. Rotman School of Management. University of Toronto.

Duffie, D. and J. Pan, 1997. An overview of value at risk. The Journal of Derivatives 4, pg. 7-49

Gneiting, T. and A.E. Raftery, 2005. Strictly Proper Scoring Rules, Prediction, and Estsimtion. Technical Report No. 463R. University of Washington.

Granger, C.W.J. and M.H. Pesaran, 2000a. A Decision Theoretic Approach to Forecast Evaluation. In W.S.Chon, W.K.Li, and H.Tong (eds.) Statistics and Finance: An Interface, Imperial College Press, London, 261-278

Granger, C.W.J. and M.H. Pesaran, 2000b. Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting 19, 537-560.

Jarque, C. M. and Bera, A. K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6, 255–259.

Jorion, P., 1997. Value at Risk. McGraw-Hill, New York

J.P. Morgan, 1995. RiskMetrics Technical Document, Third Edition. New York: JP Morgan.

Kerkhof , J and B. Melenberg, 2002. Backtesting for Risk-Based Regulatory Capital. Working paper, CertER.

Kupiec, N.H., 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives 3, 73-84

Lawford, S., 2005. Finite-Sample Quantiles of the Jarque-Bera Test. Applied Economics Letters 12, 351-354

Lopez J.A., 1999a. Regulatory Evaluation of Value-at-Risk Models. Journal of Risk 1, 37-64.

Lopez J.A., 1999b, Methods for Evaluating Value-at-Risk Models. Federal Reserve Bank of SanFrancisco Economic Review 2, 3-17

Market Risk Amendment, 1997. Amendment to the capital accord to incorporate market risks, Basle

Poon, S. and C. Granger. 2003. Forecasting volatility in financial markets: A review. Journal of Economic Literature 41: 478-539.

Read, R.C. and N.A. Cressie, 1988. Goodness-of-fit statistics for discrete multivariate data. Springer-Verlag, New York

Rosenblatt, M., 1952. Remarks on a Multivariate Transformation. Annals of Mathematical Statistics 23. 470-472.

Sawa, T., 1978. Information criteria for discriminating among alternative regression models. Econometrica 46, 1273-1291

Selten, R., 1998. Axiomatic Characterization of the Quadratic Scoring Rule. Experimental Economics 1, 43-62.

Stuart, A., Ord, J. K., and Arnold, S., 1999. Kendall's advanced theory of statistics: Classical inference and the linear model (Volume 2A, 2nd Edition). New York, NY: Oxford University Press.

Szegö, G. 2002. Measures of risk. Journal of Banking & Finance 26. 1253-1272

Tasche, D., 2002. Expected shortfall and beyond. Journal of Banking and Finance 26, 1519-1533

West, K.D., 1996. Asymptotic Inference about Predictive Ability. Econometrica 64, 1067-1084

Yamai, Y. and T. Yoshiba, 2002. Comparative Analyses of Expected Shortfall and Value-at-Risk (3): Their Validity under Market Stress. Monetary and Economic Studies. October. 181-237

Table 1: Finite-Sample Critical Values

Sample Size	**125**	**250**	**500**	**1000**
Asymptotic X^2 *(1)*		**3.842**		
LR_{uc}(99)	2.513	5.025	4.813	4.091
LR_{uc}(95)	4.093	4.040	3.888	3.805
LR_{ind}(99)	0.332	0.345	2.163	2.633
LR_{ind}(95)	5.665	5.531	5.754	6.178
Asymptotic X^2 *(2)*		**5.992**		
LR_{cc}(99)	4.199	5.005	4.821	4.738
LR_{cc}(95)	7.058	7.329	7.128	6.122
ES(99)	5.576	5.645	5.961	6.073
ES(95)	6.087	6.368	5.767	5.570
JB_{SRM}	5.122	5.538	5.807	5.815
Asymptotic X^2 *(3)*		**7.815**		
LR_{SRM}	8.972	8.790	8.276	7.674

Note: The finite-sample critical values estimation is based on a minimum of 20,000 simulations, and the significant level is 5%.

