Backtesting trading risk of commercial banks using expected shortfall

Woon K. Wong

Investment Management Research Unit, Cardiff Business School, Aberconway Building, Colum Drive, Cardiff CF10 3EU, United Kingdom

Received 8 January 2007; accepted 19 November 2007
Available online 19 February 2008

Abstract

This paper uses saddlepoint technique to backtest the trading risk of commercial banks using expected shortfall. It is found that four out of six US commercial banks have excessive trading risks. Monte Carlo simulation studies show that the proposed backtest is very accurate and powerful even for small test samples. More importantly, risk managers can carry out the proposed backtest based on any number of exceptions, so that incorrect risk models can be promptly detected before any further huge losses are realized.

© 2007 Elsevier B.V. All rights reserved.

JEL classification: G32

Keywords: Value-at-Risk; Expected shortfall; Backtesting; Saddlepoint technique

1. Introduction

Since Value-at-Risk (VaR) was sanctioned by the Basle Committee in 1996 for market risk capital requirement through the so called internal models, VaR has become the standard measure for financial market risk. Major commercial banks and other financial entities such as hedge funds disclose VaR on a regular basis. Also, risk management consulting firms and software systems have been set up to meet the demand generated by the VaR-based risk management. Despite the popularity of VaR, many commercial banks and other financial institutions suffered severe trading losses during the fall of 1998. This prompted a report by the BIS Committee on the Global Financial System in 1999, in which it is remarked that “last autumn’s events were in the ‘tails’ of distributions and that VaR models were useless for measuring and monitoring market risk.” Such risk in the use of VaR is termed the “tail risk” by the literature; see Yamai and Yoshiba (2005). Given the role of commercial banks as principal dealers in the ever growing over-the-counter derivatives markets, their trading accounts have grown rapidly and become progressively more complex. Though most of them are hedged, during times of extreme market movements, correlations between various derivatives and their hedged counterparts may break down, resulting in the failure of VaR models in measuring and monitoring market risk; see for example Jorion (2000). On the other hand, Longin and Solnik (2001) and Campbell et al. (2002) provide evidence of increased correlation in bear markets, with reduced diversification and more heavy capital losses. Thus it is important to use a risk measure that takes into account the extreme losses beyond VaR, and expected shortfall (ES) proposed by Artzner et al. (1997, 1999) is an alternative risk measure that can remedy this shortcoming of VaR.

Essentially, VaR is a quantile measure which, at a given confidence level, describes the loss that can occur over a given period due to exposure to market risk. ES is the...
expected loss conditional on the loss being above the VaR level. Despite the fact that ES is a coherent risk measure whereas VaR is not subadditive, the former is absent in Basel II. One main reason for this is that the backtest of ES is harder than that of VaR; see Kerkhof and Melenberg (2004, footnote 8). For instance, the test statistics of existing ES backtests, namely the censored Gaussian approach of Berkowitz (2001) and the functional delta approach of Kerkhof and Melenberg (2004), rely on large sample for convergence to the limiting distributions. This is undesirable because test samples tend to be small in practice and regulation requires backtest to be carried out on past 250 observations. In this paper, I propose to use saddlepoint or small sample asymptotic technique to compute under the null hypothesis the required p-value based on sample expected shortfall. The advantage of the saddlepoint technique is that the required p-value can be calculated accurately for any given number of exceptions under the null hypothesis. This is of great significance because the Capital Accord stipulates VaR at 99% level, which implies that exceptions rarely occur in practice. Monte Carlo simulation study confirms that the proposed technique is not only very accurate but also very powerful even for small samples.

Current VaR backtesting as stipulated by the Basle Committee focuses on checking number of violations of the VaR level. While it is advantageous for such backtesting procedure to be applicable to any distribution, it ignores valuable information conveyed by the sizes of losses exceeding the VaR level, and thus lacks statistical test power. Indeed, the report by the Basle Committee for Banking Supervision (1996b, page 5) acknowledges that, “tests of this type are limited in their power to distinguish an accurate model from an inaccurate model... This limitation has been a prominent consideration in the design of the framework presented here, and should also be prominent among the considerations of national supervisors in interpreting the results of a bank’s backtesting program.” Therefore, various methods of ES backtesting, noticeably by Berkowitz (2001) and Kerkhof and Melenberg (2004), have been proposed. Though parametric assumptions for the null distributions are made in the ES backtests, considerable powers have been gained. The proposed backtest of ES using saddlepoint technique, however, have several advantages over these two existing approaches. First, as mentioned above, both existing ES backtests rely on asymptotic test statistics that might be inaccurate when sample size is small. If the Basle Committee were to employ the backtest of ES to determine the risk capital requirement, such lack of accuracy might be unacceptable because banks may be penalized based on incorrect inference. Second, not only Monte Carlo simulation study shows that the saddlepoint backtest has correct size, it is also more powerful than the two existing ES backtests. Finally, perhaps the most important of all from a practical risk management viewpoint, the proposed saddlepoint technique allows one to compute the required p-value conditional on the number of exceptions instead of the size of the full sample. This means that it is possible to detect failure of a risk model based on just one or two exceptions before any more data are observed. This is extremely useful since prompt action is often required in order to avert extreme financial losses due to market risk.

The proposed backtest of ES is used to test for non-normal or excessive trading risk at the six commercial banks studied by Berkowitz and O’Brien (2002). For the sample banks, internal VaR is calculated using structural model which takes into consideration all positions in the trading portfolio. Since the saddlepoint technique is conditional on number of realized exceptions but not on sample size, ES backtests based on the bank’s internal VaR model are carried out for the full sample period from January 1998 to March 2000 as well as for the three months sub-sample period from August 1998 to October 1998 (when the financial markets were highly volatile). In both cases, four out of six commercial banks’ trading accounts are found to have sample ES that are significantly larger than a normal null hypothesis. This serves as a warning for risk modeling based on normal distribution which is quite common in practice; see for example Lotz and Stahl (2005).

