Published online ahead of print January 28, 2009

MANAGEMENT SCIENCE
*Articles in Advance*, pp. 1–15
ISSN 0025-1909 | EISSN 1526-5501

informs

DOI 10.1287/mnsc.1080.0964




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

Jeremy Berkowitz
Department of Finance, University of Houston, Houston, Texas 77004, jberkowitz@uh.edu

Peter Christoffersen
McGill University, Montreal, Quebec H3A 2T5, Canada; and CREATES, School of Economics and Management, University of Aarhus, DK-8000 Aarhus C, Denmark, peter.christoffersen@mcgill.ca

Denis Pelletier
Department of Economics, College of Management, North Carolina State University, Raleigh, North Carolina 27695, denis_pelletier@ncsu.edu

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 data set includes the actual 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 data set, 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 conditional autoregressive value-at-risk test of Engle and Manganelli (2004) performs best overall, but duration-based tests also perform well in many cases.

*Key words*: risk management; backtesting; volatility; disclosure
*History*: Received July 30, 2007; accepted October 24, 2008, by John Birge, special issue editor. Published online in *Articles in Advance*.

## 1. Introduction

In the financial services industry, a primary concern of money managers is the ongoing level of risk in their portfolios. For decades, the textbook measure of portfolio risk was the standard deviation or "volatility." However, by the 1990s banks began widespread adoption of value at risk (VaR) as an internal definition of portfolio risk, where the VaR is defined as the lower end of a 99% confidence interval. It is now arguably the single most prevalent financial risk measure used in banking and is becoming increasingly common even in nonfinancial firms (see, for example, Jorion 2006 for an extensive overview of VaR).

The widespread use of VaR as an internal measure of risk was given regulatory recognition under the 1996 Market Risk Amendment to the Basel Accord (Basel Committee on Banking Supervision 1996). Under this system, banks are allowed to have their regulatory required capital based on the bank's own internal VaR forecasts. Although VaR began as a way to measure risk, it is now also used as a management tool. A large bank has a fixed amount of capital that can be allocated by management to traders. To manage overall risk, each trader is typically given a trading limit of some kind. Those trading limits are now typically based on the trader's portfolio VaR. To a certain extent, traders and portfolio managers even use VaR to guide portfolio choice. If a manager observes VaR increasing, it may signal an undesired increase in risk and trigger the closing of a position. For all of these reasons, both financial services firms themselves as well as Federal Reserve and FDIC regulators have an enormous incentive to make sure that bank's VaR forecasts are accurate.

In this paper, we provide an integrated, unifying framework for assessing the accuracy of VaR forecasts. Our approach includes the existing tests proposed by Christoffersen (1998) and Christoffersen and Pelletier (2004) as special cases. In addition, we describe some new tests, which are suggested by our framework. To provide guidance as to which of these many tests have the best finite-sample size and power properties, we conduct a thorough Monte Carlo horserace where the profit and loss (P/L)-generating processes are based on four real P/L series.

We obtained the actual daily profit and loss 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. Although of interest in its own

right, the desk-level data set also provides important guidance for choosing realistic P/L-generating processes in our Monte Carlo comparison of backtesting methods.

In addition to the daily P/L data, we obtained the corresponding daily, one-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 et al. (2007) and Perignon and Smith (2008) 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 10 days ahead, banks typically simply use the square root of 10 to scale the one-day-ahead VaR.

Our umbrella framework for testing the accuracy of a 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 (see Durlauf 1991).

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/nonviolations on their lagged values and other lagged variables such as the previous day's VaR.

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.

Last, 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 function. 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 cannot rely on the asymptotic distribution of the tests to conduct inference. We instead rely on Dufour's (2006) 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 conditional autoregressive value-at-risk (CaViaR) test of Engle and Manganelli (2004) performs best overall, but that duration-based tests also perform well in many cases.

This paper proceeds as follows. In §2, we discuss the use of VaR as a managerial and operational tool within financial services firms. In §3, we present the actual desk-level daily P/Ls and VaRs from several business lines from a large international bank. Section 4 gives an overview of existing methods for backtesting VaR estimates, and it suggests a few new approaches as well. Section 5 presents the results of a detailed horserace among the methods in terms of size and power properties in the finite sample. In §6, we report the results from applying the test to our unique desk-level data sample, and we also assess the ability of VaRs to forecast P/L volatility.

## 2. VaR as a Managerial Tool

The one-day 1% VaR of a given portfolio is a dollar amount, such that daily portfolio loss will be worse than the VaR only 1% of the time. This provides a simple one-dimensional snapshot of the downside risk of the profit and loss distribution. This simplicity is a key reason for its widespread adoption, although it clearly represents a somewhat limited amount of information about the P/L distribution. A key advantage of VaR is that it does not rely on any assumptions of asset return normality.

### 2.1. Risk Controls Using VaR

A typical large commercial or investment bank will have its trading operations organized in a set of trading desks. The organization typically includes a desk for equities, one for currencies, one for fixed-income, and one for derivatives. The risk management team in the bank has to monitor in real time that each trading desk stays within the predefined risk limits imposed by management.