1. The sample sizes are 125, 250, 500, and 1000.

2. LRuc(99) indicates the critical value of a 99% VaR in the unconditional coverage test. LRuc(95) indicates the critical value of a 95% VaR in the unconditional coverage test. LRind(99) indicates the critical value of a 99% VaR in the independent test. LRind(95) indicates the critical value of a 95% VaR in the independent test. LRcc(99) indicates the critical value of a 99% VaR in the conditional coverage test. LRcc(95) indicates the critical value of a 95% VaR in the conditional coverage test. ES(99) indicates the critical value of a 99% ES in the density forecast backtesting. ES(95) indicates the critical value of a 95% ES in the density forecast backtesting. JB_{SRM} indicates the critical value of the Jarque and Bera test on a SRM in the density forecast backtesting. LR_{SRM} indicates the critical value of the likelihood ratio test on a SRM in the density forecast backtesting.

Table 2A: The Hypothesis-based Backtesting Techniques - Left Tail
Size and Power: True DGP - T(6) model

size

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.057	0.030	0.066	0.030	0.049	0.051	0.036	0.051	0.043	0.057
250	0.016	0.041	0.066	0.036	0.054	0.053	0.052	0.040	0.026	0.043
500	0.047	0.050	0.050	0.070	0.053	0.052	0.058	0.054	0.052	0.046
1000	0.048	0.068	0.065	0.074	0.045	0.106	0.053	0.072	0.067	0.050

Power: Normal

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.035	0.029	0.156	0.082	0.047	0.055	0.575	0.675	0.071	0.518
250	0.074	0.061	0.189	0.107	0.050	0.050	0.695	0.729	0.218	0.696
500	0.128	0.088	0.064	0.162	0.054	0.057	0.880	0.923	0.397	0.726
1000	0.225	0.130	0.056	0.265	0.046	0.136	0.937	0.998	0.860	0.938

Power: Skew-t

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.338	0.290	0.036	0.231	0.055	0.232	0.582	0.608	0.615	0.048
250	0.352	0.422	0.053	0.391	0.050	0.364	0.704	0.718	0.725	0.049
500	0.603	0.747	0.071	0.631	0.065	0.638	0.765	0.804	0.963	0.068
1000	0.881	0.955	0.067	0.894	0.037	0.940	0.889	0.982	0.999	0.088

Power: t(4)

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.041	0.053	0.051	0.024	0.053	0.064	0.018	0.092	0.037	0.050
250	0.009	0.060	0.046	0.025	0.052	0.078	0.038	0.128	0.023	0.047
500	0.063	0.097	0.045	0.085	0.059	0.091	0.154	0.301	0.075	0.078
1000	0.070	0.173	0.060	0.097	0.056	0.206	0.409	0.583	0.144	0.151

Power: GPD

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.053	0.032	0.065	0.032	0.050	0.052	0.057	0.036	0.042	0.054
250	0.013	0.043	0.061	0.034	0.049	0.048	0.044	0.043	0.028	0.046
500	0.052	0.054	0.048	0.073	0.053	0.051	0.071	0.063	0.056	0.053
1000	0.046	0.064	0.059	0.074	0.050	0.111	0.056	0.074	0.061	0.055

Notes:
1. The Table compares the Monte Carlo rejection rate of alternative backtesting methods - the binomial test, the interval forecast backtesting, and the density forecast backtesting, over 2000 simulations with four different sample sizes - 125, 250, 500 and 1000 observations. In each simulation, a data generating process is comprised of an unconditional student's t distribution with 6 degree of freedoms, labeled "t(6)" model. The alternative risk models are an unconditional student's t distribution with 4 degree of freedoms, labeled "t(4)" model, an unconditional normal distribution, labeled "Normal" model, an unconditional Skew-t distribution, labeled "Skew-t" model and an unconditional Generalized Pareto Distribution, labeled "GPD" model.

2. ‘Size’ indicates that the forecast model coincides with the true model "t(6)" and the null hypothesis is therefore true.
3. The panels labeled ‘power’ display size-adjusted rejection rates for alternative forecast models. The size-adjected rejection rates are calculated based on the finite-sample critical value in Table 1. For the backtesting procedures, the desired size is 0.05.

