CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Evaluating Value-at-Risk Models with Desk-Level Data

Jeremy Berkowitz, Peter Christoffersen, and Denis Pelletier

CEA@Cass Working Paper Series

WP–CEA–12-2007

Evaluating Value-at-Risk Models with Desk-Level Data[#]

Jeremy Berkowitz – University of Houston
jberkowitz@uh.edu

Peter Christoffersen – McGill University
peter.christoffersen@mcgill.ca

Denis Pelletier – North Carolina State University[*]
denis_pelletier@ncsu.edu

July 31, 2007

Abstract

We present new evidence on disaggregated profit and loss (P/L) and Value-at-Risk (VaR) forecasts obtained from a large international commercial bank. Our dataset includes daily P/L generated by four separate business lines within the bank. All four business lines are involved in securities trading and each is observed daily for a period of at least two years. Given this unique dataset, we provide an integrated, unifying framework for assessing the accuracy of VaR forecasts. We use a comprehensive Monte Carlo study to assess which of these many tests have the best finite-sample size and power properties. Our desk-level data set provides importance guidance for choosing realistic P/L generating processes in the Monte Carlo comparison of the various tests. The Caviar test of Engle and Manganelli (2004) performs best overall but duration-based tests also perform well in many cases.

JEL Codes: G21, G32

Keywords: Risk Management, Backtesting, Volatility, Disclosure.

[#] Christoffersen acknowledges financial support from FQRSC, IFM2, and SSRHC and Pelletier acknowledges financial support from the NCSU Enterprise Risk Management Initiative and the Edwin Gill Research Grant.

[*] Address correspondence to: Denis Pelletier, Department of Economics, College of Management, North Carolina State University, Raleigh, NC 27695-8110. Phone: 919-513-7408. Fax: 919-515-5613.

1. Introduction

Value-at-Risk (VaR) is arguably the leading measure of portfolio risk in use at major commercial banks. Despite criticisms of VaR's statistical properties (e.g. Artzner, Delbaen, Eber, and Heath (1999)) and theoretical concerns that widespread usage of VaR could increase systemic risk (Basak and Shapiro (2002)), financial institutions and their regulators increasingly rely on VaR as a measure of portfolio risk. Recent work by Danielsson et al (2005) suggest that in practice some of the arguments against VaR may indeed not be very important.

Under the "internal models approach" of the Basle Accord on banking (Basle Committee on Banking Supervision, 1996), financial institutions have the freedom to specify their own model to compute their Value-at-Risk. The accuracy of VaR models can be checked and must be checked by assessing the accuracy of the forecasts—a procedure known as backtesting.

Model validation in general and backtesting in particular is an important component of the Supervisory Review Process (the Second Pillar) in Basel II (Basle Committee on Banking Supervision, 2004). The lively public policy debate sparked by Basle II has focused attention on banks' procedures for backtesting. While no particular technique for backtesting is currently suggested in Basle II, the potential for a supervisor endorsed backtesting technique has clear implications for banks who need to apply VaR models that can pass the supervisors' tests. It is thus crucially important for institutions and regulators alike to assess the quality of the methods employed.

In this paper, we provide further empirical evidence on the accuracy of actual VaR models in forecasting portfolio risk. We obtained the daily profit and loss (P/L) generated by four separate business lines—or desks—from a large, international commercial bank. Each of the business line's P/L series is observed daily for a period of more than two years. For two of the business lines, we have over 600 daily observations while for the other two we have over 800 observations yielding a panel of 2,930 observations. While of interest in its own, the desk-level data set also provides importance guidance for choosing realistic P/L generating processes in our subsequent Monte Carlo comparison of the various backtesting methods.

All four business lines are involved in securities trading but the exact nature of each business line is not known to us. Each series is constructed and defined in a consistent manner but the series are normalized to protect the bank's anonymity.

In addition to the daily P/L data, we obtained the corresponding daily, 1-day ahead VaR forecasts computed using Historical Simulation. For each business line within the bank, and for each day, the VaR forecasts are estimates of the 1% lower tail. Our data set complements that of Berkowitz and O'Brien (2002) who obtained daily bank-wide P/L and VaR data, but who were not able to obtain any information on separate business lines within the same bank. In recent work, Perignon, Deng and Wang (2006), and Perignon and Smith (2006) also analyze bank-level VaRs. They find that one-day ahead VaR based on Historical Simulation is the industry standard. For the longer horizons required by supervisory bodies, such as ten-day ahead, banks typically use the square root of the horizon to scale the one-day ahead VaR.

Given our rich dataset, we provide an integrated, unifying framework for assessing the accuracy of VaR forecasts. Our approach includes the existing tests proposed Christoffersen (1998) and Christoffersen and Pelletier (2004) as special cases. In addition, we describe some new tests which are suggested by our framework.

In order to provide some guidance as to which of these many tests have the best finite-sample size and power properties, we conduct a thorough horserace in a Monte Carlo experiment where the P/L generating processes are motivated by the properties of the real-life P/Ls.

Testing the accuracy of a Value-at-Risk (VaR) model is based on the observation that the VaR forecast is a (one-sided) interval forecast. Violations – the days on which portfolio losses exceed the VaR – should therefore be unpredictable. In particular, the violations form a martingale difference sequence. The martingale hypothesis has a long and distinguished history in economics and finance (Durlauf (1991)). Related work dates back at least to the random walk theory of stock prices. The risk-neutral pricing methods of Harrison and Kreps (1979) and Harrison and Pliska (1981) are based on the martingale representation theorem.

As a result of this extensive toolkit, we are able to cast all existing methods of evaluating VaR under a common umbrella of martingale tests. This immediately

suggests several testing strategies. The most obvious is a test of whether any of the autocovariances are nonzero. The standard approach to test for uncorrelatedness is by estimating the sample autocovariances or sample autocorrelations. In particular, we suggest the well-known Ljung-Box test of the violation sequence's autocorrelation function.

The second set of tests are inspired by Campbell and Shiller (1987) and Engle and Manganelli (2004). If the violations are a martingale difference sequence, then they should be uncorrelated by any transformation of the variables available when the VaR is computed. It suggests a regression of the violations/non-violations on their lagged values and lagged variables such as previous VaRs.

A third set of tests are adapted from Christoffersen and Pelletier (2004) who focus on hazard rates and durations. These tests are based on the observation that the number of days separating the violations (i.e., the durations) should be unpredictable.

Lastly, a fourth set of tests is taken from Durlauf (1991). He derives a set of tests of the martingale hypothesis based on the spectral density functions. This approach has several features to commend it. Unlike variance ratio tests, spectral tests have power against any linear alternative of any order. Spectral density tests have power to detect any second moment dynamics. Variance ratio tests are typically not consistent against all such alternatives.

Because the violation of the VaR is, by construction, a rare event, the effective sample size in realistic risk management settings can be quite small. It follows that we can't rely on the asymptotic distribution of the tests to conduct inference. We instead rely on Dufour (2006)'s Monte Carlo testing technique which yields tests with exact level, irrespective of the sample size and the number of replications used. Our results suggest that the Caviar test of Engle and Manganelli (2004) performs best overall but that duration-based tests also perform well in many cases.

In summary, the key contribution of the paper is to use the unique, real-life, desk-level P/L data to construct realistic data generating processes for a thorough Monte Carlo comparison of existing and new VaR backtesting methods which we nest in a common framework.

The paper proceeds as follows. In Section 2 we present some new evidence on desk-level P/Ls and VaRs from a large international bank and we discuss the pros and cons of market risk management using Historical Simulation. Section 3 gives an overview of existing methods for backtesting VaR estimates and it suggests a few new approaches as well. Section 4 presents the results of a detailed horserace among the methods in terms of size and power properties in finite sample. Section 5 presents the empirical results of applying the various testing methods to our desk-level data sample. Section 6 concludes.

2. Desk Level P&L and VaR at a Commercial Bank

We collected the daily profit and loss (P/L) generated by four separate business lines from a large, international commercial bank. All four business lines are involved in securities trading but the exact nature of each business line is not known to us. Each series is constructed and defined in a consistent manner but they series are normalized to protect the bank's anonymity. We do not observe the aggregate P&L summed across the business desks.