Before the advent of VaR, such risk limits were typically set in the form of notional limits and/or stop-loss limits. Examples of notional limits include a maximum allowed amount invested in a particular currency, in bonds of a certain maturity, or in equities

from a particular industry. Such notional limits are problematic for several reasons, including the fact that they are not easily comparable across asset classes.

The stop-loss limits instead force the desk to unwind positions when the accumulated loss on a position has reached a preset level. Stop-loss limits are comparable across assets, but they suffer from being backward-looking in nature, only measuring risk once the loss is realized.

Using VaR limits as risk controls has the advantage that a forward-looking risk measure is used. The VaR is forward-looking by definition because it reports for example the maximum loss over the next day with 99% probability. Empirical evidence on the forward-looking nature of VaR estimates is provided in Taylor (2005), who shows how VaR estimates can be used to forecast future volatility. Furthermore, VaR limits are comparable across asset classes because the VaR of a position reflects both the notional size of the position as well as the risk per dollar invested. Blanco and Blomstrom (1999) provide a more detailed discussion of the advantages of VaR-based risk limits.

### 2.2. VaR-Based Portfolio Choice

VaR-based risk controls as described above form a passive use of VaR. That is, they do not inform the trading desk how to optimally trade when facing VaR risk limits nor do they tell management how to set optimal VaR-based limits.

In theory, VaR can be used for portfolio choice if it is used as a constraint for the optimal investment policy. For example, optimal portfolio weights can be found by maximizing the expected return or expected utility of terminal wealth subject to a maximum VaR. Basak and Shapiro (2001) have argued on theoretical grounds against the use of VaR as a portfolio optimization constraint; it can encourage excessive risk taking because VaR does not penalize extreme losses. They recommend using an expected shortfall, also known as a CVaR constraint, instead. Alexander and Baptista (2004) also compare the use of VaR and CVaR constraints in portfolio selection and find that CVaR generally dominates VaR except in the absence of a risk-free asset.

However, these critiques of VaR as a portfolio optimization constraint have since been challenged in Cuoco et al. (2008). They show that if the VaR is recomputed dynamically using available information, as is realistic, then the risk exposure of a trader using VaR constraints is always lower than the unconstrained trader.

### 2.3. Regulatory Uses

Under U.S. banking regulations, commercial banks engaged in trading risky financial assets are required to maintain a minimal level of safe assets as a cushion against unforeseen risk. Since the 1996 Market Risk Amendment to the Basel Accord, qualifying banks can opt to set this required capital level as a function of their VaR. Banks are permitted to use their own internal models to calculate their VaR. Backtesting has been given further relevance by its prominence in the discussion of the Supervisory Review Process (the Second Pillar) in Basel II (Basel Committee on Banking Supervision 2004).

Although no particular technique for backtesting is currently suggested in the Basel Accord, Lopez (1999) notes that the required capital for market risk includes a multiplier based on the unconditional number of VaR violations. In this paper, we develop backtesting techniques that assess both the unconditional VaR and bunching in VaR violations. The results of our horserace show the potential for supervisor endorsement of these more advanced backtesting techniques.

## 3. Desk-Level P/L and VaR at a Commercial Bank

We collected the actual daily P/L generated by four separate business lines from a large, international commercial bank. The P/L is based on the change in position values recorded at the close of each day and it does not include brokerage fees or commissions. Each series is constructed and defined in a consistent manner but the series are normalized to protect the bank's anonymity.[1]

For two of the business lines we have over 600 daily observations, whereas for the other two we have over 800 observations, yielding a panel of 2,930 observations. All four business lines are involved in securities trading but the exact nature of each business line is not known to us. We do know that there is very little overlap in assets across business lines. We also know that the different business lines are run by different employees and that all business lines rely on historical-simulation-based VaR systems for risk management. We do not observe the aggregate P&L summed across the business desks.

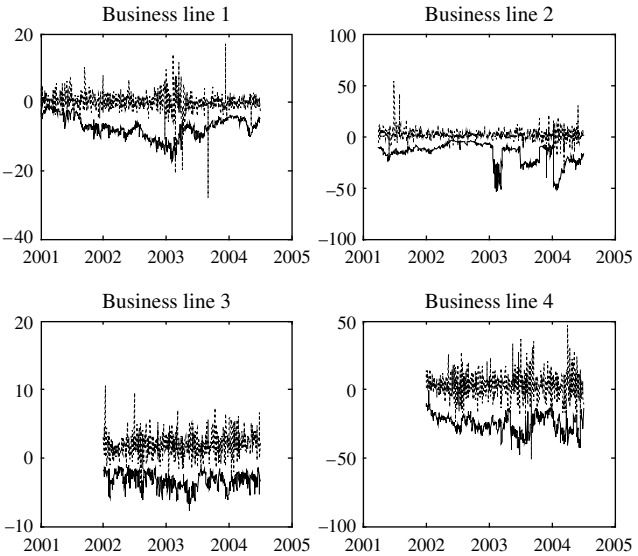
In addition to the daily revenue data, we obtained the corresponding one-day-ahead value-at-risk forecasts. The VaR forecasts are estimates of the 99% lower tail and are calculated for each business line within the bank. The bank relies on historical simulation for computing VaR.