4. The first two columns report the rejection rates of a 99% VaR and a 95% VaR using the binomial test. 'UC' indicates the unconditional coverage test in the binomial backtesting. Columns 3 and 4 report the rejection rates of a 99% VaR and a 95% VaR using the interval forecast backtesting. 'IND' indicates the independent test, and 'CC' indicates the conditional coverage test. The last four columns present the rejection rates of a 99% ES, a 95% ES and a SRM in the density forecast backtesting. 'LR' indicates the likelihood ratio test and 'JB' indicates the Jarque and Bera test. LR statistic tests the first two moments that depart from normality and JB statistic tests the third and forth moments that depart from normality.

Table 2B: The Probability Forecast Backtesting - Left Tail Size and Power: True DGP - T(6) model

	Lopez		Dowd		SRMP
Sample size:125	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
t (6)	0.020	0.095	0.020	0.072	0.040
Normal	0.035	0.093	0.032	0.083	0.077
t (4)	0.018	0.120	0.021	0.073	0.047
Skew-t	0.086	0.361	0.041	0.112	0.133
GPD	0.020	0.097	0.021	0.074	0.047

Sample size:250	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
t (6)	0.019	0.094	0.022	0.074	0.023
Normal	0.038	0.096	0.034	0.085	0.065
t (4)	0.018	0.134	0.023	0.075	0.024
Skew-t	0.126	0.574	0.042	0.112	0.093
GPD	0.019	0.094	0.023	0.074	0.024

Sample size:500	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
t (6)	0.020	0.093	0.021	0.074	0.012
Normal	0.051	0.104	0.034	0.087	0.057
t (4)	0.022	0.096	0.022	0.075	0.016
Skew-t	0.211	0.991	0.048	0.127	0.076
GPD	0.020	0.092	0.023	0.075	0.013

Sample size:1000	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
t (6)	0.020	0.101	0.022	0.074	0.007
Normal	0.072	0.135	0.034	0.075	0.051
t (4)	0.021	0.223	0.023	0.085	0.013
Skew-t	0.387	1.859	0.043	0.114	0.067
GPD	0.020	0.101	0.022	0.258	0.007

Notes:

1. The Table reports sets of comparative accuracy results of the Quadratic probability score (QPS) using the probability forecast backtesting over 2000 simulations with four different sample sizes - 125, 250, 500 and 1000 observations. In each simulation, a data generating process is comprised of an unconditional student's t distribution with 6 degree of freedoms, labeled "t(6)" model. The alternative risk models are an unconditional student's t distribution with 4 degree of freedoms, labeled "t(4)" model, an unconditional normal distribution, labeled "Normal" model, an unconditional Skew-t distribution, labeled "Skew-t" model and an unconditional Generalized Pareto Distribution, labeled "GPD" model.

2. The first and second columns report the QPS of a 99% VaR and a 95% VaR. The QPS calculation is based on the loss function that suggested by Lopez (1999a), and labeled as ‘Lopez’. The third and fourth columns represent the QPS of a 99% ES and a 95% ES. The QPS calculation is based on the loss function that proposed by Dowd (2004), labeled as ‘Dowd’. The last column reports the QPS of a SRM. The QPS calculation is based on the loss function that we suggest in this paper, and label as "SRMP".

3. Quadratic probability score (QPS) has a negative orientation (small values indicate more accurate forecast).

Table 3A: The Hypothesis-based Backtesting Techniques - Left Tail
Size and Power: True DPG - GARCH-t(6) Model

size

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.028	0.049	0.027	0.026	0.046	0.039	0.017	0.042	0.064	0.068
250	0.053	0.060	0.030	0.031	0.065	0.049	0.016	0.050	0.076	0.105
500	0.056	0.070	0.150	0.042	0.111	0.067	0.016	0.046	0.092	0.224
1000	0.063	0.076	0.235	0.061	0.231	0.083	0.009	0.027	0.099	0.484

Power: Homoskedastic - t(6)

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.207	0.324	0.144	0.165	0.038	0.234	0.209	0.428	0.675	0.082
250	0.276	0.491	0.226	0.213	0.046	0.396	0.330	0.581	0.823	0.116
500	0.479	0.576	0.290	0.409	0.090	0.543	0.619	0.676	0.960	0.219
1000	0.626	0.698	0.317	0.580	0.216	0.672	0.722	0.806	0.979	0.459