This paper is organized as follows: Section 2 introduces the sample ES statistic for backtesting and the saddlepoint technique to calculate the p-value of the proposed backtest is derived in Section 3. Section 4 presents the Monte Carlo simulation results whereas Section 5 applies the proposed backtest to the trading accounts at six large banks in US. Section 6 briefly illustrates how the proposed ES backtesting could help risk management to be more responsive. Finally, Section 7 concludes.

2. Backtesting expected shortfall for tail risk

In this section, some preliminary definitions of VaR and ES are first provided, followed by the definition of the sample ES statistic to be used in backtesting. Difficulties of existing methods for backtesting ES and the motivation for the use of small sample asymptotic technique are explained.

2.1. Some preliminary definitions

Let R be the random profit or loss of a portfolio over a holding period. For simplicity, R is assumed to have a
cumulative distribution function (CDF) that is absolutely continuous. Then for \(0 < x < 1\), VaR is defined as the maximum potential loss realized by \(R\) over the holding horizon at \((1-x)\) confidence level. If \(F(r)\) is the CDF of \(R\), then a formal definition of VaR can be written as

\[
\text{VaR} = -q(x) = -F^{-1}(x).
\]

That is, VaR may be regarded as the negative of left \(x\)-quantile of the distribution, which will be simply denoted as \(q\) from now onwards. Let \(f(r)\) denotes the probability density function (PDF) of \(R\). ES at \((1-x)\) confidence level is then defined by

\[
\text{ES} = -E(R|R < q) = -x^{-1} \int_{-\infty}^{q} r f(r) \, dr.
\]

ES is appealing in that it sums all values of \(r\), weighted by \(f(r)\), from minus infinity to \(q\), thus takes into account of the sizes of losses beyond the VaR level. Economically it tells us the average of losses when VaR is violated. The idea of ES for risk management first appears in Rappoport (1993) in the financial industry. It is formally proposed by Artzner et al. (1997) as a coherent risk measure.

### 2.2. Sample statistic for backtesting expected shortfall

Now I shall consider the sample ES statistic which will be used for backtesting. Given a sample \(R_1, \ldots, R_T\) with \(n > 0\) exceptions, I use the simple average as an estimator of the true ES:

\[
\text{ES}_n = -\frac{1}{n} \sum_{i=1}^{n} R_{(i)},
\]

where for \(i = 1, \ldots, T-1\), \(R_{(i)}\) is the order statistic such that \(R_{(i)} \leq R_{(i+1)}\), and the subscript \(n\) denotes the fact that the sample expected shortfall is based on \(n\) observations. Given a sample of size \(T\), one may set \(n = [T \alpha]\)\(^5\) if ES is to be estimated ex post using a fixed number of exceedances. However, such definition is inconsistent with the current VaR backtesting procedure stipulated by the Basle Committee in which an exception is realized if the loss is larger than the associated VaR forecast. Since VaR as a measure for risk is already a common usage, it would be sensible for backtesting purpose to measure ES as the mean of exceptions in accordance to the Capital Accord.

Therefore \(n\) in (1) refers to the number of VaR violations that varies from sample to sample. \(n\) is moreover a small number since the Basle Committee sets \(\alpha = 0.01\) as the day-to-day smaller fluctuations are of less concern. Backtest of ES is thus rendered to be difficult by the fact that a breach of VaR is rare in practice. Though the functional delta approach\(^6\) advocated by Kerkhof and Melenberg (2004) improves somewhat the convergence speed of (1) to its limiting distribution, backtests of risk models are often required to be carried out on small samples. This would be undesirable if the Basle Committee were to use the backtest of ES to determine the risk-based capital requirement. Raising the multiplication factor based on inaccurate \(p\)-value will not be fair to the banks.

Similar small sample problem also occurs in the case of the censored Gaussian likelihood ratio approach\(^7\) suggested by Berkowitz (2001). Moreover, this approach is a two-sided test of joint hypothesis that the mean and variance conform to that of a standard normal distribution. Since the null hypothesis could be rejected on the ground that the observed tail is too thin or there is too much capital to cover the risk, (being a two-tailed test) it is not suitable for regulatory backtesting. This is because financial fluctuations by their nature are very complicated and it is likely that a true and perfect model can never be found. While a risk manager may not know the perfect model, it is feasible that she can always allow for possible inadequacy in her risk model by marking up the risk forecast. Such practice is evident from Berkowitz and O’Brien (2002) who observe that banks’ VaR forecasts are conservative. This is also in line with the one-tailed definition and the corresponding backtesting procedure of VaR by the Capital Accord; see Basle Committee on Banking Supervision (1996a).

Unlike the functional delta and censored Gaussian approaches, the proposed saddlepoint technique makes use of small sample asymptotic distribution that works reasonably well even for \(n = 1\).\(^8\) The key idea is that under fairly general conditions, a full expansion of the true density of (1) can be derived by means of steepest decent and thus the name saddlepoint; see Daniels (1954). Monte Carlo simulation study in Section 4 shows that for any realized number of exceptions \(n > 0\), the saddlepoint technique can accurately calculate the \(p\)-value of the backtest of ES. This has very important implication: in backtesting for possible breakdown of a risk model, a decision can be formally based on as little as one or two exceptions.

---

\(^5\) \([x]\) means the largest integer number less than or equal to \(x\).

\(^6\) The ES estimator given by (1) may be regarded as a “functional” of the “statistic” which refers to the given sample distribution. In the same way that Taylor’s expansion can be used to establish asymptotic normality of a differentiable function of statistics that are asymptotically normal, the functional delta approach can be used to establish the asymptotic normality of (1) since the sample distribution converges to the true distribution as sample size tends to infinity; see Shao (1998, chapter 5).

\(^7\) In this approach, all losses that are smaller than VaR are set equal to VaR, thus give rise to a censored distribution.

\(^8\) When \(n = 1\), a precise CDF can actually be obtained simply by using the conditional distribution function; see the first paragraph of Section 4.1 for the case of normal distribution.
3. Saddlepoint technique

Since the Basle Committee sets the chance of extreme loss at 1% for market risk, the number of exceptions in a year is small and usual asymptotic result approximates the true distribution poorly. This section adopts the saddlepoint technique to derive the small sample asymptotic distribution for the sample ES statistic given by (1) under a standard normal null hypothesis. Monte-Carlo simulations in Section 4 show that the technique performs very well and the proposed backtest of ES has good power in detecting non-normal trading risk or incorrect risk models.