In addition to the daily revenue data, we obtained the corresponding 1-day ahead Value-at-Risk forecasts. The VaR forecasts are estimates of the 99% lower tail and are calculated for each business line within the bank. The bank relies on Historical Simulation for computing VaR.

Suppose revenue is denoted by R_t. The p percent Value-at-Risk (VaR) is the quantity VaR_t such that

$$F\left(R_{t+1} < VaR_t \mid \Omega_t\right) = p \tag{1}$$

where Ω_t is the risk manager's time-t information set. The VaR is the p^{th} percentile of the return distribution. The probability p is referred to as the coverage rate. By definition, the coverage rate is the probability that the lower tail VaR will be exceeded on a given day.

In our dataset the tail percentile of the bank's VaR is set at p =.01 which yields a one-sided, 99% confidence interval. This is quite far in the tail but is typical of the VaR forecasts at commercial bank (e.g., Berkowitz and O'Brien (2002)).

The daily P/L (dashed) and associated VaR (solid) are plotted over time in Figure 1. Business line 1 is observed from January 2, 2001 through June 30, 2004, business line 2 is observed from April 2, 2001 and lines 3 and 4 from January 3, 2002. Several interesting observations are apparent in Figure 1. First, notice that bursts of volatility are apparent in each of the P/L series (e.g. mid-sample for line 1 and end-sample for line 2) but these bursts are not necessarily synchronized across business lines. Second, note the occasional and very large spikes in the P/Ls. These are particularly evident for line 1 and 2. Third, the bank VaRs exhibit considerable short-term variability (line 3), sometimes they show persistent trends away from the P/Ls (line 1) and even what looks like regime-shifting without corresponding moves in the associated P/L (line 2). This can happen in a case where the bank took a large position on an asset that had volatile P/L in the recent past, thus not affecting the current business line's P/L but increasing its Historical Simulation VaR which is based on reconstructed—or pseudo—P/L series.

Table 1 reports the first four sample moments of the P/Ls and VaRs along with the exact number of daily observations. Of particular interest are the skewness and kurtosis estimates. Skewness is evident in business line 1 (negative) and line 2 (positive) but much less so in business lines 3 and 4. Excess kurtosis is evident in all four business lines and dramatically so in lines 1 and 2. The skewness statistics confirm the occasional spikes in the P/Ls in Figure 1. For completeness, the descriptive statistics for the VaRs are also reported in Table 1.

The occasional bursts of volatility apparent in the P/Ls in Figure 1 are explored further in Figure 2 where we demean the P/Ls and plot their daily absolute values over time. While the spikes in P/Ls dominate the pictures, episodes of high volatility is evident in each of the series, although perhaps less so in business line 3.

Violations of the VaR should be happening randomly over time and should not be clustered over time. For example, if it can be predicted that volatility will be increasing in the near future, then the model used to compute the VaR should take this information into account and adjust the VaR accordingly. In other words, if the model used to compute the VaR is correctly specified, then violations should only happen because of unpredictable events.

3. A Unified Framework for VaR Evaluation

Under the 1996 Market Risk Amendment to the Basle Accord effective in 1998 qualifying financial institutions have the freedom to specify their own model to compute their Value-at-Risk. It thus becomes crucially important for regulators to assess the quality of the models employed by assessing the forecast accuracy—a procedure known as "backtesting".

The accuracy of a set of VaR forecasts can be assessed by viewing them as one-sided interval forecasts. A violation of the VaR is defined as occurring when the *ex post* return is lower than the VaR. Specifically, we define violations

$$(2) \qquad I_{t+1} = \begin{cases} 1, & if\ R_{t+1} < VaR_t(p) \\ 0, & otherwise \end{cases}$$

i.e. a sequence of zeros and ones. By definition, the conditional probability of violating the VaR should always be

$$(3) \qquad \Pr(I_{t+1} = 1 \mid \Omega_t) = p$$

for every time-t. The critical upshot is that no information available to the risk manager at the time the VaR was made should be helpful in forecasting the probability that the VaR will be exceeded. If it were, then this information should be incorporated into constructing a better VaR with unpredictable violations. Below, we will refer to tests of this property as conditional coverage (CC) tests.

A. Autocorrelation Tests

Christoffersen (1998) notes that property (3) implies that any sequence of violations, $\{I_t\}$, should be an i.i.d. Binomial random variable with mean p. In order to formally test this, Christoffersen (1998) embeds the null hypothesis of an i.i.d. Binomial within a general first-order Markov process.

If $\{I_t\}$ is a first-order Markov process the one-step-ahead transition probabilities $pr(I_{t+1} \mid I_t)$ are given by

$$(4) \qquad \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

where π_{ij} is the transition $pr(I_{t+1} = j \mid I_t = i)$.

Under the null, the violations have a constant conditional mean which implies the two linear restrictions, $\pi_{01} = \pi_{11} = p$. A likelihood ratio test of these restrictions can be computed from the likelihood function

$$L\left(I;\pi_{01},\pi_{11}\right) = \left(1-\pi_{01}\right)^{T_0-T_{01}} \pi_{01}^{T_{01}} \left(1-\pi_{11}\right)^{T_1-T_{11}} \pi_{11}^{T_{11}}$$

where T_{ij} denotes the number of observations with a j following a i and T_i is the number of i is the number of ones or zeros in the sample.

We note that all the tests we consider are carried out conditioning on the first observation. While the tests all have know asymptotic distributions we will rely on finite sample p-values as discussed below.

In this paper, we extend and unify the existing tests by noting that the de-meaned violations $\{ I_t - p \}$ form a martingale difference sequence (m.d.s.). By definition of the violation, equations (2)-(3) immediately imply that

$$E\left[(I_{t+1} - p) \mid \Omega_t\right] = 0 \tag{5}$$

where Ω_t is the information set of the risk manager up to time-*t*. The de-meaned violations form an m.d.s. with respect to the time-t information set. This will be an extremely useful property because it implies that the violation sequence is uncorrelated at *all leads and lags*.

For any variable Z_t in the time-t information set, we then must have,

$$E\left[(I_{t+1} - p) \otimes Z_t\right] = 0 \tag{6}$$

which is familiar as the basis of GMM estimation.

This motivates a variety of tests which focus on the white noise or martingale property of the sequence. Since white noise has zero autocorrelations at all leads and lags, the violations can be tested by calculating statistics based on the sample autocorrelations.

Thus, specifying Z_t to be the most recent de-meaned violation, we have

$$E\left[(I_{t+1} - p)(I_t - p)\right] = 0\,. \tag{7}$$

The violation sequence has a first-order autocorrelation of zero, under the null. It is this property which is exploited by the Markov test of Christoffersen (1998).

More generally, if we set $Z_t = I_{t-k}$ for any $k \geq 0$,

(8) $$E\left[(I_{t+1} - p)(I_{t-k} - p)\right] = 0$$

which says that the de-meaned violation sequence is in fact white noise. We write this null hypothesis compactly as

(9) $$(I_{t+1} - p) \overset{iid}{\sim} \left(0,\, p(1-p)\right).$$

A natural testing strategy is to check whether any of the autocorrelations are not zero. Under the null all the autocorrelations are zero

$$H_0 : \gamma_k = 0, \quad k > 0$$

and the alternative hypothesis of interest is that

$$H_a : \gamma_k \neq 0, \quad \text{for some } k.$$

The Portmanteau or Ljung-Box statistics, for example, have known distribution which can be compared to critical values under the white noise null. The Ljung-Box statistic is a joint test of whether the first *m* autocorrelations are zero. We can immediately make this into a test of of a VaR model by calculating the autocorrelations of $(I_{t+1} - p)$ and then calculating

$$LB(m) = T(T+2)\sum_{k=1}^{m} \frac{\gamma_k^2}{T-k}$$

which is asymptotically chi-square with *m* degrees of freedom.