Suppose revenue is denoted by $R_t$. The $p$% VaR is the quantity $\mathrm{VaR}_t$ such that

$$F(R_{t+1} < \mathrm{VaR}_t \mid \Omega_t) = p, \tag{1}$$

[1] The normalization that we employ does not imply that the P/L variance is one. However, the data is normalized by a constant and thus does not affect our results or the analysis in any way.

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




*Note.* We plot the P/Ls (dashed lines) and one-day, 1% VaRs (solid lines) from the four business lines.

where $\Omega_t$ 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 data set, the tail percentile of the bank's VaR is set at $p = 0.01$, 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, to June 30, 2004; business line 2 is observed from April 2, 2001; and business lines 3 and 4 are observed 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., midsample 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 lines 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

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

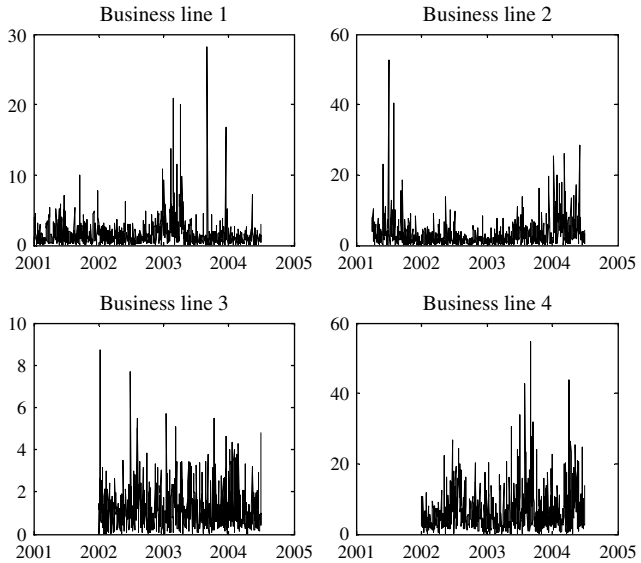
| | Desk 1 | Desk 2 | Desk 3 | Desk 4 |
|---|---|---|---|---|
| | | P/Ls | | |
| Number of observations | 873 | 811 | 623 | 623 |
| Mean | 0.1922 | 1.5578 | 1.8740 | 3.1562 |
| Standard deviation | 2.6777 | 5.2536 | 1.6706 | 9.2443 |
| Skewness | −1.7118 | 1.5441 | 0.5091 | −0.1456 |
| Excess kurtosis | 24.2195 | 19.8604 | 2.0060 | 3.6882 |
| | | VaRs | | |
| 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.* We report various descriptive statistics for the daily P/Ls and daily 1%, one-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.

observations. Of particular interest are the skewness and kurtosis estimates. Skewness is evident in business line 1 (negative) and business 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. Although the spikes in P/Ls dominate the pictures, episodes of high volatility are

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




*Note.* We subtract the sample mean from each of the four P/Ls in Figure 1 and plot the absolute value of these demeaned P/Ls.

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.

# 4. A Unified Framework for VaR Evaluation

Under the 1996 Market Risk Amendment to the Basel Accord (Basel Committee on Banking Supervision 1996) 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 nonregulatory uses of VaR presented in §2 also call for their accurate measurements.

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

$$I_{t+1} = \begin{cases} 1 & \text{if } R_{t+1} < \mathrm{VaR}_t(p), \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

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

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

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. We will refer to tests of this property as conditional coverage (CC) tests.

An unconditional coverage (UC) test of whether $\Pr(I_{t+1} = 1) = p$, under the assumption that the violations are independent, was developed in Kupiec (1995). The UC test rejects the null of an accurate VaR if the actual fraction of VaR violations in a sample is statistically different than $p$. We may expect Kupiec's (1995) test to have lower power than other tests considered in our study because it cannot capture time series dependence in the violations.

## 4.1. Autocorrelation Tests

Christoffersen (1998) notes that property (3) implies that any sequence of violations, $\{I_t\}$, should be an independent and identically distributed (i.i.d.) Bernoulli random variable with mean $p$. To formally test this, Christoffersen (1998) embeds the null hypothesis of an i.i.d. Bernoulli 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

$$\begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}, \tag{4}$$

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

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

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

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

We note that all of the tests we consider are carried out conditional on the value of the first observation. Although the tests all have known 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) and (3) immediately imply that

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

where $\Omega_t$ 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

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

which is familiar as the basis of generalized method of moments estimation.

This motivates a variety of tests that focus on the white noise or martingale property of the sequence. Because 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

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

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

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

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

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

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

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

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

and the alternative hypothesis of interest is that

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

The Portmanteau or Ljung-Box statistics, for example, have a 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 a VaR model by calculating the autocorrelations of $(I_{t+1} - p)$ and then calculating

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

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

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

for any nonanticipating function $g(\cdot)$.

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

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

where we set $g(I_{t-k}, I_{t-k-1}, \ldots, R_{t-k}, R_{t-k-1}, \ldots) = \text{VaR}_{t-k+1}$ 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 whether the $\beta$ 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 (2004).