3.1. The ES random variable and its moments

Let \( R \) be the portfolio return which has a standard normal distribution with CDF and PDF denoted by \( \Phi \) and \( \phi \), respectively. From now onwards, unless stated otherwise, \( q \) refers to the \( \alpha \)-quantile of a standard normal distribution. That is \( q = \Phi^{-1}(\alpha) \). Define a random variable \( X \) such that \( P(X \leq x) = P(R \leq x| R < q) = P(R \leq x)/P(R < q) \),

\[
P(X \leq x) = P(R \leq x| R < q) = P(R \leq x)/P(R < q),
\]

where \( x < q \). Thus \( X \) refers to the portfolio return conditional on \( R < q \) and the theoretical expected shortfall of \( R \) at confidence level \( 1 - \alpha \) is given by \( -E(X) \). Given a sample \( X_1, \ldots, X_n \) of exceedances above VaR, the sample expected shortfall can be written as

\[
ES_n = -\overline{X} = -\frac{1}{n} \sum_{i=1}^{n} X_i.
\]

The PDF of \( X \) is given by

\[
f(x) = x^{-1} \phi(x),
\]

and the corresponding CDF by

\[
F(x) = x^{-1} \Phi(x),
\]

where \( x \in (-\infty, q) \). Now consider the following results on the moments of \( f(x) \) that shall be useful for the rest of this section.

**Proposition 1.** Let \( X \) be a continuous random variable with density \( f(x) = x^{-1} \phi(x) \), \( x \in (-\infty, q) \). The moment generating function of \( X \) is then given by

\[
M(t) = x^{-1} \exp(t^2/2) \cdot \Phi(q - t),
\]

and its derivatives with respect to \( t \) are given by

\[
M'(t) = t \cdot M(t) - \exp(qt) \cdot x^{-1} \phi(q),
\]

\[
M''(t) = t \cdot M'(t) + M(t) - \exp(qt) \cdot qx^{-1} \phi(q),
\]

\[
M^{(m)}(t) = t \cdot M^{(m-1)}(t) + (m-1)M^{(m-2)}(t) - \exp(qt) \cdot q^{m-1} x^{-1} \phi(q),
\]

where \( m \geq 3 \).

**Proof.** See Appendix \( \square \)

As a corollary of Proposition 1, the mean and variance of \( X \) can be obtained easily.

**Corollary 1.** The mean of \( X \), which is negative of ES, is given by

\[
\mu_X = E(X) = -\frac{\phi(q)}{x} = -e^{-q^2/2}/x\sqrt{2\pi},
\]

and its variance is

\[
\sigma_X^2 = \text{var}(X) = 1 - \frac{q \cdot \phi(q)}{x} - \mu_X^2 = 1 - \frac{q e^{-q^2/2}}{x\sqrt{2\pi}} - \mu_X^2.
\]

**Proof.** See Appendix \( \square \)

Though \( \sqrt{n} \sigma_X^{-1}(\overline{X} - \mu_X) \) is asymptotically standard normal, the asymptotic result is hardly valid for sample sizes one would encounter in practice. In particular, \( x \) is set by the Basle Committee as 0.01 and \( n \) equals to about 2 or 3 for a year’s data. For this reason, I resort to the Lugannani and Rice (1980) formula derived by the saddlepoint technique in order to provide an accurate method to compute under the null hypothesis the required \( p \)-value for the backtest of ES.

3.2. The Lugannani and Rice formula

Using the saddlepoint technique, Lugannani and Rice (1980) provide an elegant formula that can accurately calculate the tail probability (or CDF) of the sample mean of an independently and identically distributed (IID) random sample \( X_1, \ldots, X_n \) from a continuous distribution \( F \) with density \( f \). To have an intuitive understanding of the method, we can look at the inversion formula:

\[
f_{\overline{X}}(\overline{x}) = \frac{n}{2\pi} \int_{-\infty}^{\infty} e^{i[K(t) - \alpha]} dt,
\]

where \( f_{\overline{X}}(\overline{x}) \) denotes the PDF of the sample mean and \( K(t) = \ln M(t) \) is the cumulant generating function of \( f(x) \). Then the tail probability can be written as

\[
P(\overline{X} > \overline{x}) = \int_{\overline{x}}^{\infty} f_{\overline{X}}(u) du = \frac{1}{2\pi} \int_{-\infty}^{s+i\infty} e^{i[K(t) - \alpha]} t^{-1} dt,
\]

where \( s \) is chosen to be the saddlepoint \( \sigma \) satisfying

\[
K'(\sigma) = \overline{x}.
\]

By using appropriate transformation, the elegant Lugannani and Rice formula can be obtained, which is given below without proof.\(^9\)

**Proposition 2.** Given an IID standard normal random sample \( R_1, \ldots, R_T \) with \( n \) exceptions, construct \( X_1, \ldots, X_n \) according to (2). Let \( \sigma \) be the saddlepoint such that (8) is satisfied, and define \( \eta = \sigma \sqrt{nK''(\sigma)} \) and \( \zeta = \text{sgn}(\sigma) \sqrt{2n(\sigma^2 - K(\sigma))} \), where \( \text{sgn}(s) \) equals to zero when \( s = 0 \), or takes the value

\(^9\) Readers are referred to Daniels (1987) for proof of the Lugannani and Rice formula.
of 1(-1) when $s < 0(s > 0)$. Then the tail probability of less than or equal to the sample mean $\bar{x} \neq \mu_X$\textsuperscript{10} is given by

$$P(\bar{X} \leq \bar{x}) = \begin{cases} \Phi(z) - \phi(z) \cdot \left(1 - \frac{1}{\sqrt{n}} + O(n^{-3/2})\right) & \text{for } \bar{x} < q, \\ 1 & \text{for } \bar{x} \geq q. \end{cases}$$

To obtain the tail probability, the saddlepoint $\sigma$ is first obtained by solving for $t$ in the following equation\textsuperscript{11}

$$K'(t) = \frac{M'(t)}{M(t)} = t - \exp(\gamma/2) \frac{\phi(q - t)}{\phi(q - t)} = \bar{x}. \quad (10)$$

Next, (3) and (4) are used to calculate $M(\sigma)$ and its first two derivatives at $t = \sigma$, which in turn are used to evaluate $K(\sigma)$ and $K'(\sigma)$. Then it is straightforward to calculate $\eta$ and $\zeta$ defined in Proposition 2, and the required probability.

