A New Approach to Comparing VaR Estimation Methods

CHRISTOPHE PÉRIGNON AND DANIEL R. SMITH

CHRISTOPHE PÉRIGNON
is an associate professor at HEC in Paris, France.
perignon@hec.fr

DANIEL R. SMITH
is an associate professor in the faculty of business at Simon Fraser University in Vancouver, Canada, and a visiting fellow in the faculty of business in the Queensland University of Technology in Brisbane, Australia.
drsmith@sfu.ca

We develop a novel backtesting framework based on multidimensional value-at-risk (VaR) that focuses on the left tail of the distribution of bank trading revenues. Our coverage test is a multivariate generalization of Kupiec's unconditional test. Applying our method to actual daily bank trading revenues, we find that nonparametric VaR methods, such as GARCH-based methods or filtered historical simulation, work best for bank trading revenues.

Value-at-risk (VaR) is the standard measure for market risk used by financial institutions and banking regulators. VaR quantifies the loss that a bank can face on its trading portfolio within a given period and for a given confidence interval. Since the first widely publicized appearance of the term VaR in a G-30 report in July 1993, numerous statistical methods have been proposed to compute this market risk measure, such as RiskMetrics and historical simulation (HS). Furthermore, over the past 10 years, an increasing number of banks have been setting their regulatory capital using in-house VaR measures. The quest for the most accurate VaR method is of great interest to regulators and risk managers in charge of developing banks' proprietary risk models. Unfortunately, current backtesting procedures for VaR models are known to lack power.

The key methodological contribution of our article is the development of a new backtesting framework for VaR. It is based on multidimensional VaR, which is a vector of VaRs measured with the same horizon but different coverage probabilities or confidence levels. Our basic idea is that the accuracy of a given VaR method should not be assessed using only a single coverage probability or, in other words, a single observation on the trading revenue distribution. Instead, we focus on the left tail of the distribution of the trading revenues and consider K different confidence levels. Our coverage test is a multivariate generalization of the unconditional coverage test of Kupiec [1995], which remains the standard backtesting procedure. Furthermore, our test can be implemented with any combination of coverage probabilities or any time horizon—for example, [1% 2% ... 10%], [0.1% 0.2% 0.3%] or 1 day, 10 days.

To date, almost all empirical comparisons of VaR methods have been based on hypothetical positions in individual assets, interest rates, exchange rates, or stock indices. For instance, Ferreira and Lopez [2005] consider an equally weighted portfolio of short-term fixed-income positions in the U.S. dollar, German deutschemark, and Japanese yen using daily interest rate and foreign exchange data from 1979 through 2000.[1] The reason for this practice is because actual trading positions and daily trading revenues of banks are typically unknown to the public. Unlike previous studies, we use actual daily trading revenues from several large banks to identify the most accurate VaR method. To the best of our knowledge, our study is the first to compare VaR methods using actual bank trading revenues.

Comparing VaR methods on the basis of hypothetical holdings is of limited interest for banks. Indeed, knowing that a given VaR method outperforms its competitors is a useful piece of information if and only if the data used in the horse race (e.g., the S&P 500 Index) is closely related to the actual composition of bank trading portfolios. Recent empirical evidence suggests that this is unlikely to be the case. Berkowitz and O'Brien [2007] show that regressing actual daily U.S. bank trading revenues on market factors (e.g., stock index, interest rate, credit spread, exchange rate) leads to 1) low R-squares, 2) time-varying regression coefficients, and 3) very different regression coefficients across banks. Their findings suggest that banks' trading positions are 1) complex and affected by nonstandard risk factors,[2] 2) frequently rebalanced, and 3) very different across banks. This latter conclusion is confirmed by Jorion [2007] in his study of quarterly trading revenues of 11 U.S. banks. He reports a modest correlation across banks' aggregate trading revenues (i.e., the average correlation is 0.163) and across banks' trading revenues by category (i.e., from an average correlation of –0.039 for commodity trading to 0.149 for equity trading).

In this article, we compare existing VaR methods empirically using actual daily trading revenues from five international banks: Bank of America, Credit Suisse First Boston, Deutsche Bank, Royal Bank of Canada, and Société Générale. Using real data allows us to identify the VaR methods that work best in an empirically relevant context. Other appealing features of our dataset include its high frequency; its cross-sectional dimension that allows us to compare results across banks and/or domestic regulatory jurisdictions; its four-year time span that permits assessment of the performance of different VaR methods in different market conditions; and the fact that the sample banks are not anonymous.

The main conclusion of our empirical study is that parametric VaR methods, such as GARCH-based models and filtered HS work best for bank trading revenues. Our findings contrast with current practice at commercial banks since most proprietary VaR models rely on nonparametric methods, such as HS. We also find that the multivariate test can detect misspecification missed by the univariate tests. In particular, HS for the Bank of America is not rejected by any of the univariate tests but is rejected by the multivariate test.

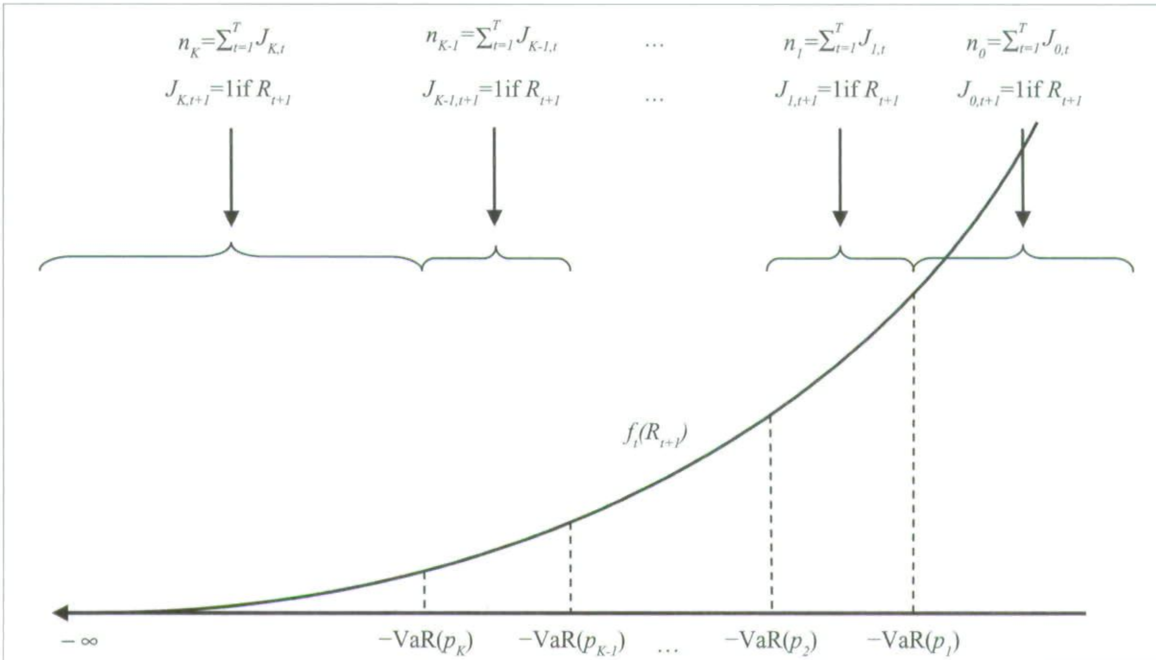
THE MULTIVARIATE UNCONDITIONAL COVERAGE TEST

The unconditional coverage test of Kupiec [1995] allows one to test whether the actual violation rate, which is the number of days when the trading loss is greater than VaR(p) divided by the sample size, is equal to the coverage probability p. While this test remains the reference test in financial risk management, it displays low statistical power when implemented with typical datasets, such as one year of daily data (see Jorion [2006]).[3] An alternative testing procedure suggested by Berkowitz [2001] allows one to test the entire distribution using the Rosenblatt transformation. If the conditional density is correctly specified then inverting the model cumulative density function using the normal density will produce a normal random variable. Berkowitz [2001] also suggests restricting the analysis to the left tail of the distribution since risk managers are generally unconcerned about the interior or the right tail of the distribution.