Power: GARCH - normal

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.346	0.450	0.071	0.246	0.033	0.450	0.527	0.627	0.905	0.467
250	0.500	0.567	0.310	0.450	0.067	0.540	0.747	0.767	0.975	0.533
500	0.517	0.733	0.346	0.621	0.100	0.633	0.867	0.870	0.999	0.833
1000	0.692	0.767	0.692	0.692	0.233	0.733	0.933	0.980	1.000	0.967

Power: EWMA

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES(99%)	ES (95%)	SRM	
Sample size	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
125	0.344	0.427	0.184	0.383	0.052	0.429	0.584	0.628	0.949	0.340
250	0.455	0.635	0.288	0.412	0.054	0.548	0.747	0.844	0.986	0.524
500	0.483	0.729	0.383	0.512	0.106	0.689	0.871	0.908	1.000	0.773
1000	0.608	0.738	0.438	0.661	0.228	0.815	0.931	0.994	1.000	0.957

Notes:

1. The Table compares the Monte Carlo rejection rate of alternative backtesting methods - the binomial test, the interval forecast backtesting, and the density forecast backtesting, over 2000 simulations with four different sample sizes - 125, 250, 500 and 1000 observations. In each simulation, a data generating process is comprised of a GARCH(1,1) model with t(6) innovations, labeled "GARCH-t(6)" model. The alternative risk models are a GARCH(1,1) model with normal innovations, labeled "GARCH-normal" model, a homoskedastic model with t(6) innovations, labeled "Homoskedasitic-t(6)" model, and an Exponential Weight Moving Average model, labled "EWMA" model.

2. 'Size' indicates that the forecast model coincides with the true model "t(6)" and the null hypothesis is therefore true.

3. The panels labeled 'power' display size-adjusted rejection rates for alternative forecast models. The size-adjected rejection rates are calculated based on the finite-sample critical value in Table 1. For the backtesting procedures, the desired size is 0.05.

4. The first two columns report the rejection rates of a 99% VaR and a 95% VaR using the binomial test. 'UC' indicates the unconditional coverage test in the binomial backtesting. Columns 3 and 4 report the rejection rates of a 99% VaR and a 95% VaR using the interval forecast backtesting. 'IND' indicates the independent test, and 'CC' indicates the conditional coverage test. The last four columns present the rejection rates of a 99% ES, a 95% ES and a SRM in the density forecast backtesting. 'LR' indicates the likelihood ratio test and 'JB' indicates the Jarque and Bera test. LR statistic tests the first two moments that depart from normality and JB statistic tests the third and forth moments that depart from normality.

Table 3B: The Probability Forecast Backtesting - Left Tail Size and Power: True DPG - GARCH -t(6) Model

	Lopez		Dowd		SRMP
Sample size:100	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
GARCH - t(6)	0.123	0.245	0.184	0.353	0.170
Homoskedastic-t(6)	0.098	0.384	0.163	0.394	0.187
GARCH - normal	0.150	0.254	0.215	0.372	0.352
EWMA	0.125	0.230	0.190	0.333	0.361
Sample size:250	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
GARCH - t(6)	0.199	0.709	0.214	0.468	0.144
Homoskedastic-t(6)	0.147	0.657	0.183	0.414	0.170
GARCH - normal	0.512	0.790	0.358	0.582	0.456
EWMA	0.472	0.700	0.344	0.559	0.461
Sample size:500	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
GARCH - t(6)	0.302	0.977	0.175	0.352	0.107
Homoskedastic-t(6)	0.174	0.893	0.177	0.410	0.176
GARCH - normal	1.607	2.300	0.517	0.789	0.725
EWMA	1.993	2.722	0.559	0.836	0.960
Sample size:1000	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
GARCH - t(6)	0.350	1.126	0.112	0.387	0.085
Homoskedastic-t(6)	0.214	1.111	0.166	0.402	0.136
GARCH - normal	3.159	4.464	0.539	0.821	0.746
EWMA	4.319	5.835	0.595	0.884	1.083

Notes:

1. The Table reports sets of comparative accuracy results of the Quadratic probability score (QPS) using the probability forecast backtesting over 2000 simulations with four different sample sizes - 125, 250, 500 and 1000 observations. In each simulation, a data generating process is comprised of a GARCH(1,1) model with t(6) innovations, labeled "GARCH-t(6)" model. The alternative risk models are a GARCH(1,1) model with normal innovations, labeled "GARCH-normal" model, a homoskedastic model with t(6) innovations, labeled "Homoskedasitic-t(6)" model, and an Exponential Weight Moving Average model, labled "EWMA" model.