Fig. 1 depicts the theoretical CDF of the standardized statistic $z = n^{1/2} \sigma \bar{x}^{-1}(\bar{x} - \mu_X)$ calculated using the Lugannani and Rice formula for various $n$. Since the formula is valid only when the sample mean is less than $q$, the CDF is assigned a value of 1 when $z > n^{1/2} \sigma_X^{-1}(q - \mu_X)$. As predicted by the central limit theorem, the finite sample size CDF approaches the standard normal CDF when $n$ gets large.

\textsuperscript{10} Though the case of $\bar{x} = \mu_X$ is hardly encountered in practice, the corresponding tail probability is provided for completeness:

$$P(\bar{X} \leq \mu_X) = \frac{1}{2} + \frac{K''(0)}{6\sqrt{2n}[K'(0)]^4} + O(n^{-3/2}).$$

\textsuperscript{11} The saddlepoint $\sigma$ is unique; see Daniels (1987).

3.3. Hypothesis testing

Given $n$ realized exceptions, one can compute the sample expected shortfall $ES_n = -\bar{X}$ and carry out a one-tailed backtest to check whether the sample ES is too large. The null and alternative hypotheses can be written as

$$H_0 : ES_n = ES_0 \text{ versus } H_1 : ES_n > ES_0,$$

where $ES_0$ denotes theoretical expected shortfall under the null hypothesis. Under the standard normal null hypothesis with $z = 0.01$, $ES_0 = -\mu_X = 2.6652$. The $p$-value of the hypothesis test is simply given by the Lugannani and Rice formula as

$$p-value = P(\bar{X} \leq \bar{x}). \quad (11)$$

The null hypothesis is rejected if $ES_n$ is significantly larger than 2.6652, the value of $ES_0$ under the standard normal assumption. In this case, the risk model may be described as failing to "capture" the risk or simply there is excessive risk in the empirical distribution. On the other hand, if the null hypothesis is accepted, one may say that the risk model provides sufficient risk coverage.\textsuperscript{12} Finally, it is trivial that when there is no exception, that is $n = 0$, the null hypothesis is accepted.

Strictly speaking, above hypothesis formulation tests whether a given sample of exceedances above VaR is consistent with the tail of distribution under the null hypothesis, which in our case refers to the standard normal distribution. It cannot be over emphasized the fact that it is possible to have non-normal distributions with values of expected shortfall less than those of standard normal distribution.

\textsuperscript{12} Strictly speaking, if the null hypothesis is accepted, it means that there is no evidence suggesting insufficient risk coverage.
3.4. Non-normal null hypothesis

In this paper, backtest of ES is carried out under the null hypothesis that the underlying distribution is normal. This is because normality is widely assumed in practice\(^{13}\) and, if the null hypothesis is a non-normal distribution, one can always follow the suggestion of Berkowitz (2001) and Kerkhof and Melenberg (2004) to transform the distribution into a normal one. The idea is that if the observed returns are fatter-tailed than the model presumes, the transformed data will be also fatter-tailed than the standard normal hypothesis. It must be emphasized that, however, the saddlepoint technique can also be applied directly to non-normal null hypothesis. Consider the case of a random variable \(X\) with absolutely continuous PDF \(f(x)\) and CDF \(F(x)\). Then the moment generating function is given by

\[
M(t) = \exp(qt) - t \int_{-\infty}^{\theta} F(x) \exp(tx) \, dx,
\]

and its first two derivatives are

\[
M'(t) = q \exp(qt) - \int_{-\infty}^{\theta} F(x) \exp(tx)(1 + xt) \, dx,
\]

\[
M''(t) = q^2 \exp(qt) - \int_{-\infty}^{\theta} F(x) \exp(tx)(2 + xt)x \, dx.
\]

Once the saddlepoint value is obtained using (8), the moment’s derivatives can be evaluated. Then it is straightforward for the required \(p\)-value to be computed using the Lugannani and Rice formula described in Section 3.2.

4. Monte Carlo simulations

In this section, Monte Carlo simulation experiments are carried out to investigate the empirical test sizes and powers of various backtests. Only 99% confidence level VaR and ES are considered. IID samples of various sizes under the null and various alternative hypotheses are generated and the corresponding \(p\)-values are calculated. 100,000 simulations are carried out to calculate the empirical test sizes and Bernoulli likelihood ratio tests are applied to check if they differ significantly from the theoretical test sizes. For powers of test, only rejection rates at 5% significance level based on 10,000 simulations are reported. It is found that the proposed backtest using the saddlepoint technique yields very accurate test sizes and has the most power in rejecting the false models.

4.1. Empirical test sizes of ES backtest using saddlepoint technique

Here I first investigate the empirical test sizes of ES backtesting using saddlepoint technique for number of exceptions \(n = 1, 2, 5\) and 10 under the standard normal null hypothesis. For each simulation, rejection is made if the \(p\)-value calculated using (9) is less than the various significance levels ranging from 0.5% to 5% on both sides of the distribution.\(^{14}\) Panel A of Table 1 depicts the rejection rates for both tails based on 100,000 simulations. Bernoulli likelihood ratio tests are carried out and it indicates that the empirical test size differs significantly at 1% level from its theoretical value. It can be seen that only the empirical test sizes on the far right tails show incorrect rejection rates when \(n = 1\). This is nevertheless unimportant for two reasons. First, backtest for the purpose of risk management concerns only on the left tail. Second, for \(n = 1\), one can easily obtain precise CDF by using the conditional distribution function \(z^{-1} \Phi(x)\) under the null hypothesis. From now onwards, when \(n = 1\), the corresponding \(p\) and critical value will be calculated using the conditional distribution function.