We suggest a hybrid approach that combines the most attractive features of these two tests. On the one hand, our test maintains the simplicity of Kupiec's test since it only relies on VaR and on the number of exceptions. On the other hand, it uses more information about the left tail, which can improve the power of the test. Specifically, we suggest a test that focuses on a range of VaRs evaluated at a collection of coverage probabilities in the left tail of the distribution. A nice feature of our approach is that it fits well with current practice at commercial banks since it only requires banks to disclose a *series* of one-day ahead VaRs with different coverage probabilities, instead of a *single* one-day ahead VaR, as they currently do.

Bank revenue in period t+1 is given by R_{t+1}, and the VaR with a coverage probability p_i is denoted by $VaR_{t+1|t}(p_i)$ and satisfies $\Pr(R_{t+1} \leq -VaR_{t+1|t}(p_i) \mid \psi_t) = p_i$, where ψ_t denotes the information set at time t. We focus on multivariate VaR with K different coverage probabilities, indexed by i in descending order, $p_1 > p_2 > \cdots > p_K$. These VaRs become more extreme as i increases, $VaR_{t+1|t}(p_1) < VaR_{t+1|t}(p_2) < \cdots < VaR_{t+1|t}(p_K)$, which we illustrate graphically in Exhibit 1.

EXHIBIT 1
Presentation of the Multivariate Unconditional Coverage Test



Notes: This exhibit illustrates the construction of the multivariate coverage test presented in Equation (4). It displays the left tail of the forecasted revenue distribution $f_t(R_{t+1})$ along with K VaRs corresponding to K coverage probabilities, $p_1 > p_2 > \cdots > p_{K-1} > p_K$. The J variables are Bernouilli variables equal to one if the actual revenue falls between two adjacent VaRs. The LR_{MUC} test requires the θ_i and $\hat{\theta}_i$ K-dimensional vectors, n_0, and n_i, for $i = 1,\ldots,K$, to be known. We obtain the key inputs as follows: $\theta_i = p_i - p_{i+1}$, $\hat{\theta}_i = (1/T)\Sigma_{t=1}^{T} J_{i,t}$ for $i = 1,\ldots,K$ and $n_i = \Sigma_{t=1}^{T} J_{i,t}$ for $i = 0,\ldots,K$.

Associated with each of the K VaR numbers is an indicator variable for revenues falling in each disjoint interval (see Exhibit 1):

$$J_{i,t+1} = \begin{cases} 1 & \text{if } -VaR_{t+1|t}(p_{i+1}) < R_{t+1} \leq -VaR_{t+1|t}(p_i) \\ 0 & \text{otherwise} \quad \text{for } i = 1,\ldots,K \end{cases} \tag{1}$$

augmented with $p_{K+1} = 0$, $VaR_{t+1|t}(p_{K+1}) = +\infty$, and $J_{0,t+1} = \prod_{i=1}^{K}(1 - J_{i,t+1})$. Note that $J_{i,t+1}$ can also be expressed as a function of the more traditional exception indicator:

$$I_{i,t+1} = \begin{cases} 1 \text{ if } R_{t+1} \leq -VaR_{t+1|t}(p_i) \\ 0 \text{ if } R_{t+1} > -VaR_{t+1|t}(p_i) \end{cases} \quad \text{for } i = 1,\ldots,K \tag{2}$$

with $J_{i,t+1} = I_{i,t+1} - I_{i+1,t+1}$. The $J_{i,t+1}$ Bernoulli random variables equal one with probability $\theta_i = p_i - p_{i+1}$ when $i > 0$, and $J_{0,t+1} = 1$ with probability $\theta_0 = 1 - p_1$. However they are clearly not independent since only one J variable may be equal to one at any time period, $\Sigma_{i=0}^{K} J_{i,t+1} = 1$. We will collect the K elements θ_i into the K-dimensional parameter vector $\theta = (\theta_1,\ldots,\theta_K)'$.

We can test the joint hypothesis of the specification of the VaR model using a Likelihood Ratio test based on $N_{t+1} = i$ when $J_{i,t+1} = 1$ for $i \geq 0$, which summarizes all of the information in the various exception indicators. The density of N is given by $g(n_t;\, \theta) = \Pr(N_t = n_t;\, \theta) = (1-1'\theta)^{J_{0,t}} \Pi_{i=1}^{K} \theta_i^{J_{i,t}}$, giving the log-likelihood function:

$$LL(N;\theta) = \sum_{t=1}^{T}\left(J_{0,t}\ln(1-1'\theta) + \sum_{i=1}^{K} J_{i,t}\ln(\theta_i)\right) \tag{3}$$

Our multivariate unconditional coverage test is a likelihood ratio test LR_{MUC} that the empirical θ significantly deviates from the hypothesized θ. Formally, it is given by:

$$LR_{MUC} = 2\left(\left[n_0 \ln(1-1'\hat{\theta}) + \sum_{i=1}^{K} n_i \ln(\hat{\theta}_i)\right] - \left[n_0 \ln(1-1'\theta) + \sum_{i=1}^{K} n_i \ln(\theta_i)\right]\right)$$
$$= 2\left(\ln\left[(1-1'\hat{\theta})/(1-1'\theta)\right]^{n_0} + \sum_{i=1}^{K} \ln(\hat{\theta}_i/\theta_i)^{n_i}\right) \qquad (4)$$

where $n_i = \Sigma_{t=1}^{T} J_{i,t}$ and $\hat{\theta}_i$, the maximum likelihood estimator of θ with element i, is given by $\hat{\theta}_i = (1/T)\Sigma_{t=1}^{T} J_{i,t}$ (see appendix for the proof).[4] The LR_{MUC} statistic is asymptotically chi-square with K degrees of freedom. When $K = 1$, our multivariate test boils down to the univariate unconditional coverage test of Kupiec [1995].