### 4.2. 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 Bernoulli property, the probability of a violation next period is exactly equal to $\Pr(D_i = 1) = \Pr(I_{t+1} = 1) = p$. The probability of a violation in $d$ periods is

$$\Pr(D_i = d) = \Pr(I_{t+1} = 0, I_{t+2} = 0, \ldots, I_{t+d} = 1). \tag{13}$$

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

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

This allows us to rewrite (13) as

$$\Pr(D_i = d) = (1-p)\cdots(1-p)(p) = (1-p)^{d-1}p. \tag{14}$$

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 without duration dependence 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

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

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 nonconstant probabilities of a violation,

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

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

$$p_d = ad^{b-1}$$

with $b \le 1$ 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 zero, 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 zero, then the last duration, $D_{N(T)}$, is the number of days after the last one 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 probability density function (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

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

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 are

$$\begin{aligned}H_0&: b = 1 \quad \text{and} \quad a = p;\\ H_a&: b \ne 1 \quad \text{or} \quad a \ne p.\end{aligned}$$

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

### 4.3. 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 whether the sample spectrum is "close" to the theoretical flat line.

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

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

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

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

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,

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

which should be approximately a flat line.

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

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

for each frequency $\omega \in [0, \pi]$. The $\widehat{U}(\omega)$ are deviations of the sample spectral distribution from the 45° 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

$$\text{CVM} = \sum_{\omega=0}^{\pi} \widehat{U}(\omega)^2, \tag{21}$$

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

Another common test statistic dates to Bartlett (1955), who showed that the supremum

$$\sup_{\omega} \widehat{U}(\omega)^2 \tag{22}$$

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 that is not white noise.

### 4.4. 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 (separate realizations from the same process). For the Ljung-Box test, we could compute the autocorrelations using all the series, treating them as multiple nonoverlapping sequences from the same underlying process. For likelihood-based tests such as the duration tests in §4.2, we could sum the log-likelihoods for each series. All of the above are based on a likely unrealistic independence assumption.

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, §4.4), or multivariate spectral tests as in Paramasamy (1992). Duration-based tests could be extended by considering competing risk models following Cameron and Trivedi (2005, Chap. 19). Perhaps the easiest way to use information from all of the business lines is offered by the regression approach of the CaViaR test. We can simply use variables from other business lines, such as their P/Ls, as explanatory variables. The conditional coverage test would then consist in testing that the coefficients of the explanatory variables (such as P/Ls) are zero and the probability of getting a violation is equal to $p$. For the Kupiec test, a multivariate version of the unconditional coverage test is developed in Perignon and Smith (2007).

## 5. Size and Power Properties

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

### 5.1. Effective Size of the Tests

To assess the size properties of the various methods, we simulate i.i.d. Bernoulli samples with probabilities $p = 1\%$ and 5%. For each Bernoulli probability, we consider several different sample sizes, from 250 to 1,500. 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 generalized autoregressive conditional heteroskedasticity (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 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) oversized in finite samples. The Markov is somewhat undersized and the Weibull test oversized. The geometric test is extremely oversized for the smallest sample. The CaViaR test is undersized. The CVM test is undersized

**Table 2 Size of 10% Asymptotic CC Tests**

| Sample | LB(1) | LB(5) | Markov | Weibull | Geometric | CaViaR | KS | CVM | Kupiec |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 1% VaR | | | | | |
| 250 | 0.025 | 0.100 | 0.050 | 0.110 | 0.531 | 0.046 | 0.039 | 0.052 | 0.122 |
| 500 | 0.044 | 0.134 | 0.068 | 0.176 | 0.233 | 0.069 | 0.066 | 0.112 | 0.068 |
| 750 | 0.067 | 0.165 | 0.066 | 0.162 | 0.158 | 0.070 | 0.092 | 0.124 | 0.100 |
| 1,000 | 0.076 | 0.147 | 0.076 | 0.157 | 0.119 | 0.067 | 0.094 | 0.125 | 0.117 |
| 1,250 | 0.102 | 0.146 | 0.055 | 0.128 | 0.111 | 0.075 | 0.112 | 0.140 | 0.116 |
| 1,500 | 0.101 | 0.131 | 0.064 | 0.127 | 0.095 | 0.071 | 0.112 | 0.137 | 0.121 |
| | | | | 5% VaR | | | | | |
| 250 | 0.081 | 0.108 | 0.128 | 0.134 | 0.098 | 0.093 | 0.102 | 0.106 | 0.115 |
| 500 | 0.068 | 0.101 | 0.128 | 0.125 | 0.076 | 0.103 | 0.091 | 0.083 | 0.098 |
| 750 | 0.069 | 0.102 | 0.166 | 0.140 | 0.068 | 0.102 | 0.097 | 0.086 | 0.115 |
| 1,000 | 0.089 | 0.097 | 0.209 | 0.142 | 0.072 | 0.113 | 0.095 | 0.095 | 0.112 |
| 1,250 | 0.092 | 0.093 | 0.161 | 0.149 | 0.061 | 0.121 | 0.096 | 0.098 | 0.104 |
| 1,500 | 0.087 | 0.098 | 0.152 | 0.160 | 0.063 | 0.114 | 0.095 | 0.097 | 0.095 |

*Notes.* 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 tests with 1 and 5 lags, respectively. Markov is a first-order Markov test. Weibull and geometric are duration-based tests. CaViaR is a regression-based test. Kupiec is a test of unconditional coverage. Please see the text for details on each test.

for the smallest sample size and oversized for the larger samples. Finally, the Kupiec (1995) unconditional test and the KS test have good size properties beyond the smallest sample sizes.