We may also want to consider whether violations can be predicted by including other data in the risk manager's information set such as past returns. Under the null hypothesis, it must be that

(10) $$E\left[(I_{t+1} - p)\, g(I_t, I_{t-1}, ..., R_t, R_{t-1}, ...)\right] = 0\,.$$

for any non-anticipating function $g(\cdot)$.

In analogy with Engle and Manganelli (2004), we might consider the *n*th-order autoregression

(11) $$I_t = \alpha + \sum_{k=1}^{n} \beta_{1k} I_{t-k} + \sum_{k=1}^{n} \beta_{2k} g(I_{t-k}, I_{t-k-1}, \ldots, R_{t-k}, R_{t-k-1}, \ldots) + u_t$$

where we set $g(I_{t-k}, I_{t-k-1}, \ldots, R_{t-k}, R_{t-k-1}, \ldots) = VaR_{t-k+1}$ and n=1.

Estimating this autoregression by ordinary least squares would leave us having to deal with heteroskedasticity to make valid inference because the hit sequence is binary.

We instead assume that the error term u_t has a logistic distribution and we estimate a logit model. We can then test with a likelihood ratio test if the coefficients are statistically significant and whether $\Pr(I_t = 1) = e^{\alpha} /(1 + e^{\alpha}) = p$. We refer to this test as the Caviar test of Engle and Manganelli.

B. Hazard Rates and Tests for Clustering in Violations

Under the null that VaR forecasts are correctly specified, the violations should occur at random time intervals. Suppose the duration between two violations is defined as

$$D_i = t_i - t_{i-1} \tag{12}$$

where t_i denotes the day of the violation number i. The duration between violations of the VaR should be completely unpredictable. There is an extensive literature on testing duration dependence (e.g., Kiefer (1988), Engle and Russel (1998), Gourieroux (2000)) which makes this approach particularly attractive.

Christoffersen and Pelletier (2004) and Haas (2005) apply duration-based tests to the problem of assessing VaR forecast accuracy. In this section we expand upon their methods. The duration-based tests can be viewed as another procedure for testing whether the violations form a martingale difference sequence.

Using the Binomial property, the probability of a violation next period is exactly equal to $pr\left(D_i = 1\right) = pr\left(I_{t+1} = 1\right) = p$. The probability of a violation in d periods is

$$pr\left(D_i = d\right) = pr\left(I_{t+1} = 0, I_{t+2} = 0, ..., I_{t+d} = 1\right). \tag{13}$$

Under the null of an accurate VaR forecast, the violations are distributed

$$I_{t+1} \sim iid\ \left(p, p(1-p)\right).$$

This allows us to rewrite (13) as

$$pr(D_i = d) = (1-p)...(1-p)(p)$$

$$= \left(1-p\right)^{d-1} p. \tag{14}$$

Equation (14) says that the density of the durations declines geometrically under the null hypothesis.

A more convenient representation of the same information is given by transforming the geometric probabilities into a flat function. The hazard rate defined as

$$\lambda(D_i) = \frac{pr(D_i = d)}{1 - pr(D_i < d)} \tag{15}$$

is such a transformation. Writing out the hazard function $\lambda(D_i)$ explicitly

$$\frac{(1-p)^{d-1} p}{1 - \sum_{j=0}^{d-2} (1-p)^j p} = p \tag{16}$$

collapses to a constant after expanding and collecting terms.

We conclude that under the null, the hazard function of the durations should be *flat* and equal to p. Tests of this null are constructed by Christoffersen and Pelletier (2004). They consider alternative hypothesis under which the violation sequence, and hence the durations, display dependence or clustering. The only (continuous) random distribution is the exponential, thus under the null hypothesis the distribution of the durations should be

$$f_{\exp}(D; p) = pe^{-pD}$$

The most powerful of the two alternative hypotheses they consider is that the durations follow a Weibull distribution where

$$f_W(D; a, b) = a^b b D^{b-1} \exp^{-(aD)^b}$$

This distribution is able to capture violation clustering. When $b < 1$, the hazard, i.e. the probability of getting a violation at time D_i given that we did not up to this point, is a decreasing function of D_i.

It is also possible to capture duration dependence without resorting to the use of a continuous distribution. We can introduce duration dependence by having non-constant probabilities of a violation,

$$\begin{aligned} pr(D_i = d) &= pr(I_{t+1} = 0, I_{t+2} = 0, ..., I_{t+d} = 1) \\ &= (1-p_1)(1-p_2)\cdots(1-p_{d-1})\, p_d \end{aligned}$$

where

$$p_d = pr(I_{t+d} = 1 \mid I_{t+d-1} = 0, \ldots, I_{t+1} = 0)$$

In this case, one must specify how these probabilities p_d vary with d. We will set

$$p_d = ad^b$$

with $b \leq 0$ in order to implement the test. We refer to this as the Geometric test below.

Except for the first and last duration the procedure is straightforward, we just count the number of days between each violation. We then define a binary variable C_i which tracks whether observation *i* is censored or not. Except for the first and last observation, we always have $C_i = 0$. For the first observation if the hit sequence starts with 0 then D_1 is the number of days until we get the first hit. Accordingly $C_1 = 1$ because the observed duration is left-censored. The procedure is similar for the last duration. If the last observation of the hit sequence is 0 then the last duration, $D_{N(T)}$, is the number of days after the last 1 in the hit sequence and $C_{N(T)} = 1$ because the spell is right-censored.

The contribution to the likelihood of an uncensored observation is its corresponding p.d.f. For a censored observation, we merely know that the process lasted at least D_1 or $D_{N(T)}$ days so the contribution to the likelihood is not the p.d.f. but its survival function $S(D_i) = 1 - F(D_i)$. Combining the censored and uncensored observations, the log-likelihood is

$$\ln L(D;a,b) = C_1 \ln S(D_1) + (1 - C_1) \ln f(D_1) + \sum_{i=2}^{N(T)-1} \ln f(D_i) + C_{N(T)} \ln S(D_{N(T)}) + (1 - C_{N(T)}) \ln f(D_{N(T)})$$

Once the durations are computed and the truncations taken care of, then the likelihood ratio tests can be calculated in a straightforward fashion. The null and alternative hypotheses for the test is

$$H_0 : \quad b = 1 \text{ and } a = p$$
$$H_a : \quad b \neq 1 \text{ or } a \neq p$$

The only added complication is that the ML estimates are no longer available in closed form, they must be found using numerical optimization.

C. Spectral Density Tests

Another method for testing the martingale property is to examine the shape of the spectral density function. There is a long standing literature on using the spectral density for this purpose because white noise has a particularly simple representation in the frequency domain -- its spectrum is a flat line (e.g., Durlauf (1991)). Statistical tests are constructed by examining if the sample spectrum is "close" to the theoretical flat line.

The spectral density function is defined as a transformation of the autocovariance sequence,

$$(17) \qquad f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-ik\omega} .$$

For a white noise process, all the autocovariances equal zero for any $k \neq 0$. This means that for the hit sequence the spectral density collapses to

$$(18) \qquad f(\omega) = \frac{1}{2\pi} p(1-p)$$

for all $\omega \in [0, \pi]$.

The spectral density function is constant and proportional to the variance. Equivalently, the spectral *distribution function* is a 45° line. The asymptotic theory centers on the convergence of the random, estimated spectral density function using a functional central limit theorem.

The sample spectrum (or periodogram) is given by replacing the population autocovariances with the finite-sample estimates,

$$(19) \qquad \hat{f}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \hat{\gamma}_k e^{-ik\omega}$$

which should be approximately a flat line.

It is often convenient to de-mean the sample spectral density and take the partial sums

$$(20) \qquad \hat{U}(\omega) = \sum_{\omega=0}^{\pi} \left(\frac{\hat{f}(\omega)}{\hat{\sigma}^2} - \frac{1}{\pi} \right)$$

for each frequency $\omega \in [0, \pi]$. The $\hat{U}(\omega)$ are deviations of the sample spectral distribution from the 45 degree line. If the violations are white noise, the deviations should be small.