Since the number of exceedances varies from sample to sample in practice, Panel B lists the empirical test sizes for standard normal samples of sizes ranging from 125 to 1000. Only those samples with at least one exception are considered. For example, for sample of size \(T = 250\), first simulation yields 3 exceptions and the required \(p\)-value is calculated using (9) with \(n = 3\). Second simulation yields 2 exceptions and the required \(p\)-value is obtained with \(n = 2\). If the third simulation does not produce any exception, no \(p\)-value is calculated and the simulation is abandoned. When 100,000 simulations are completed, the empirical test size is obtained by dividing the number of rejections by the number of nonzero-exception simulations.\(^{15}\) The accuracy of the saddlepoint technique is again demonstrated by the results in Panel B.

Finally, Panel C considers the case when the number of exceptions is not known (for example, a risk report that discloses only the mean violations but not the number of exceptions). In this case, the \(p\)-value can be calculated as

\[
P(\bar{X} \leq \bar{x}) = B(0 | T) + \sum_{n=1}^{\infty} p(\bar{X} \leq \bar{x} | n) \cdot B(n | T),
\]

where \(n\) is the number of exceptions and \(B(\cdot | T)\) is the Binomial probability density function with 0.01 probability of success and \(T\) trials. As can be observed from Panel C, the empirical test sizes differ from their true values only at the right tail when \(T\) is small. This is because for sample sizes \(T = 125, 250, 500\) and 1000, \(B(0 | T) = 0.285, 0.081, 0.007\) and 4.32 \times 10^{-5}\), respectively. Hence as long as the theoretical test size is larger than the corresponding probability value when \(n = 0\), the empirical test sizes are very close to the theoretical ones.

---

\(^{13}\) Due to the working of central limit theorem, trading account’s returns at the bank level are often approximated by a normal distribution in practice; see Lotz and Stahl (2005).

\(^{14}\) Though only the left tail of the distribution is required for the backtesting of risk model, right tail is also included to illustrate the good approximation achieved by the saddlepoint technique.

\(^{15}\) For 99% confidence level (\(x = 0.01\)), the probability of getting a zero-exception sample is 0.99\(^{7}\) where \(T\) is sample size.
Table 1

Empirical test sizes

<table>
<thead>
<tr>
<th>Panel A: given a fixed number of VaR violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: given realized number of VaR violations, ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
</tr>
<tr>
<td>125</td>
</tr>
<tr>
<td>250</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>1000</td>
</tr>
</tbody>
</table>

Panel A and Panel B show similar behavior. Panel C: given a fixed number of observations \( T \), \( n \) unknown

<table>
<thead>
<tr>
<th>Panel C: given a fixed number of observations ( T, n ) unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
</tr>
<tr>
<td>125</td>
</tr>
<tr>
<td>250</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>1000</td>
</tr>
</tbody>
</table>

For the study of test powers, 10,000 simulations are used and the rejection rates reported in Panels B, C and D of Table 2 are all based on 5% significance level. In practice, financial time series are found to exhibit excess kurtosis, have longer left tails and time-varying conditional variances. I use three non-normal processes to replicate this behavior. First, Panel B considers the case of Student \( t \) distribution with 5 degrees of freedom. Student \( t \) distribution is well known for its heavy tails but is symmetric. Next in Panel C, the IID NIG samples have excess kurtosis and are skewed to the left. The parameters of the NIG distribution are \( \alpha = 0.8 \), \( \beta = -0.05 \), \( \delta = \alpha (1 - \rho)^{3/2} \) and \( \mu = -\rho \delta (1 - \rho)^{1/2} \) where \( \rho = \beta / \alpha \). Finally in Panel D, the time-varying conditional variance of financial returns are simulated by a GARCH(1,1) process \( R_t = \sigma_t Z_t \) with volatility equation given by \( \sigma_t^2 = 0.05 + 0.15 \rho_i^2 + 0.8 \sigma_i^2 \) and \( Z_t \) distributed as IID \( N(0, 1) \). Generally speaking, the ES-based backtests are more powerful in detecting incorrect risk models, and amongst the ES-backtests, the saddlepoint technique is the most powerful. Though the Kupiec’s likelihood ratio test is the most powerful in detecting non-normal tail risk when the underlying process is GARCH(1,1) at \( T = 250 \), the test has large empirical test size equals 9.61%. Finally, it is observed that in comparison

4.2. Empirical test sizes and powers of other backtests

Empirical test sizes of other backtests are first investigated. This is followed by study of their test powers if the samples are distributed as IID Student \( t \), normal inverse Gaussian (NIG) or follow the GARCH process. ES-SDP, ES-Delta and ES-Censor refer to the backtests of ES using the saddlepoint, functional delta and censored Gaussian methods, respectively. These are size-based backtests whereas Bernoulli and Kupiec (see Kupiec (1995)) are frequency-based backtests. Same as above, \( ^a \) denotes empirical test size significantly differs from theoretical size at 1% level.

\[ \text{Table 2 are all based on 5\% significance level. In practice, financial time series are found to exhibit excess kurtosis, have longer left tails and time-varying conditional variances. I use three non-normal processes to replicate this behavior. First, Panel B considers the case of Student } t \text{ distribution with 5 degrees of freedom. Student } t \text{ distribution is well known for its heavy tails but is symmetric. Next in Panel C, the IID NIG samples have excess kurtosis and are skewed to the left. The parameters of the NIG distribution are } \alpha = 0.8, \beta = -0.05, \delta = \alpha (1 - \rho)^{3/2} \text{ and } \mu = -\rho \delta (1 - \rho)^{1/2} \text{ where } \rho = \beta / \alpha. \text{ Finally in Panel D, the time-varying conditional variance of financial returns are simulated by a GARCH(1,1) process } R_t = \sigma_t Z_t \text{ with volatility equation given by } \sigma_t^2 = 0.05 + 0.15 \rho_i^2 + 0.8 \sigma_i^2 \text{ and } Z_t \text{ distributed as IID } N(0, 1). \text{ Generally speaking, the ES-based backtests are more powerful in detecting incorrect risk models, and amongst the ES-backtests, the saddlepoint technique is the most powerful. Though the Kupiec’s likelihood ratio test is the most powerful in detecting non-normal tail risk when the underlying process is GARCH(1,1) at } T = 250, \text{ the test has large empirical test size equals 9.61%. Finally, it is observed that in comparison} \]