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 test is somewhat undersized, whereas the Kupiec, 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 §6.

### 5.2. Finite-Sample Power of the Tests

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 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,

$$R_{t+1} = \sigma_{t+1}((d-2)/d)^{1/2} z_{t+1},$$

$$\sigma_{t+1}^2 = \omega + \alpha\sigma_t^2(((d-2)/d)z_t - \theta)^2 + \beta\sigma_t^2,$$

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's $t$ 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 $\alpha$ and a $\beta$ much closer to one. Variance persistence in this model is given by $\alpha(1+\theta^2)+\beta$. It is largest in business lines 2 and 4, which confirms 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 $\alpha$ and small $d$, which suggests large kurtosis. Desk 2 is characterized by high variance persistence and high unconditional kurtosis from the low $d$. Desk 3 has an unusually large negative $\theta$, which suggests that a positive P/L

**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 |
| $\theta$ | −0.245 | 0.5031 | −0.9616 | 0.0928 |
| $\beta$ | 0.7495 | 0.9284 | 0.8728 | 0.9153 |
| $\alpha$ | 0.1552 | 0.0524 | 0.0261 | 0.0723 |
| $\omega$ | 0.5469 | 0.2154 | 0.2127 | 1.6532 |
| Variance persistence | 0.9140 | 0.9941 | 0.9230 | 0.9882 |
| Unconditional std. dev. | 2.5220 | 6.0233 | 1.6624 | 11.8478 |
| logL | −1,360.76 | −1,781.25 | −825.87 | −1,855.98 |
| logL (homosked.) | −1,401.64 | −1,843.49 | −831.46 | −1,877.73 |
| $p$-value | 0.0000 | 0.0000 | 0.0108 | 0.0000 |

*Notes.* Using the 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**

| Sample | LB(1) | LB(5) | Markov | Weibull | Geometric | CaViaR | KS | CVM | Kupiec |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Business line 1 | | | | | |
| 250 | 0.196 | 0.320 | 0.186 | 0.143 | 0.326 | 0.420 | 0.327 | 0.331 | 0.129 |
| 500 | 0.229 | 0.420 | 0.191 | 0.144 | 0.264 | 0.429 | 0.411 | 0.356 | 0.061 |
| 750 | 0.300 | 0.476 | 0.190 | 0.147 | 0.276 | 0.539 | 0.461 | 0.408 | 0.036 |
| 1,000 | 0.371 | 0.519 | 0.168 | 0.182 | 0.342 | 0.618 | 0.508 | 0.473 | 0.025 |
| 1,250 | 0.434 | 0.564 | 0.187 | 0.228 | 0.370 | 0.682 | 0.542 | 0.503 | 0.025 |
| 1,500 | 0.446 | 0.603 | 0.202 | 0.244 | 0.410 | 0.737 | 0.585 | 0.564 | 0.023 |
| | | | | Business line 2 | | | | | |
| 250 | 0.231 | 0.232 | 0.211 | 0.137 | 0.365 | 0.451 | 0.278 | 0.266 | 0.207 |
| 500 | 0.220 | 0.296 | 0.190 | 0.156 | 0.368 | 0.430 | 0.315 | 0.269 | 0.144 |
| 750 | 0.237 | 0.332 | 0.181 | 0.180 | 0.387 | 0.480 | 0.341 | 0.281 | 0.097 |
| 1,000 | 0.281 | 0.361 | 0.173 | 0.218 | 0.421 | 0.532 | 0.390 | 0.323 | 0.070 |
| 1,250 | 0.281 | 0.400 | 0.160 | 0.265 | 0.475 | 0.580 | 0.381 | 0.327 | 0.090 |
| 1,500 | 0.280 | 0.423 | 0.160 | 0.304 | 0.507 | 0.617 | 0.426 | 0.371 | 0.085 |
| | | | | Business line 3 | | | | | |
| 250 | 0.077 | 0.117 | 0.073 | 0.079 | 0.137 | 0.333 | 0.113 | 0.116 | 0.069 |
| 500 | 0.068 | 0.153 | 0.063 | 0.074 | 0.081 | 0.329 | 0.128 | 0.108 | 0.024 |
| 750 | 0.090 | 0.160 | 0.053 | 0.054 | 0.055 | 0.410 | 0.126 | 0.112 | 0.015 |
| 1,000 | 0.106 | 0.146 | 0.036 | 0.051 | 0.044 | 0.526 | 0.137 | 0.121 | 0.011 |
| 1,250 | 0.131 | 0.127 | 0.039 | 0.049 | 0.047 | 0.611 | 0.137 | 0.130 | 0.009 |
| 1,500 | 0.147 | 0.126 | 0.031 | 0.042 | 0.038 | 0.686 | 0.150 | 0.147 | 0.005 |
| | | | | Business line 4 | | | | | |
| 250 | 0.250 | 0.264 | 0.234 | 0.159 | 0.406 | 0.471 | 0.313 | 0.302 | 0.213 |
| 500 | 0.240 | 0.337 | 0.214 | 0.184 | 0.414 | 0.452 | 0.382 | 0.298 | 0.160 |
| 750 | 0.280 | 0.382 | 0.204 | 0.212 | 0.429 | 0.510 | 0.403 | 0.322 | 0.114 |
| 1,000 | 0.333 | 0.419 | 0.202 | 0.267 | 0.503 | 0.574 | 0.449 | 0.375 | 0.085 |
| 1,250 | 0.317 | 0.458 | 0.196 | 0.343 | 0.544 | 0.612 | 0.455 | 0.392 | 0.112 |
| 1,500 | 0.329 | 0.510 | 0.202 | 0.389 | 0.597 | 0.655 | 0.488 | 0.427 | 0.114 |