Durlauf (1991) derives the asymptotic distribution of a variety of statistics based on these deviations. The Cramér-Von Mises (CVM) test statistic is the sum of squared deviations

$$(21) \qquad CVM = \sum_{\omega=0}^{\pi} \hat{U}(\omega)^2$$

and it converges to a known distribution whose critical values can be tabulated numerically.

Another common test statistic dates to Bartlett, who showed the supremum

$$(22) \qquad \sup_{\omega} \hat{U}(\omega)^2$$

converges to the Kolmogorov-Smirnov (KS) statistic.

These test statistics have several attractive features. Unlike some tests of white noise (e.g., variance ratio tests), spectral tests have power against any linear alternative of any order. That is, the test has power to detect any second moment dynamics (see Durlauf, (1991)). Both the CVM and KS statistics diverge asymptotically if I_t is any stationary process which is not white noise.

D. Multivariate Tests

The tests described above only use information about one hit sequence at a time. In a case where we have Values-at-Risk and P/L for different business lines we might be interested in jointly testing if property (3) holds for all the hit sequences. In this way, we could hope that the tests would have more power because we would be effectively increasing the sample size.

A first approach to study simultaneously the hit sequences could be to simply "stack" the series together, assuming that the series are independent across desks. For the Ljung-Box test we could compute the autocorrelations using all the series, treating them as multiple non-overlapping sequences from the same underlying process. For likelihood-based tests such as the duration tests in Section B, we could sum the log-likelihoods for each series.

A second approach would be to capture the dependence across the series by considering multivariate generalizations of the previous tests. Recall from equation (3) that no information available to the risk manager at the time the VaR is made should be

helpful in forecasting a VaR violation. Thus, if the VaR models are correctly specified, then past observations from the hit sequence of one business line, which are clearly available to the risk manager, should not help predict violations of another business line. One could then consider using multivariate Box-Pierce tests as in Lütkepohl (1993, Section 4.4), or multivariate spectral test as in Paramasamy (1992). Duration-based tests could be extended by considering competing risk models following Cameron and Trivedi (2005, Chapter 19). Perhaps the easiest way to use information from all the business lines is offered by the regression approach of the caviar test. We can simply use variables form other business lines, such as their P/L's as explanatory variables. The Conditional Coverage test would then consist in testing that the coefficients of the explanatory variables (such as P/L's) are zero and the probability of getting a violation is equal to *p*.

4. Size and Power Properties

Given the large variety of backtesting procedures surveyed in Section 3, it is important to give users guidance as to their comparative size and power properties in a controlled setting.

A. Effective Size of the Tests

In order to assess the size properties of the various methods, we simulate i.i.d. Bernoulli samples with probabilities p = 1% and 5% respectively. For each Bernoulli probability, we consider several different sample sizes, from 250 to 1500. Rejection rates under the null are calculated over 10,000 Monte Carlo trials. If the asymptotic distribution is accurate in the sample sizes considered then the rejection frequencies should be close to the nominal size of the test, which we set to 10%. In the Caviar test we generate the required VaR regressors via a GARCH model with innovations that are independent of the simulated hit sequence. This way we perform a test that is true to the Caviar idea while ensuring that the null hypothesis is true.

Table 2 contains the actual size of the conditional coverage (CC) tests when the asymptotic critical values are used. The number of observations in each simulated sample is reported in the first column. The top panel shows the finite sample test sizes for a 1% VaR. We see that the LB(1) test tends to be undersized and the LB(5) tends to be

oversized. The Markov test tends to be undersized and the Weibull test tends to be oversized. The Geometric test is extremely oversized for the smallest sample. The Caviar test is undersized. The CVM test is undersized for the smallest sample size and oversized for the larger samples. Finally, the KS test has good size beyond the smallest sample sizes where it is undersized.

The results in the bottom panel cover the 5% VaR. In this case the LB(1) test is slightly undersized whereas the LB(5) is very close to the desired 10%. The Markov and Weibull tests are both oversized. The Geometric is somewhat undersized, whereas the Caviar, KS and CVM tests now are very close to the desired 10% level.

The overall conclusion from Table 2 is that for small sample sizes and for the 1% VaR which is arguably the most common in practice, the asymptotic critical values can be highly misleading. When computing power below we therefore rely on the Dufour (2006) Monte Carlo testing technique which is described in detail in Section 5.

B. Finite Sample Power of the Tests

In order to perform a power comparison, we use a flexible and simple GARCH specification as a model of the P/L process. GARCH models are some of the most widely used models for capturing variance dynamics in daily asset returns. See Andersen et al (2006) for a recent survey. We estimate the parameters for each business line separately in order to model the volatility persistence in each series.

The GARCH model allows for an asymmetric volatility response or "leverage effect". In particular, we use the NGARCH(1,1)-t(d) specification,

$$R_{t+1} = \sigma_{t+1}((d-2)/d)^{1/2} z_{t+1}$$
$$\sigma^2_{t+1} = \omega + \alpha\sigma^2_t(((d-2)/d)z_t - \theta)^2 + \beta\sigma^2_t$$

where R_{t+1} is the daily demeaned P/L and the innovations z_t are drawn independently from a Student's t(d) distribution. The Student-t innovations enable the model to capture some of the additional kurtosis.

Table 3 reports the maximum likelihood estimates from the GARCH model for each business line. As usual we get a small but positive α and a β much closer to 1. Variance persistence in this model is given by $\alpha(1+\theta^2)+\beta$. It is largest in business lines 2 and 4 which confirm the impression provided by Figure 2. The last three lines of

Table 3 report the log likelihood values for the four GARCH models along with the log likelihood values for the case of no variance dynamics, where $\alpha = \beta = \theta = 0$.

Looking across the four GARCH estimates we see that Desk 1 is characterized by a large α and small d which suggests are large conditional kurtosis. Desk 2 is characterized by high variance persistence and high unconditional kurtosis from the low d. Desk 3 has an unusually large negative θ which suggests that a positive P/L increases volatility by more than a negative P/L of the same magnitude. Desk 4 has an unusually large unconditional volatility and a relatively high persistence as noted earlier.

For the power simulation exercise, we will assume that the correct data-generating processes are the four estimated GARCH processes. We must also choose a particular implementation for the VaR calculation. Following industry practice (see Perignon and Smith (2006)) and the approach used by the bank that provided us with the VaR data in Figure 1, we rely on Historical Simulation or "bootstrapping". The Historical Simulation VaR on a certain day is simply the unconditional quantile of the past T_e daily observations. Specifically

$$VaR_{t+1}^{p} = percentile\left(\{R_s\}_{s=t-T_e+1}^{t}, 100\,p\right)$$

For the purposes of this Monte Carlo experiment, we choose T_e=250 corresponding to 250 trading days. The VaR coverage rate p we study is either 1% (as in Section 2) or 5%. We look at one-day ahead VaR again as in Section 2. When computing the finite-sample p-values we use 9,999 simulations and we perform 10,000 Monte Carlo simulations for each test. Section 5 provides the details of the p-value simulation.

Table 4 shows the finite sample power results for the 1% VaR from Historical Simulation for various samples sizes when using the GARCH DGP processes corresponding to each of the four business lines.

For all the sample sizes in all the four business lines in Table 4, the Caviar test performs the best. For business line 1, the LB(5), the KS and the CVM tests perform well also. For business line 2, the Geometric test also performs well. For business line 3 only the Caviar test has good power. For business line 4, the LB(5) and the KS tests perform well in addition to the Caviar test.

Consider next Table 5 which shows reports the finite sample power calculations for the 5% VaR. For business line 1 the LB(5) and the Caviar are best. For business line 2

the Caviar test is best for small samples and the Geometric best for larger samples. For business line 3 the power is again low everywhere except for in the Caviar test. For business line 4 the Caviar is again best for small samples and the Geometric is best for large samples.

Considering Table 4 and 5 overall it appears that the Caviar test is best for 1% VaR testing whereas for 5% VaR testing the Geometric test is sometimes better than Caviar. It is also important to note that in business line 3 where all the tests have trouble showing power only the Caviar test has a decent performance. Clearly, these results suggest that the Caviar test should be included in any arsenal of backtesting procedures.