We illustrate our test using a hypothetical sample of 500 historical observations ($T = 500$) of concurrent trading revenues and VaR estimates computed using three coverage probabilities, $p_1 = 5\%$, $p_2 = 2.5\%$, and $p_3 = 1\%$. We assume that VaR(1%) is violated 8 times, VaR(2.5%) 9 times, and VaR(5%) 21 times. This implies that the distribution of the exceptions is such that there are 12 occurrences when the loss is between VaR(2.5%) and VaR(5%) and one occurrence when the loss is between VaR(1%) and VaR(2.5%), that is $n_1 = 12$, $n_2 = 1$, and $n_3 = 8$. According to the univariate coverage test LR_{UC}, the VaR model cannot be rejected with any of the coverage probabilities, that is, p-value(1%) = 0.215, p-value(2.5%) = 0.292, p-value(5%) = 0.399. However, the multivariate coverage test suggests the opposite:

$$LR_{MUC} = 2\left(\ln\left[\frac{1-(0.024+0.002+0.016)}{1-(0.025+0.015+0.010)}\right]^{479} + \ln\left(\frac{0.024}{0.025}\right)^{12} + \ln\left(\frac{0.002}{0.015}\right)^{1} + \ln\left(\frac{0.016}{0.010}\right)^{8}\right)$$

which is $LR_{MUC} = 10.544$, with a p-value of 0.014. While the univariate tests suggest that the bank's VaR model is correct, our multivariate test clearly rejects the null hypothesis that the VaR model is well specified.

VaR METHODS

Our selection of VaR methods is driven by current practice at commercial banks. In a recent international survey of VaR usage, Pérignon and Smith [2008] show that HS is at the heart of most VaR models currently in use at commercial banks. They find that 73% of the firms that disclose their VaR methodology in their 2005 annual reports use HS or related techniques. HS is a flexible, nonparametric technique that forecasts future potential price changes using actual shocks on state variables that occurred in the past (Christoffersen [2003], Campbell [2005], Jorion [2006], and Pritsker [2006]). When directly applied to aggregate trading revenues, VaR(p) is the empirical p-percentile of the trading revenue distribution.[5] For instance, with a sample of 500 observations, the VaR(5%) is given by the 25th smallest trading revenue.

The second most popular VaR method is Monte-Carlo simulation and other related parametric methods. This family of techniques is reported to be used by 21.6% of the disclosing banks surveyed in Pérignon and Smith [2008]. In our tests, we employ several prominent parametric methods that aim to capture time variation in the conditional variance of the trading revenues, h_t. In the RiskMetrics method, h_t is modeled using an exponentially weighted moving average forecast of the form:

$$h_{t+1} = \lambda \cdot h_t + (1-\lambda) \cdot R_t^2 \qquad (5)$$

where the decay factor $\lambda = 0.94$. In this model, the weight on an observation that is d-days old is $(1-\lambda) \cdot \lambda^{d-1}$, which is decaying fairly quickly through time. Indeed, the weight is 0.060 for $d = 1$, 0.034 for $d = 10$, and 0.003 for $d = 50$. We also use the more general GARCH(1,1) model (Bollerslev [1986]):

$$h_{t+1} = \omega + \alpha \cdot R_t^2 + \beta \cdot h_t \qquad (6)$$

where the conditional variance of trading revenues, $h_{t+1} = E_t(R_{t+1}^2)$, is modeled as a GARCH process with standardized innovations that are, alternatively, normal or Student-t random variables. We estimate the GARCH model by maximum likelihood, and in the GARCH-T model we estimate the GARCH parameters and the degrees of freedom (ν) jointly by maximizing the log-likelihood. To forecast time $t+1$ conditional variance in Equations (5) and (6), we estimate the parameters using observations up to and including time t. For the three parametric methods, we measure VaR as follows:

$$VaR_{t+1|t}^{RM}(p) = \sqrt{\hat{h}_{t+1}} Z_p \tag{7}$$

$$VaR_{t+1|t}^{\text{GARCH}}(p) = \sqrt{\hat{h}_{t+1}} Z_p \tag{8}$$

$$VaR_{t+1|t}^{\text{GARCH-T}}(p) = \sqrt{\hat{h}_{t+1}} \cdot T_p(\hat{v}_t) \cdot \sqrt{\frac{\hat{v}_t - 2}{\hat{v}_t}} \tag{9}$$

where Z_p denotes the p-th percentile of a standardized normal variable and $T_p(\hat{v}_t)$ denotes the p-th percentile of a Student-t random variable with $\hat{v}_t$ degrees of freedom.

Besides the purebred parametric or nonparametric methods, we consider two semi-parametric methods: filtered HS of Barone-Adesi et al. [1999, 2002] and the hybrid HS method proposed by Boudoukh et al. [1998]. In the filtered HS method a parametric GARCH model is initially filtered, which generates a sequence of standardized returns $\hat{z}_t = \frac{R_t}{\sqrt{\hat{h}_t}}$ where $\hat{h}_t$ denotes the in-sample fitted conditional volatility estimate from the GARCH model. The VaR is then estimated as:

$$VaR_{t+1|t}^{FiltHS}(p) = \sqrt{\hat{h}_{t+1}} \hat{Z}_p \tag{10}$$

and $\hat{Z}_p$ is the empirical p-th percentile of the fitted standardized returns $\hat{z}_t$ over the previous 250 trading days (see also Hull and White [1998] for a similar method). The hybrid HS method assigns exponentially declining probability weights to past trading revenues. In particular each of the last K trading revenues R_{t-i} are assigned the probability $\omega_i = \frac{1-\lambda}{1-\lambda^K}\lambda^i$ for $i = 1,..., K$ (the term $\frac{1-\lambda}{1-\lambda^K}$ ensures that the weights sum to one). We then sort the K returns and the associated weights, which we denote by $R_{(j)}$ and $\omega_{(j)}$. To compute the p-th percentile we then start with the lowest return and keep accumulating probabilities to this level and linearly interpolate between adjacent points to arrive at the appropriate percentile. The motivation behind the two semi-parametric techniques is that they allow for nonnormal data while attempting to capture the conditional heteroskedasticity of the data.

SIZE AND POWER OF COVERAGE TESTS

To understand the size (probability of incorrectly rejecting the null hypothesis) and power (probability of rejecting a false null hypothesis) properties of the various univariate and multivariate coverage tests we undertake a series of Monte Carlo experiments. We first study the size of the tests by simulating a time series of revenues and evaluating the performance of the VaR model computed under the *true* data-generating process (see Panel A of Exhibit 2). We simulate 2,000 different sample paths with a range of sample sizes including $T = \{250, 500, 1000, 2500\}$ revenues, corresponding to 1, 2, 4, and 10 years of daily data. This situation is equivalent to directly simulating the exception indicators. Since the null hypothesis is true, the rejection frequencies in Panel A should ideally be 5% (the nominal size of the test) in all cases. In practice, we find that all tests have good finite sample properties, particularly when the sample size is moderately large. The exception is VaR(1%) when the sample size is 250. We note that this is precisely the situation underpinning the backtesting strategy suggested by regulators to evaluate banks' proprietary VaR models.