2. The first and second columns report the QPS of a 99% VaR and a 95% VaR. The QPS calculation is based on the loss function that suggested by Lopez (1999a), and labeled as ‘Lopez’. The third and fourth columns represent the QPS of a 99% ES and a 95% ES. The QPS calculation is based on the loss function that proposed by Dowd (2004), labeled as ‘Dowd’. The last column reports the QPS of a SRM. The QPS calculation is based on the loss function that we suggested in this paper.

3. Quadratic probability score (QPS) has a negative orientation (small values indicate more accurate forecast).

Table 4

Summary Statistics for Daily Returns of Futures Series

	S&P 500	*FTSE 100*	*NIKKEI 225*
Mean	0.000	0.000	0.000
Standard Deviation	0.011	0.012	0.014
Skewness	0.044	-0.169	-0.105
Kurtosis	5.455	5.869	4.854
Kolmogorov - Smirnov	0.483	0.481	0.481
P - Value	0.000	0.000	0.000

Note: The summary statistics are presented for each futures index. With the exception of skewness and kurtosis coefficients, all values are expressed in percentage form. The skewness statistic is a measure of distribution asymmetry with symmetric returns having a value of zero. The kurtosis statistic measures the shape of a distribution vis-á-vis a normal distribution with Gaussian density function having a value of three. Normality is formally examined with Kolmogorov-Smirnov test which indicates a null hypothesis of normality is rejected at standard confidence levles for all series.

Table 5A: Power of Hypothesis-based Backtesting Techniques - Left Tail

S&P 500

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES (99%)	ES(95%)	SRM	
	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
GARCH-normal	0.217	0.711	0.081	0.298	4.973	6.684	6.185**	5.811**	8.328**	4.262
GARCH - t	0.943	1.127	0.048	0.991	5.697	6.824	5.676	5.995**	9.246**	3.398
EWMA	5.419**	4.365**	1.474	6.893**	6.208**	7.573**	5.967**	6.839**	8.611**	5.658

FTSE 100

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES (99%)	ES(95%)	SRM	
	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
GARCH-normal	0.000	0.394	0.121	0.121	3.219	3.613	0.489	0.669	8.737**	3.320
GARCH - t	0.000	0.043	0.121	0.121	2.452	2.495	1.955	4.160	11.284**	1.461
EWMA	2.613	3.992**	2.163**	4.875**	0.890	1.882	5.982**	6.357**	8.534**	5.531

NIKKEI 225

	Binomial Test		Interval Forecast Backtesting				Density Forecast Backtesting			
	VaR(99%)	VaR(95%)	VaR(99%)		VaR(95%)		ES (99%)	ES(95%)	SRM	
	UC	UC	IND	CC	IND	CC	LR	LR	LR	JB
GARCH-normal	0.190	2.277	0.170	0.360	2.092	4.368	1.561	5.253	10.591**	14.848**
GARCH - t	0.217	0.711	0.081	0.298	1.249	1.960	1.411	8.641**	15.051**	4.566
EWMA	1.538	0.394	2.598**	4.137	1.030	1.424	12.334**	14.221**	8.423**	26.381**

Notes:

1. The Table reports the Likelihood values of three hypothesis-based backtesting methods - the binomial test, the interval forecast backtesting and the density forecast backtesting based on three futures index contracts - the S&P 500, the FTSE 100 and the NIKKEI 225. The data generate process are comprised of a GARCH(1,1) model with t innovations, labeled as "GARCH-t" and a GARCH(1,1) model with normal innovations, labeled as "GARCH-normal", and an EWMA model, labeled as "EWMA".
2. The first two columns report the likelihood value of a 99% VaR and a 95% VaR using the binomial test. 'UC' indicates the unconditional coverage test in the binomial backtesting. Columns 3 and 4 report the likelihood value of a 99% VaR and a 95% VaR using the interval forecast backtesting. 'IND' indicates the independent test, and 'CC' indicates the conditional coverage test. The last four columns present the likelihood value of a 99% ES, a 95% ES and a SRM in the density forecast backtesting. 'LR' indicates the likelihood ratio test and 'JB' indicates the Jarque and Bera test. LR statistic tests the first two moments that depart from normality and JB statistic tests the third and forth moments that depart from normality.
3. "**" represents the null hypothesis is rejected at 5% significant level, based on the finite-sample critical value at Table 1.