*Notes.* 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 tests with 1 and 5 lags, respectively. Markov is a first-order Markov test. Weibull and geometric are duration-based tests. CaViaR is a regression-based test. Kupiec is a test of unconditional coverage. Please see the text for details on each test.

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. Overall, our GARCH estimates are similar to ones obtained by Perignon and Smith (2007) with aggregate bank data, but our estimates of the Student's $t(d)$ degree of freedom are in the lower range of the usual values obtained with various financial returns.

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 2008) 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,

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

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 §3) or 5%. We look at the one-day-ahead VaR again as in §3. When computing the finite-sample $p$-values we use 9,999 simulations, and we perform 10,000 Monte Carlo simulations for each test. Section 6 provides the details of the $p$-value simulation.

Table 4 shows the finite-sample power results, based on a 10% significance level, for the 1% VaR from historical simulation for various samples sizes when using the GARCH data-generating processes corresponding to each of the four business lines.

For all of the sample sizes in all 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 but the geometric test is best for the larger sample sizes we examine. For business line 3, the power

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

| Sample | LB(1) | LB(5) | Markov | Weibull | Geometric | CaViaR | KS | CVM | Kupiec |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Business line 1 | | | | | |
| 250 | 0.296 | 0.385 | 0.205 | 0.161 | 0.319 | 0.447 | 0.349 | 0.344 | 0.157 |
| 500 | 0.391 | 0.528 | 0.214 | 0.183 | 0.447 | 0.517 | 0.443 | 0.464 | 0.069 |
| 750 | 0.436 | 0.633 | 0.226 | 0.231 | 0.568 | 0.611 | 0.532 | 0.553 | 0.033 |
| 1,000 | 0.484 | 0.696 | 0.251 | 0.270 | 0.679 | 0.692 | 0.586 | 0.607 | 0.019 |
| 1,250 | 0.543 | 0.762 | 0.294 | 0.325 | 0.756 | 0.761 | 0.665 | 0.675 | 0.013 |
| 1,500 | 0.593 | 0.815 | 0.328 | 0.379 | 0.819 | 0.851 | 0.720 | 0.722 | 0.010 |
| | | | | Business line 2 | | | | | |
| 250 | 0.259 | 0.358 | 0.340 | 0.322 | 0.422 | 0.583 | 0.390 | 0.383 | 0.371 |
| 500 | 0.342 | 0.508 | 0.298 | 0.366 | 0.581 | 0.617 | 0.448 | 0.449 | 0.257 |
| 750 | 0.376 | 0.597 | 0.272 | 0.435 | 0.693 | 0.662 | 0.504 | 0.506 | 0.201 |
| 1,000 | 0.419 | 0.658 | 0.276 | 0.487 | 0.784 | 0.702 | 0.558 | 0.549 | 0.164 |
| 1,250 | 0.466 | 0.721 | 0.310 | 0.548 | 0.842 | 0.741 | 0.625 | 0.609 | 0.140 |
| 1,500 | 0.504 | 0.781 | 0.335 | 0.606 | 0.900 | 0.819 | 0.681 | 0.655 | 0.140 |
| | | | | Business line 3 | | | | | |
| 250 | 0.108 | 0.113 | 0.082 | 0.068 | 0.089 | 0.299 | 0.099 | 0.099 | 0.057 |
| 500 | 0.103 | 0.120 | 0.065 | 0.040 | 0.064 | 0.351 | 0.105 | 0.112 | 0.014 |
| 750 | 0.103 | 0.125 | 0.062 | 0.034 | 0.049 | 0.430 | 0.106 | 0.116 | 0.005 |
| 1,000 | 0.113 | 0.123 | 0.062 | 0.031 | 0.052 | 0.511 | 0.102 | 0.107 | 0.002 |
| 1,250 | 0.114 | 0.125 | 0.069 | 0.029 | 0.044 | 0.583 | 0.113 | 0.122 | 0.001 |
| 1,500 | 0.107 | 0.125 | 0.065 | 0.033 | 0.053 | 0.713 | 0.117 | 0.116 | 0.001 |
| | | | | Business line 4 | | | | | |
| 250 | 0.288 | 0.393 | 0.331 | 0.326 | 0.446 | 0.590 | 0.409 | 0.393 | 0.353 |
| 500 | 0.353 | 0.536 | 0.282 | 0.386 | 0.626 | 0.625 | 0.470 | 0.474 | 0.235 |
| 750 | 0.398 | 0.637 | 0.267 | 0.465 | 0.746 | 0.672 | 0.545 | 0.540 | 0.171 |
| 1,000 | 0.443 | 0.705 | 0.278 | 0.540 | 0.838 | 0.729 | 0.606 | 0.592 | 0.150 |
| 1,250 | 0.501 | 0.769 | 0.317 | 0.613 | 0.888 | 0.768 | 0.667 | 0.658 | 0.125 |
| 1,500 | 0.548 | 0.824 | 0.361 | 0.677 | 0.936 | 0.837 | 0.734 | 0.713 | 0.132 |