To study the power of the tests, we generate revenues with zero mean and unit variance, without loss of generality, using a GARCH model with standardized Student-t innovations:

$$h_{t+1} = 0.05 + 0.05 \cdot e_t^2 + 0.9 \cdot h_t \tag{11}$$

where $e_t = \sqrt{h_t} \cdot u_t$ and $u_t \sim T(6.5)$. These parameter values are based on the full sample point estimates for the GARCH-T model fitted to Bank of America's trading revenues that we find in our empirical analysis in the next section. For each specific sample size, we generate 2,000 sample paths along with 250 extra revenues for each sample path. Starting from revenue 251, we compute three different incorrect VaR estimates based on the previous 250 revenues and, alternatively, on HS, RiskMetrics, and hybrid HS. We then compute T different out-of-sample VaR numbers for all three methods using the most recent 250 observations. Importantly we use the same simulations to estimate and evaluate all three VaR models.

We report the rejection frequencies at the nominal 5% level in panels B through D of Exhibit 2. In these panels, we are looking at rejection rates that are as close as possible to 1. The most striking result to take from

EXHIBIT 2
Size and Power of Coverage Tests

	Univariate Coverage Test p = 1%	Univariate Coverage Test p = 2.5%	Univariate Coverage Test p = 5%	Multivariate Coverage Test p = 1%, 2.5%, 5%
Panel A: Null Hypothesis is True				
$T = 250$	0.0910	0.0740	0.0435	0.0660
$T = 500$	0.0690	0.0600	0.0515	0.0570
$T = 1000$	0.0580	0.0420	0.0460	0.0625
$T = 2500$	0.0355	0.0505	0.0555	0.0535
Panel B: HS-VaR computed from GARCH-T Revenues				
$T = 250$	0.1060	0.0010	0. 0005	0.0135
$T = 500$	0.0240	0.0015	0.0055	0.0040
$T = 1000$	0.0650	0.0160	0.0055	0.0240
$T = 2500$	0.0990	0.0205	0.0040	0.0315
Panel C: RiskMetrics -VaR computed from GARCH-T Revenues				
$T = 250$	0.4690	0.5575	0.0115	0.1955
$T = 500$	0.6945	0.3980	0.0020	0.5805
$T = 1000$	0.6390	0.1975	0.0300	0.4805
$T = 2500$	0.9415	0.5070	0.0470	0.8985
Panel D: Hybrid HS-VaR computed from GARCH-T Revenues				
$T = 250$	1.0000	0.6045	0.0325	1.0000
$T = 500$	1.0000	0.3990	0.0390	0.9980
$T = 1000$	0.9990	0.8940	0.3195	0.9945
$T = 2500$	1.0000	1.0000	0.9410	1.0000
Panel E: GARCH-VaR computed from GARCH-Skew-T Revenues				
$T = 250$	0.1005	0.1560	0.2475	0.2325
$T = 500$	0.0635	0.1610	0.4325	0.3270
$T = 1000$	0.0465	0.1885	0.6895	0.6130
$T = 2500$	0.0565	0.5225	0.9745	0.9720

Note: This exhibit displays the rejection frequencies at the nominal 5% level of three different univariate coverage tests (1%, 2.5%, and 5% coverage probabilities) and our multivariate coverage test, which is jointly based on the three coverage probabilities. In each Monte Carlo simulation, we simulate 2000 different sample paths of revenues with different sample sizes, $T = \{250, 500, 1000, 2500\}$. In Panel A, we evaluate the performance of the VaR model computed under the true data-generating process. In Panels B through D, we generate revenues with zero mean and unit variance using a GARCH model with standardized Student-t innovations, and we estimate one-day ahead VaR estimates using HS, RiskMetrics, and hybrid HS, respectively. In Panel E, we study the ability of the tests to reject a parametric normal-based VaR model when the standardized revenues are generated from a Skew-t distribution that has by construction the same 1% quantile as a standard normal random variable.

these results relates to HS. When the number of observations used to compute HS-based VaR is small relative to the number of out-of-sample observations, the tests lack power (see Smith [2007] for a similar result and further details). The weakness of unconditional coverage tests is common to all tests considered. We find that the other two incorrect VaR models (i.e., RiskMetrics and hybrid HS) are rejected quite frequently, with rejection rates approaching 100% when the sample size is around 10 years. Interestingly the most power comes from either the univariate 1% test or the multivariate test. This finding is not surprising given that we are simulating revenues from the Student-t distribution and the disparity between the true model and hypothesized models, particularly for the normal-based RiskMetrics VaR, are most pronounced in the tails.

These results do not imply that the 1% univariate test will dominate the 5% test in all situations. To illustrate this we study the ability of the tests to reject a parametric normal-based VaR model when the standardized revenues are generated from the skew-t distribution of Hansen [1994] with five degrees of freedom and asymmetry parameter 0.14542. This latter model is chosen because it has the same 1% quantile as a standard normal random variable. As before, we simulate returns from a GARCH model (Equation 11) but this time with skewed Student-t innovations and evaluate the power of the tests to reject a normal-based GARCH model. Because of the similar left-tail behavior the 1%-based tests lack power to detect violations (see Panel E). Conversely, the 5% test performs admirably. Again the power is maximized when we focus on regions in the tail where the differences between null and true models are most pronounced. Interestingly, we find that the multivariate test works similarly to the 5% univariate test.

The main points to take from these simulation experiments are that the optimal univariate test to consider varies with the alternative model. Interestingly, we find that the multivariate test performs virtually as well as the best univariate test. Given that the econometrician does not have the luxury of knowing the alternative model and cannot select the most powerful tail test, the fact that the multivariate test performs so well recommends it. We also note that the multivariate test is also the correct way to consider a range of different VaR levels without violating the distribution theory by data mining.