\[ \text{For the study of test powers, 10,000 simulations are used and the rejection rates reported in Panels B, C and D of Table 2 are all based on 5\% significance level. In practice, financial time series are found to exhibit excess kurtosis, have longer left tails and time-varying conditional variances. I use three non-normal processes to replicate this behavior. First, Panel B considers the case of Student } t \text{ distribution with 5 degrees of freedom. Student } t \text{ distribution is well known for its heavy tails but is symmetric. Next in Panel C, the IID NIG samples have excess kurtosis and are skewed to the left. The parameters of the NIG distribution are } \alpha = 0.8, \beta = -0.05, \delta = \alpha (1 - \rho)^{3/2} \text{ and } \mu = -\rho \delta (1 - \rho)^{1/2} \text{ where } \rho = \beta / \alpha. \text{ Finally in Panel D, the time-varying conditional variance of financial returns are simulated by a GARCH(1,1) process } R_t = \sigma_t Z_t \text{ with volatility equation given by } \sigma_t^2 = 0.05 + 0.15 \rho_i^2 + 0.8 \sigma_i^2 \text{ and } Z_t \text{ distributed as IID } N(0, 1). \text{ Generally speaking, the ES-based backtests are more powerful in detecting incorrect risk models, and amongst the ES-backtests, the saddlepoint technique is the most powerful. Though the Kupiec’s likelihood ratio test is the most powerful in detecting non-normal tail risk when the underlying process is GARCH(1,1) at } T = 250, \text{ the test has large empirical test size equals 9.61%. Finally, it is observed that in comparison}\]

\[ \text{Given a sample of size } T \text{ with } b \text{ losses exceeding the magnitude of VaR, the } p \text{-value of the one-tailed Bernoulli test is given by } P(B \geq b) = 1 - P(B < b) \text{ where } B \text{ is distributed as Binomial}(T, z). \text{ Some articles use alternative } p \text{-value definition given by } 1 - P(B \leq b). \text{ However, this definition is not adopted by this paper because, say for } B \sim \text{ Binomial}(250, 0.01) \text{ and } b = 5, \text{ the } p \text{-value given by } 1 - P(B \leq 5) = 0.0412 \text{ is significant at 5\% level but } P(B = 5) = 0.0666.}\]

\[ \text{Let } g(z) = d[\delta^2 + (z - \mu)^2]^{1/2}, \text{ then the density of NIG}(x, \beta, \delta, \mu) \text{ is given by } f_{NIG}(z) = (\pi g(z))^{-1/2} \exp\{d \cdot g(z)^2 + \beta (z - \mu) \cdot \K0(2g(z))\}, \text{ where } \K0(x) \text{ is the modified Bessel function of the third kind. See, for example, Eberlein et al. (1998).}\]
to the other two non-normal alternatives, the powers against the GARCH process are relatively low.

4.3. Estimation risk

The Monte Carlo simulation studies so far assume that the parameters of forecasted distributions are free from estimation errors. In practice, this is of course not true. Therefore, the effects of estimation risk on the test sizes and powers of various backtests are studied here. Specifically, sample mean and standard deviation computed from estimation sample of size $P$ are used to standardize testing sample under the standard normal null hypothesis. Without loss of generality, only test sample size of 250 is considered. For the estimation, sample sizes of 125, 250, 500 and 1000 are used to compute the standard normal parameters, mean and standard deviation, on a rolling window basis. For example, estimation sample $R_1, \ldots, R_P$ are used to compute the sample mean and standard deviation, denoted as $\bar{\mu}_P$ and $\hat{\sigma}_P$, respectively. One-step-ahead testing observation $R_{P+1}$ is standardized as $\hat{Z}_{P+1} = \hat{\sigma}_P^{-1}(R_{P+1} - \bar{\mu}_P)$. The next standardized testing observation $Z_{P+2}$ is obtained in the same way using rolling sample $R_2, \ldots, R_{P+1}$. This process is repeated until the required full testing sample is constructed. Table 3 provides the empirical test sizes and powers at 5% significance level of various backtests based on 99% VaR. It is obvious from the results in Panel A that small estimation sample results in over-sized distortion in the proposed saddlepoint technique and under-sized effect in the other two ES backtesting methods. When estimation sample increases to 1000, the saddlepoint technique gives correct size but the under-sized problems are marginally more severe for the other two ES backtests. Size-corrected test powers are presented in the rest of the table. Overall, test powers decrease with estimation sample, suggesting that small sample estimation risk results in higher test powers even after size-adjustments (see Table 4).

5. Application of ES backtesting for non-normal trading risk

As an example of applying the saddlepoint backtest, I apply in this section the ES backtesting to the profit and loss (P&L) of trading accounts at six large banks studied by Berkowitz and O’Brien (2002). Given the proprietary nature of the banks’ P&L and VaR data, they are not easily available. However, the proposed backtest is easy to implement...
and sub-samples, the results of backtests show that four out of six banks
the VaR calculated by the bank’s structural model. Except for the case of
tentiality. Mean violation refers to the average of the losses that exceed
1998. The daily data are standardized to have unit variance for confi-
(Panel B) refers to the financial turbulent period from August to October
covers from January 1998 through March 2000 whereas the sub-sample
data of 6 large commercial banks in US. The full sample period (Panel A)
This table applies the proposed backtest of ES to the daily profits and loss
Panel B: sub sample from August 1998 to October 1998
Panel A: full sample from January 1998 to March 2000

<table>
<thead>
<tr>
<th>Bank</th>
<th>No. of exceptions</th>
<th>Mean violation</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank 1</td>
<td>65</td>
<td>4.446</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Bank 2</td>
<td>64</td>
<td>3.188</td>
<td>0.0018</td>
</tr>
<tr>
<td>Bank 3</td>
<td>65</td>
<td>3.506</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Bank 4</td>
<td>573</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bank 5</td>
<td>746</td>
<td>3.101</td>
<td>0.0964</td>
</tr>
<tr>
<td>Bank 6</td>
<td>586</td>
<td>8.166</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

This table applies the proposed backtest of ES to the daily profits and loss
data of 6 large commercial banks in US. The full sample period (Panel A)
refers to the financial turbulent period from August to October
The daily data are standardized to have unit variance for confidence.
Mean violation refers to the average of the losses that exceed
the VaR calculated by the bank’s structural model. Except for the case of
are represented by positive values here.