Table 5B: The Probability Forecast Backtesting - Left Tail

S&P 500	**Lopez**		**Dowd**		**SRMP**
	VaR(99%)	VaR(95%)	ES(99%)	ES(95%)	SRM
GARCH-normal	0.004	0.004	0.085	0.085	0.080
GARCH - t	0.016	0.016	0.036	0.036	0.056
EWMA	0.144	0.144	0.246	0.246	0.096
FTSE 100					
GARCH-normal	0.000	0.000	0.095	0.095	0.064
GARCH - t	0.000	0.000	0.064	0.064	0.055
EWMA	0.064	0.064	0.341	0.341	0.076
NIKKEI 225					
GARCH-normal	0.004	0.004	0.807	0.807	0.112
GARCH - t	0.004	0.004	0.304	0.304	0.096
EWMA	0.036	0.036	1.554	1.554	0.120

Notes:

1. The Table reports sets of comparative accuracy results of the Quadratic probability score (QPS) for three index futures contracts - the S&P 500, the FTSE 100 and the NIKKEI 225, based on the probability forecast backtesting. The data generate process are comprised of a GARCH(1,1) model with t innovations, labeled as "GARCH-t", a GARCH(1,1) model with normal innovations, labeled as "GARCH-normal' and an EWMA model, labeled as "EWMA".

2. The first and second columns report the QPS of a 99% VaR and a 95% VaR. The QPS calculation is based on the loss function that suggested by Lopez (1999a), and labeled as ‘Lopez’. The third and fourth columns represent the QPS of a 99% ES and a 95% ES. The QPS calculation is based on the loss function that proposed by Dowd (2004), labeled as ‘Dowd’. The last column reports the QPS of a SRM. The QPS calculation is based on the loss function that we suggested in this paper.

3. Quadratic probability score (QPS) has a negative orientation (small values indicate more accurate forecast).

Figure 1

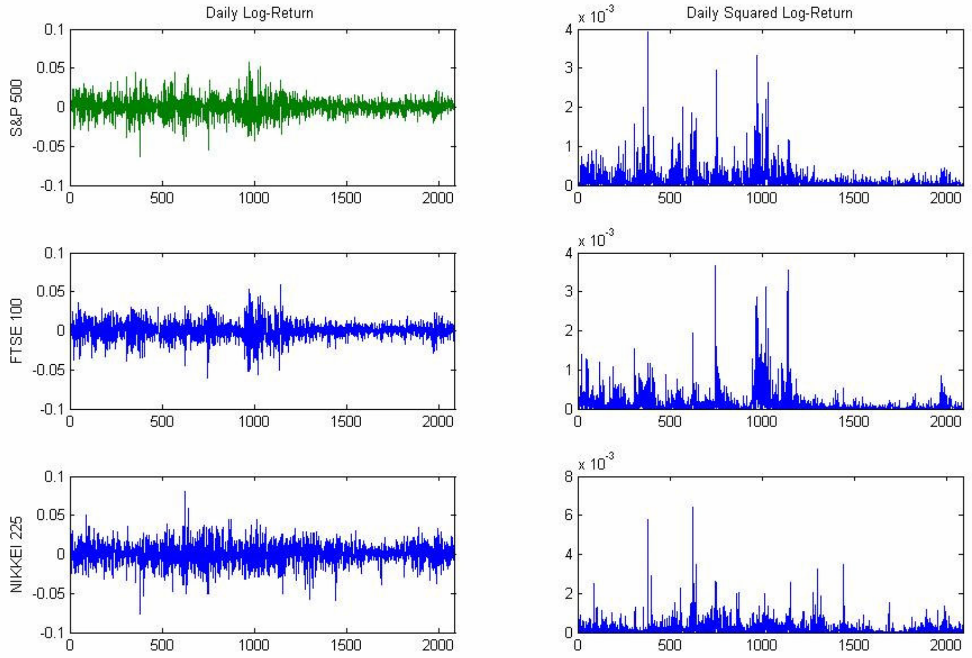

Figure 1 is a graphical representation of time series properties of daily logarithmic price changes in the three index futures and squared daily logarithmic price changes from the period 01/11/1998 through 31/10/2006.