EMPIRICAL ANALYSIS

Trading Revenues

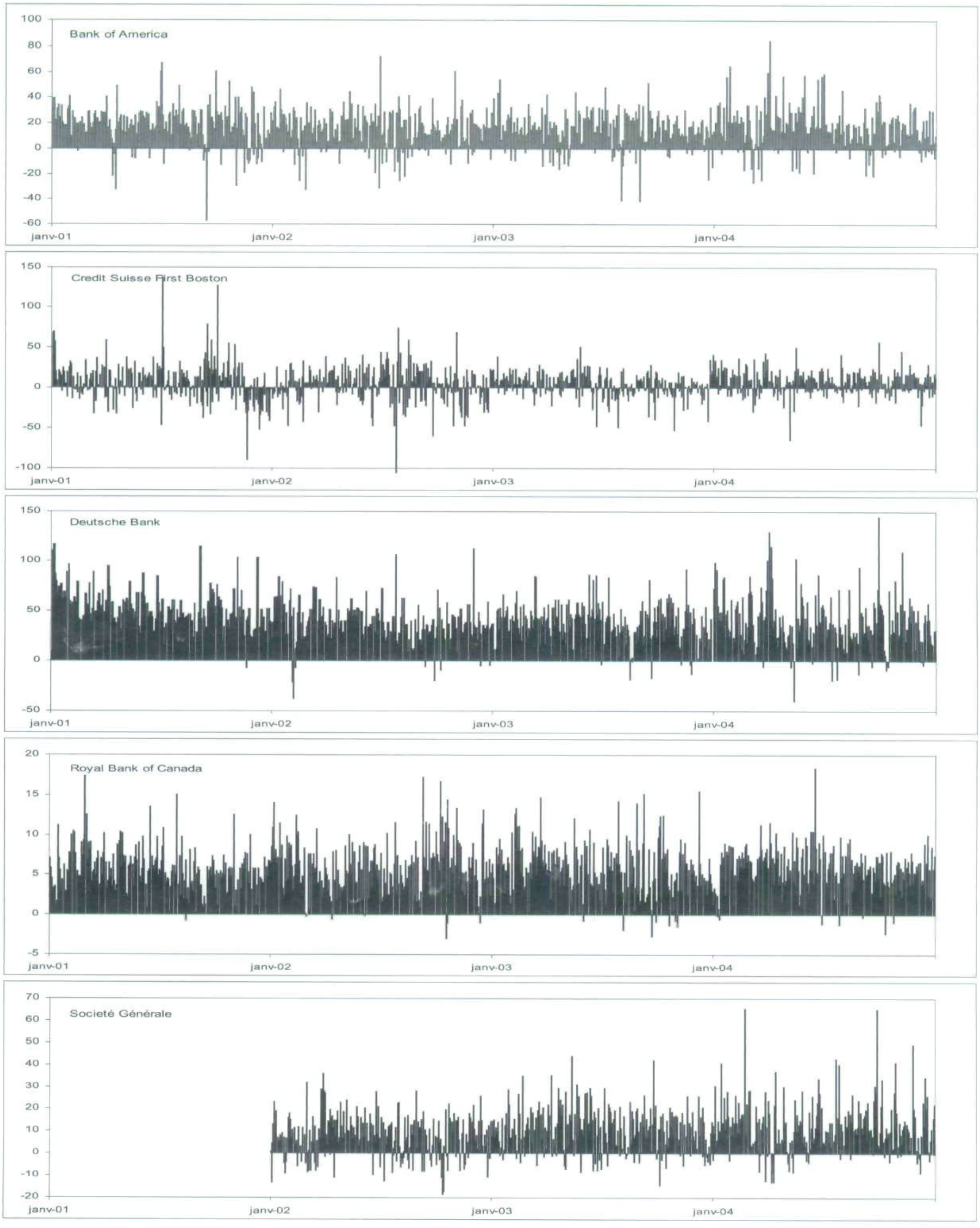
Our dataset includes actual daily trading revenues from five large international banks. The sample period is 2001–2004 for Bank of America, Credit Suisse First Boston, Deutsche Bank, and Royal Bank of Canada and 2002–2004 for Société Générale (see Exhibit 3). We extracted the data from the graphs displayed in the banks' annual reports using an innovative Matlab-based application. The data extraction process took the following steps:

- Step 1. Convert the original graph from the annual report (available as a PDF file) into a JPG file;
- Step 2. Import the JPG file into MATLAB and define it as an image;
- Step 3. Display the image in MATLAB;
- Step 4. Convert the graph scale into a MATLAB scale;
- Step 5. Add vertical lines on the image;
- Step 6. Zoom in and click on each data point. By doing so, we capture the two-dimensional coordinates of each data point;
- Step 7. Convert the MATLAB vertical coordinates into graph coordinates. At the end of the process we end up with daily values for each bank.[6]

The definition of the trading revenues is not always consistent across banks. For instance, both Royal Bank of Canada and Deutsche Bank disclose hypothetical revenues that are based on the previous day's portfolio allocation. Conversely, Bank of America, Credit Suisse First Boston, and Société Générale report actual revenues that are affected by intraday changes in the bank's holdings. Furthermore, all trading revenues include trading fees or commissions.

We present in Exhibit 4 some summary statistics on the sample firms and their trading revenues. We observe that four sample banks are comparable in size—they rank between 6th and 11th worldwide—and one bank is smaller (Royal Bank of Canada). The magnitude of trading revenues varies greatly across banks with Deutsche Bank having average trading revenue of €41 million per day which is several times larger than the other banks. Overall,

EXHIBIT 3
Daily Trading Revenues



Note: This exhibit displays the daily trading revenues of Bank of America (top panel), Credit Suisse First Boston, Deutsche Bank, Royal Bank of Canada, and Société Générale (lowest panel) between January 1, 2001, and December 31, 2004. All values are in million and expressed in local currencies.

trading revenues are highly volatile, right skewed, leptokurtic (and as a result not normal), moderately autocorrelated, and exhibit volatility clustering. For all banks, the magnitude of the most severe trading gain exceeds the magnitude of the most extreme trading loss, that is $|\min| < \max$.

Comparing VaR Methods

For each sample bank and each VaR method, we compute daily one-day ahead $VaR_{t+1|t}(p)$. For HS we follow conventional practice and use a 250-day moving window, while the other methods are estimated using an expanding window that utilizes all available data at each time point.[7] When estimating the GARCH and GARCH-T models we follow conventional practice as illustrated by the RiskMetrics model and ignore mean effects when estimating daily financial returns. The reason is quite simple: the average sample daily return on financial returns is statistically indistinguishable from zero unless one is dealing with extremely long sample periods. However, a quick perusal of Exhibit 3 indicates that the zero-mean assumption may not be the most appropriate given that the vast majority of trading revenues is positive (we note that banks include fee income along with trading returns, which should result in positive expected trading

EXHIBIT 4
Summary Statistics on Sample Banks and Trading Revenues

	Bank of America	Credit Suisse First Boston	Deutsche Bank	Royal Bank of Canada	Société Générale
Panel A: Sample Banks					
Country	US	Switzerland	Germany	Canada	France
Total Assets	1,082,243	1,016,050	1,170,277	398,981	1,000,728
Worldwide Rank	7	9	6	43	11
Domestic Rank	1	2	1	1	3
Panel B: Trading Revenues					
Sample Period	2001-2004	2001-2004	2001-2004	2001-2004	2002-2004
Number of Observations	1,008	1,031	989	998	776
Currency	USD	CHF	EUR	CAD	EUR
Mean	13.85	5.03	41.48	5.44	9.85
Variance	222.24	369.16	509.90	8.69	102.51
Skewness	0.123	0.240*	0.380*	0.457*	0.717*
Kurtosis	4.93*	9.90*	4.42*	4.19*	5.50*
Minimum	–57.39	–105.30	–39.92	–2.99	–18.56
Maximum	84.33	138.45	145.23	18.41	65.32
% Days with Neg. Revenues	13.00	37.05	3.13	2.30	14.56
Jarque–Bera Test	159.11*	1,510.30*	83.58*	93.27*	268.94*
Autocorrelation	0.064	0.124	0.423	0.099	0.119
ARCH-12	23.73*	44.99*	92.42*	32.32*	7.51