Table 3 lists for the six US large banks the number of
as it only requires the number of exceptions and the mean
violation of the banks’ trading books which are reported
by Berkowitz and O’Brien. Using frequency-based back-
tests, they find both the bank’s internal VaR model and
the reduced form normal GARCH model provide sufficient
coverage of trading risks. It is nevertheless valuable for risk
managers to investigate statistically if the exceptions exhibit
non-normal or excessive tail risk. Indeed, the proposed
backtests reveal that four out of six banks have excessive
trading risk.

5.1. Banks’ trading P&L and internal VaR

Some descriptions on the banks’ trading P&L and the
associated VaR are provided here. Readers are referred to
Berkowitz and O’Brien (2002) for more detailed information.
All of the six banks are large multinational banks
that meet the Basle “large trader” criteria and their trading
activities include trading in interest rate, foreign exchange,
equity, commodity and derivatives. To manage market
risks, large scale structural models are developed which
measure and aggregate trading risks in current positions
at a highly detailed level. Given their roles as principal
dealers in the ever growing over-the-counter derivatives
markets, their trading accounts have grown rapidly and
become progressively more complex. Thus the structural
models can have exposures to several hundred thousand
market risk factors, with individual positions numbering
in the millions. Therefore, approximations are employed
in order to turn out daily VaR for trading risk. It is
required by regulation that the daily VaR estimates are
maintained by the banks for the purposes of forecast eval-
uation or “backtesting”. Because of their proprietary nature,
there has been little empirical study of the P&L of
trading firms and the associated VaR forecasts.

5.2. Backtest of bank’s internal VaR model using expected shortfall

Some assumptions are first made regarding the bank’s
trading P&L. Let $R_t$ denotes the bank’s trading returns at
day $t$ and it is assumed that $R_t \sim N(\mu, \sigma)$. So the bank’s
trading returns is allowed to have time varying expected
mean due to daily changes in the composition of the portfolio,
but its variance is assumed to be constant. Though $\mu_t$
and $\sigma$ are not known, the reported mean violation is calcu-
lated in excess of VaR and the bank’s P&L is standardized
to have unit variance in order to protect confidentiality.
Hence, I can easily obtain the sample ES as $ES_n = \bar{x} + z_{0.01}$, where $\bar{x}$ is the mean violation in excess
of VaR given by Berkowitz and O’Brien (2002) and
$z_{0.01} = 2.326$ is the absolute value of 1% normal quantile.

Table 3 lists for the six US large banks the number of
observations, the number of exceptions, the mean violation
(or the sample ES) and the corresponding $p$-value calculated
using (9). Panel A provides the backtest results for
the full sample period from January 1998 to March 2000.
To illustrate how the $p$-value depends on the number of
exceptions, consider for instance the case of Bank 2 which
has six losses exceeded its internal VaR. According to (9)
with $n = 6$, the corresponding $p$-value is evaluated as
0.0050. Thus I infer at 1% significance level that there is
excessive trading risk for Bank 2. As another example,
the mean violation of Bank 5 is 3.101 based on one excep-
tion. Since in this case, the corresponding $p$-value is 0.0964,
the null hypothesis is accepted and the risk model is said to
“cover” the risk as measured by ES, the average of tail
losses. As can be seen from the backtest results in Panel
A, four out of six banks exhibit excessive or fatter-than-
normal market risks at 1% significance level. This is consis-
tent with the observation by Berkowitz and O’Brien that
the banks’ internal VaRs are conservative; but when excep-
tions occur, they tend to exceed the VaR models drastically.

19 See for example, page 71 of 2004 Annual Report of JPMorgan Chase & Co.
20 Alternatively, the sample ES of 3.067 is said to exceed at 1 percent significance level the theoretical ES, which equals to 2.6652 under the null hypothesis.
Berkowitz and O’Brien also find most of the exceptions took place between August and October of 1998, a turbulent period for financial markets.\footnote{Between August and October of 1998, there were the devaluation of the Russian ruble, Russian debt default and the near-collapse of LTCM.} They provide the relevant data on mean violation and number of exceptions, but do not carry out backtests on them during this period. Backtests that rely on large sample asymptotic statistic may not be applied here because the sample is too small for reliable inference. The advantage of the small sample asymptotic technique becomes obvious in this situation. Regardless of the length of observations, the proposed backtest of ES can be applied as long as there are exceptions. It can be seen from the backtest results presented in Panel B that banks that fail the backtest during the full sample also fail the test here. I therefore infer that the excessive trading risks come from this period of extreme market movements.

6. More responsive risk management decision

In this section, I provide a simple illustration of how the proposed ES backtesting can make risk management decision more responsive. In the study by Berkowitz and O’Brien (2002), a normal GARCH(1,1) is used as an alternative to the bank’s VaR model. Since the time series model attempts no accounting for changes in portfolio composition, the bank’s structural model should in principal deliver superior VaR forecasts. Berkowitz and O’Brien find, however, the normal GARCH model permits comparable risk coverage with lower VaRs, that is, less regulatory capital. They attribute this to the ability of the time series model to capture trend and time varying volatility in a bank’s P&L. Since time varying volatility as specified by a normal GARCH also generates fat tails, it is interesting to see how effective the model is in accounting for the non-normal trading risk in the equity markets. Specifically, I fit the following AR(1)-GARCH(1,1) risk model to S&P 500 daily index returns:

\[ R_t = \mu_t + \varepsilon_t = \mu + \rho R_{t-1} + \sigma_t Z_t, \]
\[ \sigma_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}, \]

where \( R_t \) is the S&P 500 return at day \( t \) and \( Z_t \) is IID standard normal process. Conditional on information available on \( t-1 \), one day ahead VaR is forecasted as \( \text{VaR}_t = 2.33\hat{\sigma}_t - \hat{\mu}_t \). For \( n \geq 0 \) realized exceptions up to time \( t \), sample expected shortfall is computed as \( \text{ES}_n = -n^{-1}\sum_{i=1}^{n} \hat{Z}_i \), where the standardized exception is given by \( \hat{Z}_i = \hat{\sigma}_i^{-1}(R_i - \hat{\mu}_i) \).