Figure 2

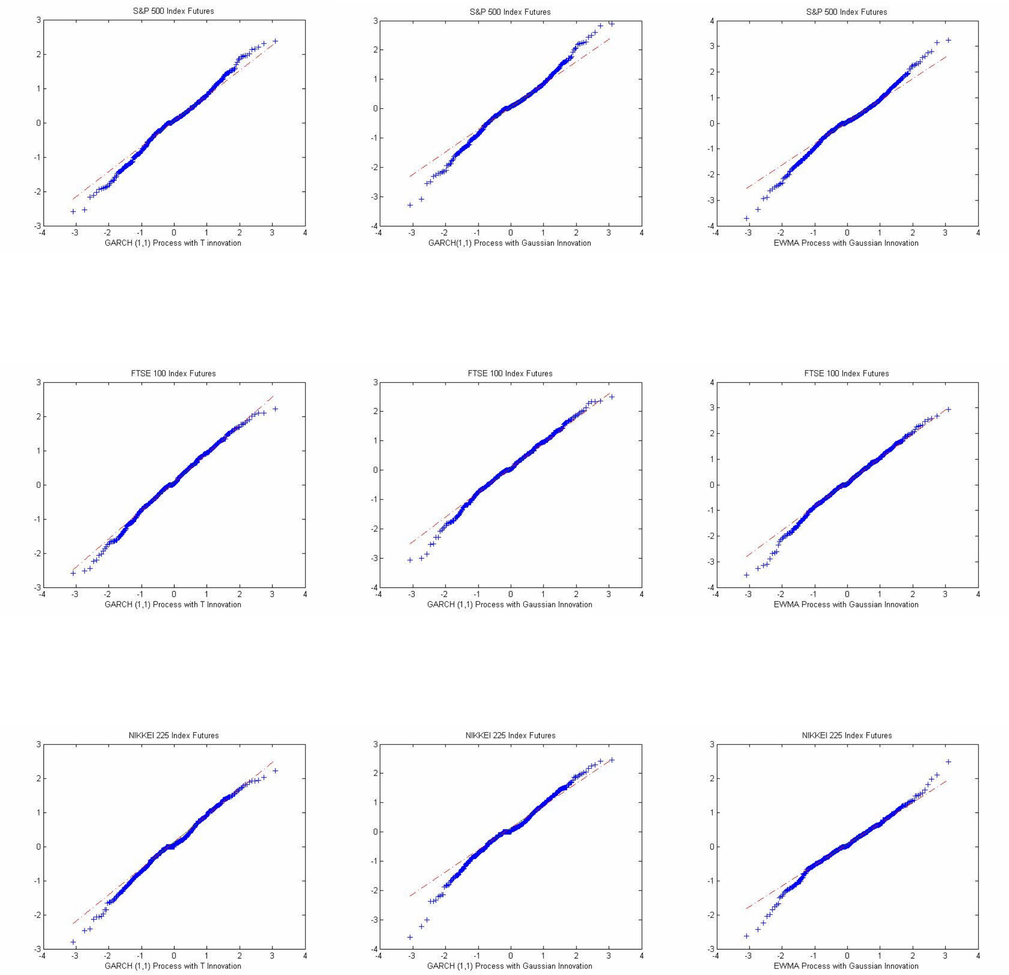

Figure 2 shows that the QQ plots of normal transform variables (*i.e.*, using Berkowitz normal transformation, as discussed in section3.2) of GARCH-t(6) process, GARCH-normal process and EWMA process with the three index futures contract. For each index futures, we scatter plot the empirical quantile of the normal transform variable from the t density forecast, and normal density forecast against the corresponding quantile of the normal distribution. The diagonal line denotes a perfect fit.