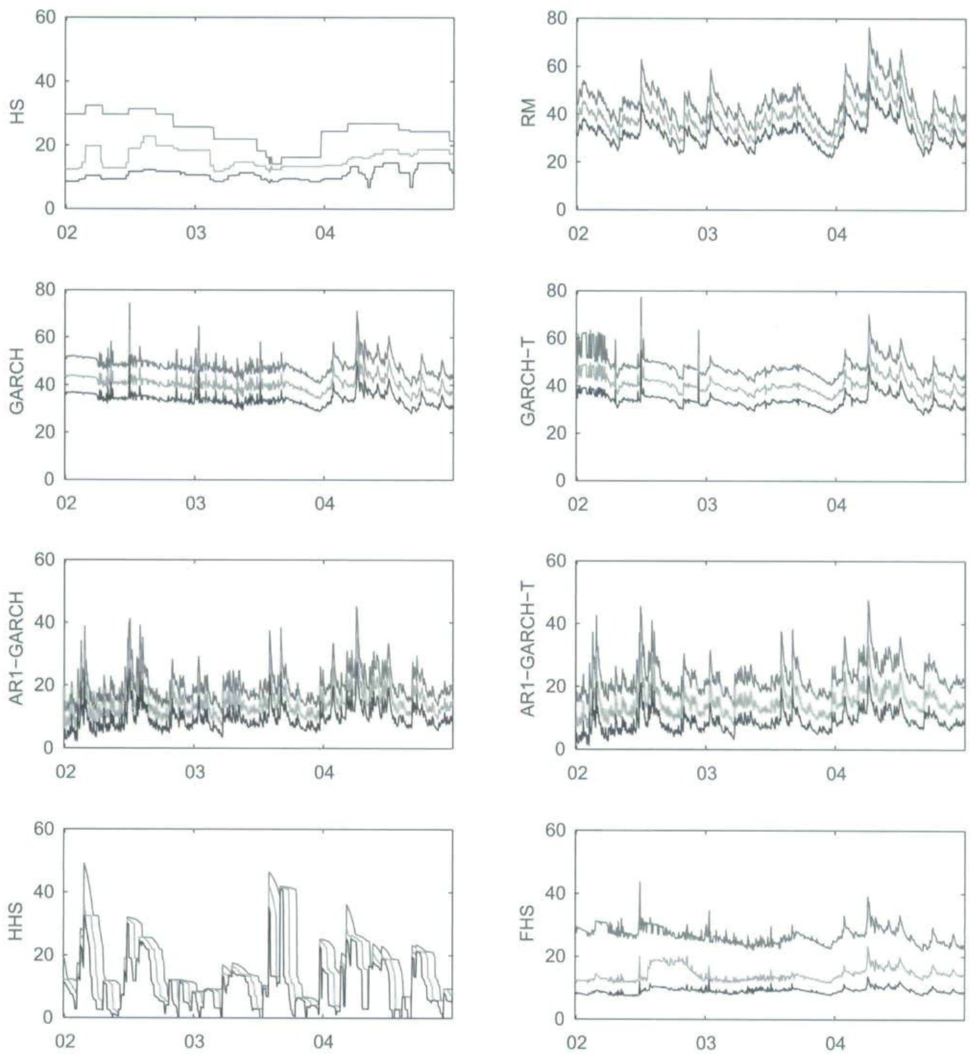
Note: This exhibit presents some summary statistics on the five sample banks (Panel A) and their trading revenues (Panel B). Banks' total assets (in millions USD) and rankings (based on total assets) are as of December 31, 2005, and were collected from Bankersalmanac.com. Figures on trading revenues are expressed in local currencies. For each trading revenues time series, we report the first four moments, minimum, maximum, the percentage of days with negative trading revenues, the Jarque–Bera normality test, the first-order autocorrelation coefficient, and the ARCH-12 test, which is a Ljung–Box autocorrelation test using 12 lags applied to the squared de-meaned revenues.

** denotes significant at the 5% confidence level.*

revenue). We thus also estimate the GARCH and GARCH-T models accounting for an AR(1) sample mean, and denote these models by AR1-GARCH and AR1-GARCH-T respectively.

We consider three coverage probabilities p within the range currently employed by banks and regulators: 1%, 2.5%, and 5%. As an illustration, we plot in Exhibit 5 the VaR estimates for Bank of America over the period January 2002–December 2004. We clearly see that different VaR methods lead to dramatically different VaR estimates, and this is true regardless of the considered coverage probability. HS estimates display sluggish dynamics whereas all the other estimates are much more reactive to volatility shocks. The inability of the HS method to respond to changes in volatility is a well-known deficiency of this approach (Christoffersen [2003], Campbell [2005], and Pritsker [2006]). Furthermore, RiskMetrics, GARCH, and GARCH-T estimates look much alike. Notice that,

EXHIBIT 5
VaR Estimates for Bank of America



Note: This exhibit displays for Bank of America the VaR(1%) (top line), VaR(2.5%) (intermediate line), and VaR(5%) (lower line) using six VaR methods: historical simulation (HS), RiskMetrics (RM), GARCH model using normal residuals (GARCH), GARCH model using Student-t residuals (GARCH-T), GARCH model with an AR(1) conditional mean and normal residuals (AR1-GARCH), an AR(1)-GARCH model using Student-t residuals (AR1-GARCH-T), and hybrid and filtered Historical Simulation (HHS and FHS). VaR figures are daily one-day ahead estimates, $VaR_{t+1|t}(p)$, computed using an expanding window including all data available before period t (HS uses a moving window of only the most recent 250 trading days covering the sample period $t-249$ to t).