Since the AR(1)-GARCH(1,1) risk model assumes IID standard normal innovation process \( Z_t \), the null hypothesis will hold if the risk model fully describes the S&P 500 return dynamics. To see how the saddlepoint ES backtesting can help risk management decision responsive, for the sake of illustration, suppose that we start the risk forecast from 1 August 1998.\footnote{3,000 observations prior to 1 August 1998 are first used to estimate the risk model and ten one-step-ahead VaRs are forecasted. The estimation window is rolled forward ten trading days and a new set of model parameters are obtained, which will in turn be used to forecast one-step ahead VaR for the next ten trading days. The process is continued until 31 July 1999 is reached.} As it turns out, there are 5 exceptions in the following twelve months ending on 31 July 1999. Fig. 2 depicts the series of mean violations (or sample ES) with first exception realized on 25 August 1998. For the first exception, \( n = 1 \) and the sample ES lies below the 5% critical value and is thus significant at 5% level. When the second exception occurs, \( n = 2 \) and the mean violation becomes significant at 1% level, an even stronger message for the normal GARCH model to be rejected. Therefore, the proposed ES backtesting facilitates prompt risk management response because accurate critical values can be calculated for any number of exceptions.

7. Conclusion

Berkowitz and O’Brien (2002) find that banks’ internal VaR models tend to be conservative but the losses are huge when exceptions occur. This motivates backtest of bank’s trading risk using expected shortfall (ES). The backtest of ES, however, is difficult to implement since the existing censored Gaussian and functional delta methods rely on asymptotic test statistics that converge poorly for small samples. The proposed saddlepoint backtest employs small sample asymptotic technique, which Monte Carlo simulation shows is highly accurate and powerful even when sample size is small. Conditional on the reported number of exceptions and mean violation for the period from January
1998 to March 2000, I find four out of the six banks studied by Berkowitz and O’Brien have excessive market risk. Moreover, since the proposed backtest depends on the number of exceptions but not the sample size, the saddlepoint backtest can also be applied to the three months sub-sample period from August to October 1998 when financial markets were extremely volatile. It is found that banks that have significant non-normal market risk in the full sample also fail the sub-sample backtests, suggesting that most of the detected tail risks can be attributed to the volatile financial markets during this period.

The proposed saddlepoint technique can compute under the null hypothesis the required p-value accurately conditional on any number of exceptions. This has two important implications. First, when there are huge losses, even though there is just one or two of them, the risk model can be evaluated on a formal statistical basis. This enables prompt action to be taken by the risk managers or regulators if necessary. Second, the proposed backtest can be used alongside the existing VaR measure. Since VaR has been sanctioned by the Basle Committee and various central banking authorities, it is now the standard risk measure. Hence it is useful for backtesting purpose to calculate ES based on the losses that exceed the forecasted VaR. This means the number of exceptions will vary from sample to sample, which poses no problem for the proposed backtest using the small sample asymptotic technique.

In short, risk management process is about communicating various aspects of risks effectively and condensing complex risk factors into simple, understandable terms. To this end, VaR has demonstrated to be a successful risk measure. However, being a quantile statistic, VaR ignores valuable information contained in the sizes of losses beyond the VaR level. By averaging the sizes of exceptions, ES communicates the degree of non-normal or tail risk in a simple way to investors, risk managers and regulators. The shortcoming that holds back ES is its backtesting, especially when sample size is small. This paper overcomes this shortcoming by using the saddlepoint technique to compute under the null hypothesis the required p-value of the test statistic. As the Basle Committee on Banking Supervision (1996b, p. 2) recognizes the fact that “the techniques for risk measurement and backtesting are still evolving, and the Committee is committed to incorporating important new developments in these areas into its framework,” it is hoped that the proposed ES backtesting will help raise the standard of future risk management.

Acknowledgment

The author would like to thank participants for suggestions and comments at The 5th NTU International Conference on Economics, Finance and Accounting—Financial Institutions and Capital Market, Taipei, Taiwan, May 23–24, 2007.

Appendix

Proof of Proposition 1. Consider the moment generating function

\[ M(t) = E(e^{\theta t}) = \int_{-\infty}^{\infty} e^{\theta x} f(x) \, dx, \]

which, after substituting \( x^{-1}\phi(x) \) for \( f(x) \) in the integral term, collect the exponential terms and simplify, will give rise to (3). For the derivatives of the moment generating function, first denote the differential operator \( \frac{d}{dt} \) by \( D_t \). Now using the standard results that \( D_t \theta(t) = \theta(t) \) and \( D_t \cdot \exp \left( \frac{t^2}{2} \right) = t \cdot \exp \left( \frac{t^2}{2} \right) \), the first derivative of the moment generating function can be written as

\[ M'(t) = t \cdot M(t) - \exp \left( \frac{t^2}{2} \right) \cdot x^{-1} \phi(q - t), \]

\[ = t \cdot M(t) - \exp(qt) \cdot x^{-1} \phi(q). \]

Second and higher order derivatives of the moment generating function given by (4) follow straight away from differentiating above result. □

Proof of Corollary 1. Letting \( t = 0 \) in \( M(t) \) and \( M'(t) \) from (4) will give rise to \( E(X) \) and \( E(X^2) \), respectively. It is easy to obtain of variance of \( X \) since \( \text{var}(X) = E(X^2) - [E(X)]^2 \). □

References

Basle Committee on Banking Supervision, 1996a. Amendment to the Capital Accord to incorporate market risk.