Tables 4 and 5 provide a couple of other conclusions. First, it is clear that the LB(5) test is better than LB(1) and Markov test. This is perhaps to be expected as the dependence in the hit sequence is not of order 1 here. Second, the Geometric test is typically substantially better than the Weibull test. This is also to be expected as the latter wrongly assumes a continuous distribution for the duration variable.

Overall the power of the best tests investigated here is quite impressive, particularly considering the small samples investigated. Admittedly, the power numbers reported in Tables 4 and 5 are affected by the fact that we have conditioned the power calculations on being able to calculate the test in the first place. Samples where the tests cannot be computed, due to a lack of VaR hits, are omitted.

C. Feasibility Ratios

For transparency we report in Table 6 the fraction of simulated samples from Tables 4 and 5 where the each test is feasible. We only report sample sizes 250, 500, and 750 for the 1% VaR and 250 for the 5% VaR as the other sample sizes had no omitted sample paths in our experiment. Table 4 shows that only in the case of 1% VaR and samples of 250 observations is the issue non-trivial. In those cases the issue is most serious for the Weibull and Geometric tests. That conclusion also holds when considering the bottom panel in Table 6 which reports the fraction of feasible samples from the size calculations in Table 2.

5. Results for Desk-level Data

In Table 7 we report the results from applying our tests to the actual observed sequences of P/Ls and Historical Simulation VaRs from the four business lines. As in the power calculations above we make use of the Dufour (2006) Monte Carlo testing technique which yields tests with correct level, regardless of sample size.

For the case of a continuous test statistic, the procedure is the following. We first generate N independent realizations of the test statistic, $LR_i, i = 1,...,N$. We denote by LR_0 the test statistic computed with the original sample. Under the hypothesis that the risk model is correct, we know that the hit sequence is i.i.d. Bernoulli with the mean equal to the coverage rate. We thus benefit from the advantage of not having nuisance parameters under the null hypothesis.

We next rank $LR_i, i = 0,1,...,N$ in non-decreasing order and obtain the Monte Carlo p-value $\hat{p}_N\left(LR_0\right)$, where

$$\hat{p}_N\left(LR_0\right) = \frac{N\hat{G}\left(LR_0\right)+1}{N+1}$$

and

$$\hat{G}_N\left(LR_0\right) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(LR_i > LR_0\right).$$

The indicator function I(·) takes on the value one if true and the value zero otherwise. We reject the null hypothesis if $\hat{p}_N\left(LR_0\right)$ is less or equal than the prespecified significance level.

When working with binary sequences, there is a non-zero probability of obtaining ties between the test values obtained with the sample and the simulated data. The tiebreaking procedure is as follows: For each test statistic, $LR_i, i = 0,1,\ldots,N$, we draw an independent realization of a uniform distribution on the $[0;1]$ interval. Denote these draws by $U_i, i = 0,1,\ldots,N$. We obtain the Monte Carlo p-value by replacing $\hat{\mathrm{G}}_{\mathrm{N}}(\mathrm{LR}_0)$ with

$$\tilde{G}_N\left(LR_0\right) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(LR_i \le LR_0\right) + \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(LR_i = LR_0\right)\mathrm{I}\left(U_i \ge U_0\right)$$

There are two additional advantages of using a simulation procedure. The first is that possible systematic biases, for example arising from the use of a continuous distribution to study discrete processes, are accounted for since they will appear both in LR_0 and LR_i. The second is that Monte Carlo testing procedures are consistent even if the parameter value is on the boundary of the parameter space. The bootstrap procedures on the other hand could be inconsistent in this case.

In Table 7 we report the results from applying our tests to the actual observed sequences of P/Ls and VaRs from the four business lines. In addition to the eight univariate test analyzed in the Monte Carlo study in Tables 2-6, we add a multivariate Caviar test denoted CavMult in Table 7. The test is run for each business line using the hit sequence as the regressand, but it uses the ex-ante VaRs from all the four business lines as regressors.

We find no rejections in the first two business lines using the univariate tests, but note that the CavMult test rejects the VaR in business line 2. Four tests reject the VaR in business line 3. Note also that in business line 3, we were unable to calculate the Weibull and the Geometric tests. This is due to the fact that business line 3 only had one VaR hit in the sample as reported in Table 1. In business line 4, two of the tests reject the risk model.

Thus, when backtesting the actual VaRs from the four business unit, not surprisingly, we find it difficult to reject the Historical Simulation VaRs. This finding could of course have several explanations. First, the samples are short. Second, random chance, that is, by design we fail to reject 10% of the time. Third, the bank may be implementing Historical Simulation with some adjustments which make it difficult to reject by the tests considered here.

6. Conclusions

With the introduction of RiskMetrics, JP Morgan (1994) sparked a revolution in the field of market risk management. RiskMetrics has many redeeming features such as dynamic volatility and correlation modeled in a very parsimonious fashion. However other aspects such as the conditional normality assumption, and the difficulty of

aggregating VaRs across business lines has increased the popularity of the model-free Historical Simulation technique.

Using new desk-level P/Ls and Historical Simulation VaR data from four business lines in a large international commercial bank we discuss the pros and cons of this trend. We find strong evidence of volatility dynamics and non-normality in daily P/Ls. Volatility dynamics are not captured in Historical Simulation and may therefore cause clustering in VaR violations which can have important economic effects such as increased risk of default. We assess the ability of external bank regulators and internal risk auditors to detect problems in Historical Simulation-based VaRs using a wide range of existing and new backtesting procedures.

The relatively sluggish dynamics of Historical Simulation is often touted as a virtue in that it avoids frequent adjustments to the associated risk-based capital. However, the internal desk-level VaRs we present here move quite rapidly at the daily level thus casting doubt on the common wisdom. The extreme spikes in P/Ls particularly evident in lines 1 and 2 highlight the dangers of relying on the normal distribution in market risk management. Fortunately, the Historical Simulation VaRs do not rely on the normal distribution, which is clearly one of its redeeming features.

Our analysis indirectly provides a number of other conclusions. First, larger backtesting sample sizes would clearly be helpful. Second, aggregate GARCH modeling may be desirably to remove clusters in the VaR violations. A viable GARCH simulation approach is suggested by Barone-Adesi, Giannopoulos and Vosper (1999). Jackson, Maude and Perraudin (1997) discuss the pros and cons of parametric versus nonparametric VaR approaches. Third, ideally the historical dataset of asset prices should be updated daily so as to capture current volatility trends. Fourth, simultaneous reporting of the VaR with several coverage rates (e.g. 1%, 2.5%, 5%) would be helpful to assess more carefully the tail distribution of the P/L and to detect omitted variance dynamics.

References

Andersen, T., T. Bollerslev, P. Christoffersen, and F. Diebold, F.X. (2006), "Volatility and Correlation Forecasting," in G. Elliott, C.W.J. Granger, and Allan Timmermann (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland, 778-878.

Artzner, P., F. Delbaen, J. Eber, and D. Heath (1999), "Coherent Measures of Risk," *Mathematical Finance*, 9, 203-228.

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

Basak, S. and A. Shapiro (2001), "Value-at-Risk Based Risk Management: Optimal Policies and Asset Prices," *Review of Financial Studies*, 14, 371-405.

Basle Committee on Banking Supervision (1996), *Amendment to the Capital Accord to Incorporate Market Risks*.

Basle Committee on Banking Supervision (2004), *International Convergence of Capital Measurement and Capital Standards: a Revised Framework.*

Berkowitz, J. (2001), "Generalized Spectral Estimation of the Consumption-Based Asset Pricing Model," *Journal of Econometrics*, 104, 269-288.

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

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics – Methods and Applications.* Cambridge.

Campbell, J.Y., and R.J. Shiller (1987), "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, 95, 1062-1088.

Christoffersen, P.F. (1998), "Evaluating interval forecasts," *International Economic Review* 39, 841-862.

Christoffersen, P.F., and S. Goncalves (2005), "Estimation Risk in Financial Risk Management," *Journal of Risk*, 7, 1-28.