EXHIBIT 6

Backtesting Results

	Bank of America		Credit Suisse First Boston		Deutsche Bank		Royal Bank of Canada		Société Générale	
	Viol %	p-value	Viol %	p-value	Viol %	p-value	Viol %	p-value	Viol %	p-value
Panel A: Univariate Coverage Test (p =1%)										
HS	0.79%	0.549	1.15%	0.676	1.35%	0.360	1.60%	0.127	0.95%	0.909
RM	**0.00%**	**0.000**	**1.79%**	**0.045**	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.001**
GARCH	**0.00%**	**0.000**	1.41%	0.280	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.001**
GARCH-T	**0.00%**	**0.000**	1.02%	0.946	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.001**
AR1-GARCH	1.58%	0.137	**2.18%**	**0.004**	**2.98%**	**0.000**	1.07%	0.850	0.38%	0.102
AR1-GARCH-T	0.92%	0.830	1.28%	0.450	1.08%	0.824	**0.27%**	**0.017**	**0.00%**	**0.001**
Hybrid HS	**2.77%**	**0.000**	**3.07%**	**0.000**	**2.71%**	**0.000**	**2.41%**	**0.001**	**2.47%**	**0.004**
Filtered HS	0.79%	0.549	1.67%	0.088	1.76%	0.061	**2.27%**	**0.003**	**0.19%**	**0.022**
Panel B: Univariate Coverage Test (p = 2.5%)										
HS	3.30%	0.179	2.18%	0.554	**3.79%**	**0.037**	3.07%	0.331	1.71%	0.219
RM	**0.26%**	**0.000**	2.31%	0.723	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
GARCH	**0.26%**	**0.000**	2.05%	0.405	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
GARCH-T	**0.26%**	**0.000**	2.05%	0.405	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
AR1-GARCH	3.03%	0.362	2.82%	0.578	**5.01%**	**0.000**	**3.88%**	**0.025**	**0.95%**	**0.009**
AR1-GARCH-T	3.03%	0.362	2.56%	0.914	**3.79%**	**0.037**	3.21%	0.234	**0.76%**	**0.003**
Hybrid HS	**4.22%**	**0.006**	**4.23%**	**0.005**	**4.19%**	**0.007**	**4.01%**	**0.015**	3.61%	0.125
Filtered HS	3.30%	0.179	2.56%	0.914	**6.90%**	**0.000**	**4.55%**	**0.001**	1.33%	0.060
Panel C: Univariate Coverage Test (p = 5%)										
HS	5.01%	0.987	4.35%	0.397	**8.39%**	**0.000**	5.88%	0.281	3.61%	0.125
RM	**0.53%**	**0.000**	**3.46%**	**0.037**	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
GARCH	**0.26%**	**0.000**	**2.81%**	**0.002**	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
GARCH-T	**0.26%**	**0.000**	**2.95%**	**0.004**	**0.00%**	**0.000**	**0.00%**	**0.000**	**0.00%**	**0.000**
AR1-GARCH	5.41%	0.610	4.48%	0.499	**9.90%**	**0.000**	6.15%	0.163	**2.47%**	**0.003**
AR1-GARCH-T	5.54%	0.502	4.74%	0.734	**10.3%**	**0.000**	6.02%	0.216	**2.09%**	**0.001**
Hybrid HS	**6.99%**	**0.017**	**6.91%**	**0.020**	6.50%	0.074	**6.95%**	**0.020**	6.46%	0.140
Filtered HS	5.67%	0.405	4.87%	0.863	**12.9%**	**0.000**	**7.09%**	**0.014**	**1.90%**	**0.000**
Panel D: Multivariate Coverage Test (p =1%, 2.5%, 5%)										
HS	-	**0.035**	-	0.391	-	**0.000**	-	0.267	-	0.193
RM	-	**0.000**	-	**0.000**	-	**0.000**	-	**0.000**	-	**0.000**
GARCH	-	**0.000**	-	**0.000**	-	**0.000**	-	**0.000**	-	**0.000**
GARCH-T	-	**0.000**	-	**0.002**	-	**0.000**	-	**0.000**	-	**0.000**
AR1-GARCH	-	0.323	-	**0.000**	-	**0.000**	-	**0.029**	-	**0.009**
AR1-GARCH-T	-	0.419	-	0.556	-	**0.000**	-	**0.001**	-	**0.000**
Hybrid HS	-	**0.000**	-	**0.000**	-	**0.001**	-	**0.003**	-	**0.012**
Filtered HS	-	0.096	-	0.074	-	**0.000**	-	**0.003**	-	**0.000**

Note: This exhibit displays for each bank and each VaR method the violation rate, i.e., the number of exceptions divided by the total number of observations, and the p-value associated with the employed coverage test. The VaR methods are historical simulation (HS), RiskMetrics (RM), GARCH model using normal residuals (GARCH), GARCH model using Student-t residuals (GARCH-T), AR(1)-GARCH model using normal residuals (AR1-GARCH), AR(1)-GARCH model using Student-t residuals (AR1-GARCH-T) and hybrid and filtered historical simulation (Hybrid HS and Filtered HS). Panels A, B, and C present the results for the univariate coverage test of Kupiec [1995] with a 1%, 2.5%, and 5% coverage probability, respectively. Panel D presents the results for the multivariate coverage test which is based jointly on the 1%, 2.5%, and 5% coverage probabilities. Daily one-day ahead $VaR_{t+1|t}(p)$ are computed using an expanding window including all data available before period t (HS uses a moving window of only the most recent 250 trading days covering the sample period $t - 249$ to t). The coverage tests are computed using three years of data, two years for Société Générale. Bold figures denote cases where the null hypothesis that the VaR model is well specified can be rejected at the 5% confidence level.

by construction, the three VaR lines in Exhibit 5 never cross each other.

We present in Exhibit 6 the results for the univariate (Panels A-C) and multivariate (Panel D) coverage tests for each bank and each VaR method. For the univariate test, we display the coverage probability, the actual violation rate, and the p-value, whereas for the multivariate test, we report the associated *p*-value. We notice that more than 30% of the time, univariate tests based on different coverage probabilities provide inconsistent results. Taking Royal Bank of Canada as an example, we see that the AR1-GARCH-T VaR model is strongly rejected (*p*-value = 0.01) with a 1% coverage probability, but it is unambiguously accepted with the 2.5% (*p*-value = 0.234) and

5% (p-value = 0.216) coverage probabilities. Indeed, when using a univariate coverage test based on a probability p, it is always possible that a misspecified VaR model leads (by chance) to the right VaR(p) whereas other VaR estimates are off by a substantial margin. Differently, our multivariate test in Panel D provides a single "reject/accept" signal for the entire left tail of the distribution.

A notable finding in Exhibit 6 is the seemingly good performance of the HS technique. However, as shown in our Monte-Carlo experiments in Exhibit 2, unconditional coverage tests lack power to reject HS when the number of observations used to compute VaR is small relative to the number of out-of-sample observations. It is interesting that HS is not rejected by the unconditional coverage tests for Bank of America, but because the 1% coverage is a little low and the 2.5% coverage is a little high, the multivariate test is able to reject the model because the trading revenue fell between the 1% and 2.5% VaR numbers on 2.51% (3.30%–0.79%) of trading days while we expect to only see this 1.5% of the time if HS is the correct VaR model.

We find that no VaR method yields satisfactory results for all sample banks. In Panel C, we see that any method that is not rejected for Credit Suisse First Boston is systematically rejected for Deutsche Bank, and vice versa. We do observe, however, that both RiskMetrics and hybrid HS perform quite poorly and are rejected in virtually every situation. Overall, we find that the best performing VaR methods are two parametric VaR methods, namely the AR1-GARCH-T and filtered HS. This result stands in contrast with the proprietary VaR methodologies currently implemented by our sample banks. Indeed, Pérignon and Smith [2008] reveal that all our sample banks, except Deutsche Bank, are currently using nonparametric methods.

CONCLUSION

We have presented a novel approach for testing VaR estimation techniques. We have shown that backtesting based on multiple points on the left tail of trading revenue improves over the ability of univariate tests to reject misspecified VaR models. We have applied our backtesting framework using actual daily bank trading revenues for a sample of five international banks. Empirically, we have found that parametric VaR methods work best for bank trading revenues. Finally, one possible extension of our methodology would be to account for clusters of exceptions in the spirit of Christoffersen [1998].