*Notes.* 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 tests with 1 and 5 lags, respectively. Markov is a first-order Markov test. Weibull and geometric are duration-based tests. CaViaR is a regression-based test. Kupiec is a test of unconditional coverage. Please see the text for details on each test.

is again low everywhere except for the CaViaR test. For business line 4, the CaViaR is again best for small samples and the geometric test is best for large samples.

Considering Tables 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.

The Kupiec test does not perform well under our simulation setup. This result is expected considering the historical simulation model used to compute the VaR. Historical simulation is by design tracking an unconditional quantile using a nonparametric approach. The major source of misspecification in this case is in the dynamics of the VaR because it is only coming from the rolling window and the Kupiec test cannot detect this.

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 because the dependence in the hit sequence is not of a first-order here. Second, the geometric test is 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 conditional tests is quite impressive. The CaViaR tests display strong power to reject inaccurate VaR, especially compared to the unconditional test. This is important because regulatory capital includes a penalty if the unconditional number of exceptions is too high, so the charge for an inaccurate VaR is implicitly dependent on an unconditional test. No formal backtesting method is currently recommended under the Basel Accord, but the evidence presented here strongly suggests a method along the lines of the CaViaR method rather than a method based on the unconditional violation rate.

### 5.3. Feasibility Ratios

For transparency we report in Table 6 the fraction of simulated samples from Tables 4 and 5 where each test is feasible. We only report sample sizes 250, 500, and 750 for the 1% VaR and 250 for the 5% VaR

**Table 6 Fraction of Samples Where Test Is Feasible (1% and 5% VaR)**

| VaR (%) | Sample | LB(1) | LB(5) | Markov | Weibull | Geometric | CaViaR | KS | CVM |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Power simulation: Business line 1 | | | | | |
| 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 | | | | | |
| 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 | | | | | |
| 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 | | | | | |
| 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 | | | | | |
| 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.* 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 tests with 1 and 5 lags, respectively. Markov is a first-order Markov test. Weibull and geometric are duration-based tests. CaViaR is a regression-based test. Please see the text for details on each test.

because 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 nontrivial. 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. We do not report results for the Kupiec test because it can always be computed.

## 6. 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 as follows: We first generate $N$ independent realizations of the test statistic, $LR_i$, $i = 1, \ldots, N$. 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 an 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 $LR_i$, $i = 0, 1, \ldots, N$ in nondecreasing order and obtain the Monte Carlo $p$-value $\hat{p}_N(LR_0)$, where

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

and

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

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

When working with binary sequences, there is a nonzero 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

**Table 7 Backtesting Actual VaRs from Four Business Lines**

| | LB(1) | LB(5) | Markov | Weibull | Geometric | CaViaR | VIX | KS | CVM | CavMult | Kupiec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Business line 1 | | | | | | |
| Test value | 0.096 | 0.483 | 0.196 | 1.014 | 1.290 | 3.227 | 0.521 | 18.748 | 2.438 | 3.721 | 0.008 |
| *p*-value | 0.460 | 0.551 | 0.963 | 0.662 | 0.376 | 0.278 | 0.832 | 0.324 | 0.395 | 0.614 | 0.911 |
| | | | | | Business line 2 | | | | | | |
| Test value | 0.032 | 0.159 | 1.458 | 3.634 | 3.838 | 4.856 | 2.187 | 11.500 | 1.350 | 12.462 | 1.395 |
| *p*-value | 0.825 | 0.838 | 0.320 | 0.235 | 0.125 | 0.131 | 0.411 | 0.467 | 0.562 | **0.040** | 0.259 |
| | | | | | Business line 3 | | | | | | |
| Test value | 0.002 | 0.008 | 6.849 | NaN | NaN | 7.561 | 27.303 | 70.365 | 70.365 | 9.800 | 6.846 |
| *p*-value | 0.992 | 0.992 | **0.018** | | | **0.033** | **0.014** | **0.053** | **0.024** | 0.103 | **0.009** |
| | | | | | Business line 4 | | | | | | |
| Test value | 0.026 | 38.572 | 0.975 | 4.424 | 4.997 | 4.104 | 3.430 | 19.627 | 5.236 | 9.352 | 0.923 |
| *p*-value | 0.785 | **0.009** | 0.369 | 0.172 | **0.060** | 0.177 | 0.284 | 0.182 | 0.180 | 0.118 | 0.330 |