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

Cuoco, D., H. He and S. Issaenko (2001), "Optimal Dynamic Trading Strategies with Risk Limits," Manuscript, The Wharton School, University of Pennsylvania.

Danielsson, J., C. de Vries, B. Jorgensen, S. Mandira, and G. Samorodnitsky, (2005), "Subadditivity Re-Examined: the Case for Value-at-Risk," Manuscript, London School of Economics.

Diebold, F.X., T.A. Gunther, and A.S. Tay (1998), "Evaluating density forecasts," *International Economic Review* 39, 863-883.

Dufour, J.-M., (2006), "Monte Carlo Tests with Nuisance Parameters : A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics," *Journal of Econometrics*, 133, 443-477.

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

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

Engle, R.F. and J. Russel (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127-1162.

Gourieroux, C. (2000). *Econometrics of Qualitative Dependent Variables*. Cambridge University Press.

Haas, M. (2005),"Improved Duration-based Backtesting of Value-at-risk," *Journal of Risk*, 8, 17–38.

Harrison, M. and D. Kreps (1979), "Martingales and Arbitrage in Multi-period Securities Markets," *Journal of Economic Theory*, 20, 381-408.

Harrison, M., and S. Pliska (1981), "Martingales and Stochastic Integrals," in the *Theory of Continuous Trading, Stochastic Processes and their Applications*, 11, 215-260.

Hendricks, D. (1996), "Evaluation of Value-at-Risk models using historical data," *Economic Policy Review*, Federal Reserve Bank of New York, April, 39-69.

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

Jorion, P. (2001). *Value-at-Risk: the New Benchmark for Controlling Market Risk* McGraw-Hill: Chicago.

Jorion, P. (2002), "How informative are Value-at-Risk disclosures," *Accounting Review*, 77, 911-931.

JP Morgan (1994), "RiskMetrics," Technical Document. New York.

Kiefer, N. (1988). "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26, 646-679.

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

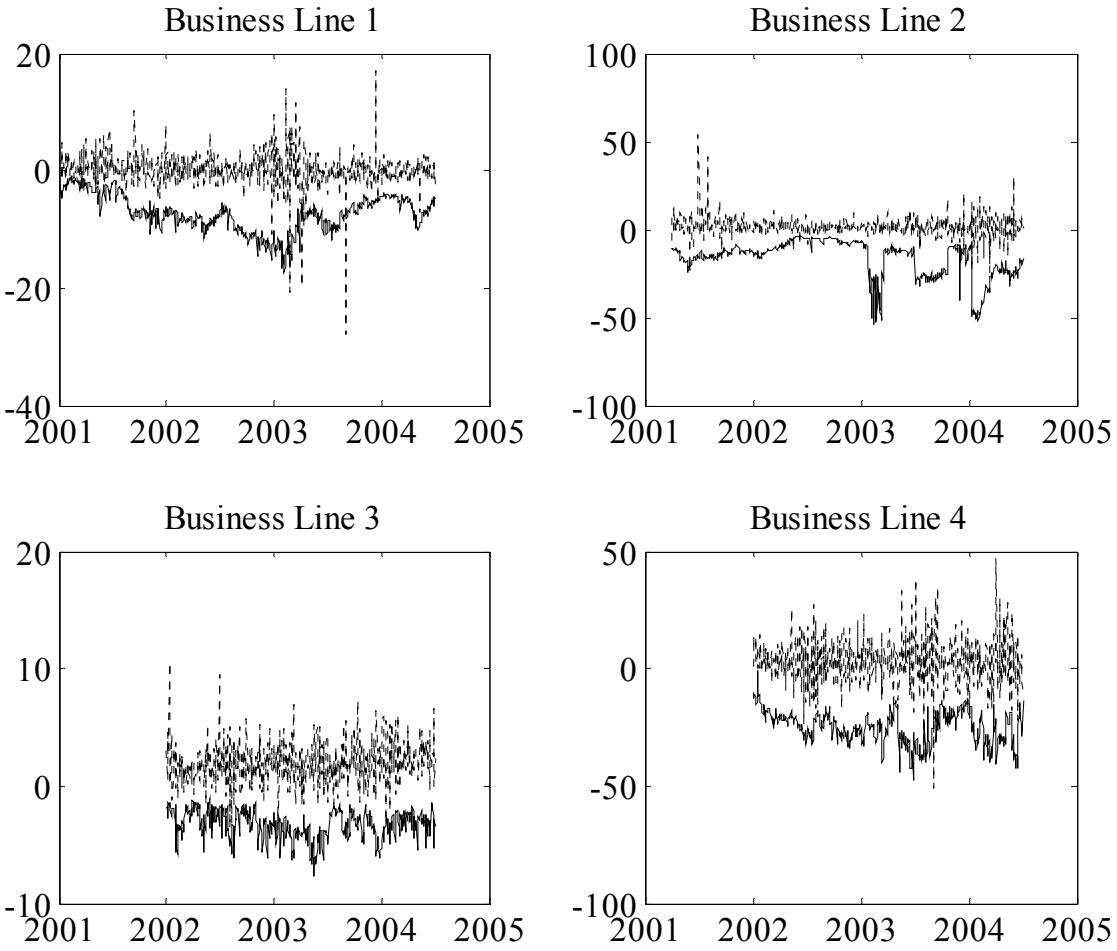
Lütkepohl, H. (1993), *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, second edition.

Paramasamy, S. (1992), "On the Multivariate Kolmogorov-Smirnov Distribution," Statistics and Probability Letters, 15, 149–155.

Perignon, C., Z. Deng and Z. Wang (2006), "Do Banks Overstate their Value-at-Risk?" Manuscript, Simon Fraser University.

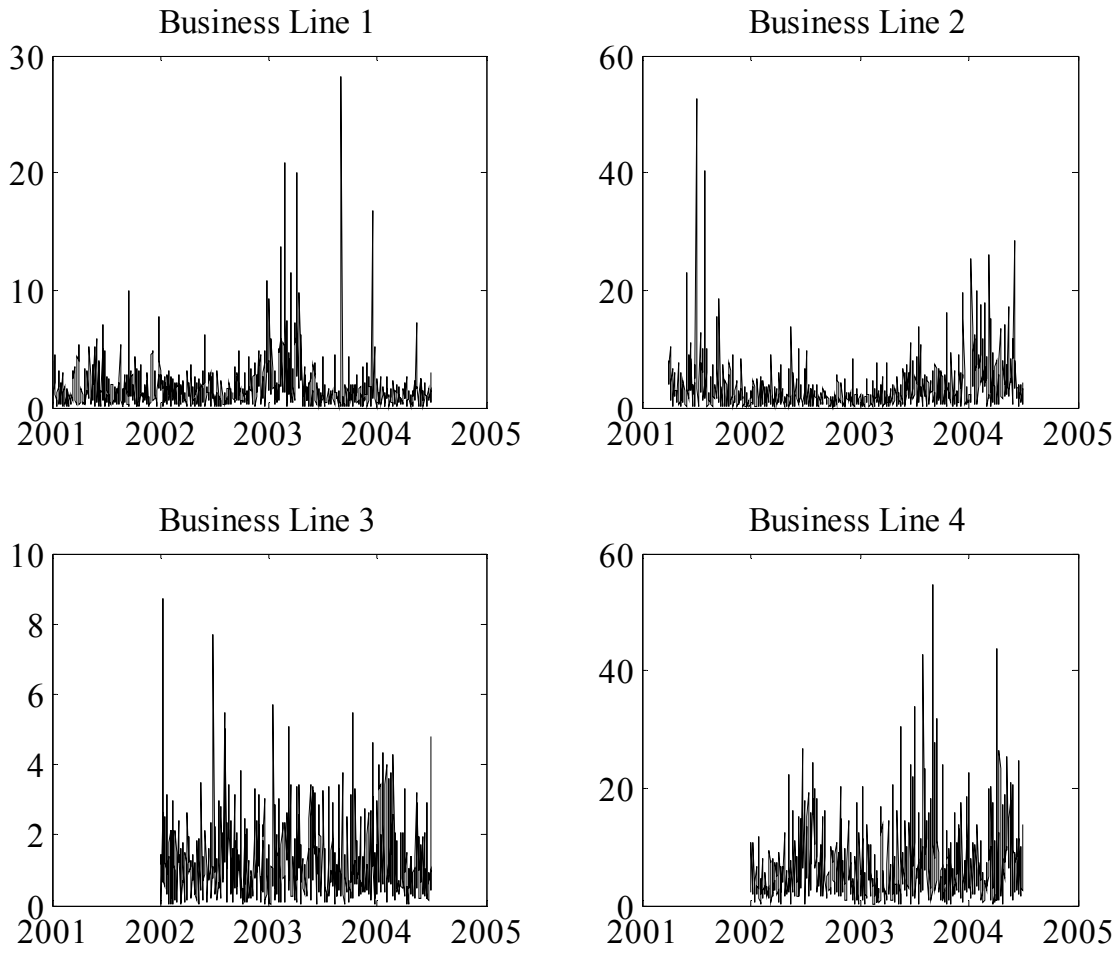
Perignon, C. and D. Smith (2006), "The Level and Quality of Value-at-Risk Disclosure by Commercial Banks," Manuscript, Simon Fraser University.