APPENDIX

Derivation of $\hat{\theta}_i$

Differentiating the log-likelihood in Equation (3) with respect to θ_i and setting equal to zero gives:

$$\alpha = \frac{T - \Sigma_{j=1}^{K}(\Sigma_{t=1}^{T} J_{j,t})}{1 - 1'\hat{\theta}} = \frac{\Sigma_{t=1}^{T} J_{i,t}}{\hat{\theta}_i} \quad \text{for } i = 1,\ldots,K \quad \text{(A1)}$$

for some α. These first-order conditions imply that $\Sigma_{t=1}^{T} J_{i,t} = \alpha \hat{\theta}_i$ for $i = 1, \ldots, K$, which we substitute into the first part of Equation (A1), giving $\alpha = \frac{T - \alpha \Sigma_{i=1}^{K} \hat{\theta}_i}{1 - 1'\hat{\theta}}$. Rearranging gives $\alpha(1 - \Sigma_{i=1}^{K} \hat{\theta}_i) = T - \alpha \Sigma_{i=1}^{K} \hat{\theta}_i$, which simplifies to $\alpha = T$ and, in turn, gives the desired result $\theta_i = \frac{1}{T} \Sigma_{t=1}^{T} J_{i,t}$. [QED]

ENDNOTES

The authors warmly thank Editor Stephen Figlewski for his useful comments and suggestions.

[1]For a similar approach, see Boudoukh, Richardson, and Whitelaw [1998], Hull and White [1998], Daníelsson [2002], Engle and Manganelli [2004], Brooks, Clare, Dalle Molle, and Persand [2005], Bao, Lee, and Saltoglu. [2006], Kuester, Mittnik, and Paolella [2006], and Pritsker [2006]. Two exceptions are the studies by Jackson, Maude, and Perraudin [1997] and Barone-Adesi, Giannopolis, and Vosper [2002] that rely on actual portfolio weights for one anonymous bank and a sample of futures exchange members, respectively.

[2]Berkowitz and O'Brien [2002] state that the trading revenues disclosed by U.S. banks typically include fee income and other income not attributable to position taking. As a result, even if the trading portfolio of bank B perfectly mimics the S&P 500 and that method M is the best VaR method for the S&P 500, it does not always imply that method M is also the best VaR method for bank B. This problem is likely to be more acute when nontrading revenues account for a significant fraction of the reported revenues and when the fee component varies with the revenues in an unspecified way.

[3]To improve the power, some authors consider conditional tests, see Christoffersen [1998], Christoffersen and Pelletier [2004], Engle and Manganelli [2004], Smith [2007], and Berkowitz, Christoffersen, and Pelletier [2008].

[4]Using the convention $0^0 = 1$, the multivariate coverage test is always defined.

[5]Our version of the HS method is a simplified one. Standard HS requires for each particular position (unknown to us) the identification of all market risk factors. The method yields a hypothetical distribution of the changes in each position value, and after aggregation, in the trading portfolio value, from which the VaR can be computed.

[6]More details about the data extraction process can be found in Pérignon and Smith [2008].

[7]The results are relatively similar when we use the minimum 250 trading days to estimate the models required by the Amendment to the Basel Accord. However, particularly in the context of our parametric models, using such a short window artificially impedes the estimation efficiency, but the results are quite similar and are available on request.

REFERENCES

Bakshi, G., and G. Panayotov. "The Capital Adequacy Puzzle." Working paper, University of Maryland, 2006.

Bao, Y., T.-H. Lee, and B. Saltoglu. "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check." *Journal of Forecasting*, 25 (2006), pp. 101–128.

Barone-Adesi, G., K. Giannopoulos, and L. Vosper. "VaR without Correlations for Portfolios of Derivative Securities." *Journal of Futures Markets*, 19 (1999), pp. 583–602.

——. "Backtesting Derivative Portfolios with Filtered Historical Simulation." *European Financial Management*, 8 (2002), pp. 31–58.

Berkowitz, J. "Testing Density Forecasts, with Applications to Risk Management." *Journal of Business and Economic Statistics*, 19 (2001), pp. 465–474.

Berkowitz, J., and J. O'Brien. "How Accurate Are Value-at-Risk Models at Commercial Banks?" *Journal of Finance*, 57 (2002), pp. 1093–1111.

——. "Estimating Bank Trading Risk: A Factor Model Approach." In M. Carey and R.M. Stulz, Eds., *The Risk of Financial Institutions*. University of Chicago Press, 2007.

Berkowitz, J., P.F. Christoffersen, and D. Pelletier. "Evaluating Value-at-Risk Models with Desk-Level Data." Working paper, McGill University, 2008.

Bollerslev, T. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31 (1986), pp. 307–327.

Boudoukh, J., M. Richardson, and R.F. Whitelaw. "The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk." *Risk*, 11 (1998), pp. 64–67.

Brooks, C., A.D. Clare, J.W. Dalle Molle, and G. Persand. "A Comparison of Extreme Value Theory Approaches for Determining Value at Risk." *Journal of Empirical Finance*, 12 (2005), pp. 339–352.

Campbell, S.D. "A Review of Backtesting and Backtesting Procedures." Working paper, Federal Reserve Board, 2005.

Christoffersen, P. F. "Evaluating Interval Forecasts." *International Economic Review*, 39 (1998), pp. 841–862.

——. *Elements of Financial Risk Management*. Academic Press, 2003.

Christoffersen, P.F., and D. Pelletier. "Backtesting Value-at-Risk: A Duration-Based Approach." *Journal of Financial Econometrics*, 2 (2004), pp. 84–108.

Daníelsson, J. "The Emperor Has No Clothes: Limits to Risk Modelling." *Journal of Banking and Finance*, 26 (2002), pp. 1273–1296.

Engle, R.F., and S. Manganelli. "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles." *Journal of Business and Economic Statistics*, 22 (2004), pp. 367–381.

Ferreira, M.A., and J.A. Lopez. "Evaluating Interest Rate Covariance Models Within a Value-at-Risk Framework." *Journal of Financial Econometrics*, 3 (2005), pp. 126–168.

Hansen, B. "Autoregressive Conditional Density Estimation." *International Economic Review*, 35 (1994), pp. 705–730.

Hull, J.C., and A. White. "Incorporating Volatility Updating into the Historical Simulation Method for Value at Risk." *Journal of Risk*, 1 (1998), pp. 5–19.

Jackson, P., D.J. Maude, and W. Perraudin. "Bank Capital and Value-at-Risk." *Journal of Derivatives* (Spring 1997), pp. 73–111.

Jorion, P. *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd edition. McGraw-Hill, 2006.

——. "Bank Trading Risk and Systematic Risk." In M. Carey and R.M. Stulz Eds., *The Risk of Financial Institutions*. University of Chicago Press, 2007.

Kuester, K., S. Mittnik, and M.S. Paolella. "Value-at-Risk Prediction: A Comparison of Alternative Strategies." *Journal of Financial Econometrics*, 4 (2006), pp. 53–89.

Kupiec, P.H. "Techniques for Verifying the Accuracy of Risk Measurement Models." *Journal of Derivatives*, 3 (1995), pp. 73–84.

Pérignon, C., and D.R. Smith. "The Level and Quality of Value-at-Risk Disclosure by Commercial Banks." Working paper, HEC Paris and Simon Fraser University, 2008.

Pritsker, M. "The Hidden Dangers of Historical Simulation." *Journal of Banking and Finance*, 30 (2006), pp. 561–582.

Smith, D.R. "Conditional Backtesting of Value-at-Risk Models." Working paper, Simon Fraser University, 2007.