Figure 1: P/Ls and 1-day, 1% VaRs for Four Business Lines



Notes to Figure: We plot the P/Ls (dashed lines) and 1-day, 1% VaRs (solid lines) from the four business lines.

Figure 2: Absolute Demeaned P/Ls for Four Business Lines



Notes to Figure: We subtract the sample mean from each of the four P/Ls in Figure 1 and plot the absolute value of these centered P/Ls.

Table 1: P/Ls and VaRs for Four Business Lines: Descriptive Statistics

	P/Ls			
	Desk 1	Desk 2	Desk 3	Desk 4
Number of Observations	873	811	623	623
Mean	0.1922	1.5578	1.8740	3.1562
Standard Deviation	2.6777	5.2536	1.6706	9.2443
Skewness	-1.7118	1.5441	0.5091	-0.1456
Excess Kurtosis	24.2195	19.8604	2.0060	3.6882

	VaRs			
	Desk 1	Desk 2	Desk 3	Desk 4
Number of Observations	873	811	623	623
Mean	-7.2822	-16.3449	-3.2922	-24.8487
Standard Deviation	3.1321	10.5446	1.1901	6.6729
Skewness	-0.3038	-1.3746	-0.6529	-0.3006
Kurtosis	-0.1525	1.6714	-0.0133	-0.1211
Observed Number of Hits	9	5	1	4
Expected Number of Hits	9	8	6	6

Notes to table: We report various descriptive statistics for the daily P/Ls and daily 1%, 1-day VaRs for each desk. The number of Hits refers to the number of days on which the ex post loss exceeded the ex ante VaR.

Table 2: Size of 10% Asymptotic CC Tests

1 % VaR

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.0253	0.0999	0.0497	0.1103	0.5306	0.0463	0.0390	0.0517
500	0.0440	0.1336	0.0676	0.1759	0.2332	0.0688	0.0664	0.1120
750	0.0669	0.1650	0.0663	0.1616	0.1582	0.0699	0.0923	0.1241
1000	0.0763	0.1465	0.0759	0.1569	0.1186	0.0673	0.0944	0.1247
1250	0.1022	0.1458	0.0550	0.1276	0.1106	0.0753	0.1120	0.1401
1500	0.1005	0.1309	0.0637	0.1273	0.0954	0.0714	0.1117	0.1372

5 % VaR

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.0805	0.1080	0.1280	0.1336	0.0976	0.0925	0.1020	0.1056
500	0.0675	0.1009	0.1284	0.1252	0.0762	0.1027	0.0907	0.0830
750	0.0685	0.1018	0.1659	0.1400	0.0678	0.1024	0.0965	0.0863
1000	0.0891	0.0965	0.2085	0.1423	0.0718	0.1131	0.0951	0.0950
1250	0.0920	0.0925	0.1607	0.1490	0.0608	0.1208	0.0956	0.0982
1500	0.0866	0.0978	0.1515	0.1596	0.0630	0.1141	0.0949	0.0974

Notes to Table: We simulate i.i.d. Bernoulli variables to assess the size properties of the various asymptotic backtesting procedures. LB(1) and LB(5) are Ljung-Box with 1 and 5 lags. Markov is a first-order Markov test. Weibull and Geometric are duration based tests. Caviar is a regression-based test. KS is Kolmogorov-Smirnov, and CVM is Cramer-von-Mises. Please see the text for details on each test.

Table 3: P/L GARCH Model Parameters and Properties

	Desk 1	Desk 2	Desk 3	Desk 4
d	3.808	3.3183	6.9117	4.7017
θ	-0.245	0.5031	-0.9616	0.0928
β	0.7495	0.9284	0.8728	0.9153
α	0.1552	0.0524	0.0261	0.0723
ω	0.5469	0.2154	0.2127	1.6532
Variance Persistence	0.9140	0.9941	0.9230	0.9882
Unconditional Stdev	2.5220	6.0233	1.6624	11.8478
LogL	-1360.76	-1781.25	-825.87	-1855.98
LogL (HomoSked.)	-1401.64	-1843.49	-831.46	-1877.73
P-value	0.0000	0.0000	0.0108	0.0000

Notes to Table: Using Maximum likelihood we estimate on each desk P/L an asymmetric GARCH(1,1) model with standardized Student's t(d) distributed innovations. The P-value reports the significance level of a test of homoskedastic t(d) returns against the heteroskedastic GARCH-t(d) alternative.

Table 4: Power of 10% Finite Sample CC Tests on 1% VaR in Four Business Lines

Business Line 1

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.1958	0.3201	0.1862	0.1431	0.3258	0.4197	0.3273	0.3305
500	0.2293	0.4200	0.1914	0.1439	0.2642	0.4286	0.4112	0.3563
750	0.2997	0.4755	0.1903	0.1474	0.2758	0.5391	0.4607	0.4075
1000	0.3709	0.5185	0.1678	0.1820	0.3418	0.6176	0.5083	0.4725
1250	0.4335	0.5635	0.1874	0.2280	0.3704	0.6823	0.5421	0.5027
1500	0.4463	0.6027	0.2018	0.2444	0.4098	0.7371	0.5852	0.5638

Business Line 2

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.2312	0.2320	0.2110	0.1373	0.3654	0.4505	0.2777	0.2662
500	0.2201	0.2960	0.1896	0.1557	0.3678	0.4298	0.3146	0.2689
750	0.2372	0.3319	0.1806	0.1795	0.3866	0.4796	0.3407	0.2810
1000	0.2806	0.3614	0.1726	0.2181	0.4214	0.5320	0.3895	0.3229
1250	0.2808	0.4001	0.1604	0.2652	0.4752	0.5804	0.3807	0.3265
1500	0.2795	0.4231	0.1604	0.3041	0.5068	0.6168	0.4263	0.3707

Business Line 3

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.0769	0.1170	0.0725	0.0794	0.1372	0.3329	0.1127	0.1158
500	0.0677	0.1532	0.0625	0.0739	0.0814	0.3291	0.1276	0.1078
750	0.0899	0.1603	0.0529	0.0537	0.0546	0.4103	0.1264	0.1122
1000	0.1059	0.1462	0.0359	0.0507	0.0438	0.5257	0.1374	0.1209
1250	0.1305	0.1273	0.0390	0.0493	0.0470	0.6109	0.1373	0.1295
1500	0.1472	0.1260	0.0311	0.0421	0.0380	0.6860	0.1498	0.1473

Business Line 4

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.2502	0.2637	0.2338	0.1593	0.4058	0.4706	0.3132	0.3016
500	0.2400	0.3367	0.2142	0.1839	0.4144	0.4523	0.3824	0.2984
750	0.2803	0.3824	0.2038	0.2118	0.4288	0.5101	0.4030	0.3216
1000	0.3326	0.4193	0.2019	0.2671	0.5026	0.5738	0.4494	0.3753
1250	0.3170	0.4580	0.1955	0.3432	0.5438	0.6124	0.4546	0.3919
1500	0.3293	0.5095	0.2018	0.3889	0.5968	0.6552	0.4881	0.4272

Notes to Table: We simulate hit sequences from GARCH P/Ls and Historical Simulation VaRs to assess the power properties of the tests. LB(1) and LB(5) are Ljung-Box with 1 and 5 lags. Markov is a first-order Markov test. Weibull and Geometric are duration based tests. Caviar is a regression-based test. KS is Kolmogorov-Smirnov, and CVM is Cramer-von-Mises. Please see the text for details on each test.

Table 5: Power of 10% Finite Sample CC Tests on 5% VaR in Four Business Lines

Business Line 1

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.2964	0.3852	0.2048	0.1613	0.3190	0.4469	0.3485	0.3442
500	0.3912	0.5275	0.2139	0.1825	0.4466	0.5172	0.4429	0.4638
750	0.4356	0.6334	0.2257	0.2305	0.5684	0.6106	0.5316	0.5534
1000	0.4836	0.6957	0.2511	0.2698	0.6794	0.6915	0.5858	0.6068
1250	0.5431	0.7621	0.2935	0.3246	0.7560	0.7609	0.6654	0.6745
1500	0.5925	0.8146	0.3279	0.3790	0.8188	0.8509	0.7199	0.7220

Business Line 2

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.2592	0.3575	0.3404	0.3222	0.4222	0.5829	0.3901	0.3832
500	0.3421	0.5077	0.2978	0.3658	0.5806	0.6167	0.4482	0.4492
750	0.3757	0.5972	0.2719	0.4346	0.6930	0.6616	0.5041	0.5059
1000	0.4194	0.6581	0.2759	0.4869	0.7838	0.7023	0.5584	0.5493
1250	0.4663	0.7208	0.3103	0.5483	0.8420	0.7405	0.6247	0.6090
1500	0.5038	0.7806	0.3345	0.6056	0.8996	0.8187	0.6811	0.6554

Business Line 3

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.1080	0.1129	0.0821	0.0676	0.0888	0.2992	0.0986	0.0991
500	0.1032	0.1199	0.0653	0.0403	0.0644	0.3508	0.1050	0.1122
750	0.1032	0.1246	0.0617	0.0343	0.0490	0.4304	0.1063	0.1160
1000	0.1126	0.1225	0.0624	0.0310	0.0524	0.5114	0.1021	0.1069
1250	0.1140	0.1251	0.0686	0.0285	0.0438	0.5830	0.1131	0.1216
1500	0.1071	0.1247	0.0653	0.0333	0.0534	0.7130	0.1171	0.1163

Business Line 4

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
250	0.2877	0.3926	0.3305	0.3260	0.4460	0.5904	0.4089	0.3926
500	0.3531	0.5360	0.2816	0.3863	0.6262	0.6254	0.4704	0.4743
750	0.3976	0.6367	0.2666	0.4652	0.7462	0.6722	0.5453	0.5397
1000	0.4433	0.7050	0.2779	0.5401	0.8378	0.7290	0.6059	0.5917
1250	0.5010	0.7694	0.3169	0.6128	0.8880	0.7678	0.6670	0.6583
1500	0.5482	0.8239	0.3609	0.6773	0.9356	0.8370	0.7340	0.7130

Notes to Table: We simulate hit sequences from GARCH P/Ls and Historical Simulation VaRs to assess the power properties of the tests. LB(1) and LB(5) are Ljung-Box with 1 and 5 lags. Markov is a first-order Markov test. Weibull and Geometric are duration based tests. Caviar is a regression-based test. KS is Kolmogorov-Smirnov, and CVM is Cramer-von-Mises. Please see the text for details on each test.

Table 6: Fraction of Samples where Test is Feasible. 1% and 5% VaR

Power Simulation: Business Line 1

VaR	Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
1%	250	0.9081	0.9081	0.9006	0.6974	0.8322	0.8998	0.9081	0.9081
1%	500	0.9984	0.9984	0.9974	0.9852	0.9918	0.9974	0.9983	0.9979
1%	750	1.0000	1.0000	1.0000	0.9998	0.9999	1.0000	0.9999	1.0000
5%	250	0.9998	0.9998	0.9998	0.9984	1.0000	0.9996	0.9999	1.0000

Power Simulation: Business Line 2

VaR	Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
1%	250	0.8693	0.8693	0.8643	0.6691	0.8167	0.8634	0.8693	0.8693
1%	500	0.9916	0.9916	0.9928	0.9654	0.9824	0.9925	0.9927	0.9929
1%	750	0.9996	0.9996	0.9999	0.9986	0.9996	0.9997	0.9997	0.9997
5%	250	0.9965	0.9965	0.9949	0.9881	0.9942	0.9958	0.9963	0.9973

Power Simulation: Business Line 3

VaR	Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
1%	250	0.9356	0.9356	0.9371	0.7077	0.8477	0.9362	0.9356	0.9356
1%	500	0.9990	0.9990	0.9998	0.9916	0.9943	0.9994	0.9990	0.9990
1%	750	1.0000	1.0000	1.0000	0.9999	0.9999	1.0000	1.0000	1.0000
5%	250	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Power Simulation: Business Line 4

VaR	Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
1%	250	0.8659	0.8659	0.8660	0.6775	0.8169	0.8645	0.8659	0.8659
1%	500	0.9935	0.9935	0.9940	0.9694	0.9839	0.9944	0.9941	0.9946
1%	750	0.9999	0.9999	0.9999	0.9989	0.9996	1.0000	0.9997	0.9997
5%	250	0.9974	0.9974	0.9971	0.9895	0.9938	0.9972	0.9963	0.9957

Size Simulation

VaR	Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM
1%	250	0.9190	0.9190	0.9190	0.6896	0.7119	0.9189	0.9190	0.9190
1%	500	0.9937	0.9937	0.9937	0.9619	0.9664	0.9938	0.9937	0.9937
1%	750	0.9992	0.9992	0.9992	0.9949	0.9964	0.9994	0.9992	0.9992
5%	250	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Notes to Table: We report the fraction of simulations where the hit sequence allowed us to compute the test statistic. LB(1) and LB(5) are Ljung-Box with 1 and 5 lags. Markov is a first-order Markov test. Weibull and Geometric are duration based tests. Caviar is a regression-based test. KS is Kolmogorov-Smirnov, and CVM is Cramer-von-Mises. Please see the text for details on each test.

Table 7: Backtesting Actual VaRs from Four Business Lines

Business Line 1

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	CavMult
Test Value	0.0955	0.4830	0.1961	1.0138	1.2896	3.2274	18.7481	2.4382	3.7213
P-Value	0.4601	0.5513	0.9628	0.6619	0.3758	0.2779	0.3241	0.3948	0.6141

Business Line 2

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	CavMult
Test Value	0.0315	0.1594	1.4576	3.6338	3.8377	4.8560	11.5003	1.3503	12.4620
P-Value	0.8252	0.8379	0.3199	0.2354	0.1251	0.1308	0.4671	0.5618	**0.0400**

Business Line 3

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	CavMult
Test Value	0.0016	0.0083	6.8487	NaN	NaN	7.5608	70.3651	70.3651	9.7996
P-Value	0.9920	0.9919	**0.0176**			**0.0330**	0.0533	**0.0242**	0.1032

Business Line 4

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	CavMult
Test Value	0.0263	38.5720	0.9752	4.4235	4.9972	4.1037	19.6270	5.2362	9.3517
P-Value	0.7845	**0.0093**	0.3687	0.1720	0.0602	0.1766	0.1815	0.1801	0.1182

Notes to Table: We report the test statistics using the hit sequences from the actual P/Ls and VaRs from the four business lines. LB(1) and LB(5) are Ljung-Box with 1 and 5 lags. Markov is a first-order Markov test. Weibull and Geometric are duration based tests. Caviar is a regression-based test. KS is Kolmogorov-Smirnov, and CVM is Cramer-von-Mises. The CavMult test uses the ex-ante VaR from all four business lines in a Caviar test. Please see the text for details on each test.