*Notes.* 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 tests with 1 and 5 lags, respectively. Markov is a first-order Markov test. Weibull and geometric are duration-based tests. CaViaR is a regression-based test. Kupiec is a test of unconditional coverage. The CavMult test uses the ex ante VaR from all four business lines in a CaViaR test. The VIX column is a CaViaR test where we use the value of the Chicago Board Options Exchange Volatility Index (VIX) as an explanatory variable instead of the VaR. NaN is to denote that the test cannot be computed. The bold numbers are *p*-values less than 10%. Please see the text for details on each test.

obtain the Monte Carlo $p$-value by replacing $\widehat{G}_N(LR_0)$ with

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

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 because 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, however, 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 tests 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 of 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. Six tests, including the Kupiec test, reject the VaR in business line 3 using a 10% significance level. 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 actual VaRs, we reject their statistical accuracy for three out of four business lines. For business line 1, the VaR based on a 250-day historical simulation approach appears to work well. The same cannot be said for the other three business lines. Our backtests indicate that the VaR models for business lines 2–4 are stastistically inaccurate and may need modification.

Our data set also allow us to explore how well this bank is able to forecast their portfolio risk at the business-line level. In the spirit of the Mincer and Zarnowitz (1969) method used in the forecasting literature, we can regress the absolute value of the P/L on the corresponding VaR. From Taylor (2005), we know that the VaR should be proportional to the standard deviation, so the $R^2$ from this regression is an indication of how much of the variability in the P/L could be forecasted by the computation of the VaR. In Table 8, we present three sets of $R^2$ values for each business line. The first $R^2$ value is the one obtained when regressing the business line's absolute P/L on its VaR (computed with historical simulation) plus an intercept. To help us assess how big the first $R^2$ is, we simulate absolute P/L data from the GARCH models used in §5, and we report the $R^2$ for the following two regression of absolute P/Ls on the 1% one-day-ahead VaR obtained with either (i) historical simulation with a 250-day rolling window or (ii) the true VaR. The first number is the $R^2$ we would expect to obtain with the true data, and the second is an indication of the upper bound we could obtain.

Our result suggest that at the business-line level the bank forecasts risk as well as, if not better than, we would expect given the historical simulation method used to compute VaR. For three of out four business lines, the $R^2$ obtained with the real data is significantly higher than the one obtained with simulated

**Table 8 Forecasting Portfolio Risk**

| | $R^2$ |
|---|---|
| Business line 1 | |
| Real data | 0.0360 |
| Simulated data and HS VaR | 0.0015 |
| Simulated data and true VaR | 0.0812 |
| Business line 2 | |
| Real data | 0.0694 |
| Simulated data and HS VaR | 0.0438 |
| Simulated data and true VaR | 0.1415 |
| Business line 3 | |
| Real data | 0.0148 |
| Simulated data and HS VaR | 0.0009 |
| Simulated data and true VaR | 0.0080 |
| Business line 4 | |
| Real data | 0.0191 |
| Simulated data and HS VaR | 0.0185 |
| Simulated data and true VaR | 0.1211 |

*Notes.* For each business line, we first regress the absolute observed P/L on the observed VaR ("real data"). In the second regression, we fit simulated GARCH P/L data on a 250-day historical simulation (HS) VaR. In the third regression, we fit the same simulated P/L data on the true simulated GARCH VaR. All regressions include a constant term.

data and historical simulation. But these $R^2$ values are much smaller than the ones where we use the true GARCH-based VaR in the regression (except for business line 3 where GARCH may fit poorly), indicating that the bank's risk management system could quite likely be improved by incorporating dynamic volatility into the VaR computations.

## 7. Conclusions

The uses of VaR in banking are many and varied. All VaR applications share, however, the need for constant evaluation of the accuracy of the VaR risk measures reported. This is true regardless of whether the VaR is used in a passive or active way, and whether it is used in internal operations or externally for regulatory purposes.

The widespread and sudden losses experienced by financial services firms during the 1998 "currency crisis," the 2000–2001 internet bubble, and the current collapse of collateralized debt securities, all serve to highlight the importance of making sure that risk measures are accurately calculated. Although having accurate VaR measures may not prevent volatility, accurate VaRs can be used to calculate risk levels and the appropriate amount of safe capital. Similarly, VaR measures cannot prevent traders from experiencing losses, but they can provide management with a sense of how risky their traders are behaving, and VaR-based trading limits can be instituted to control overall risk. Using new desk-level P/Ls from four business lines in a large international commercial bank, we find evidence of volatility dynamics and nonnormality in the desk-level data. Volatility dynamics are not captured in historical simulation and may therefore cause clustering in VaR violations.

Formal backtesting techniques show the clustering is severe enough that we can reject the accuracy of the VaR models for two of the four business lines. A third business line VaR is rejected by the Kupiec test of unconditional coverage. This suggests that the set of VaR problems discussed here can successfully be detected by external bank regulators and internal risk auditors in real-world situations. Because no formal backtesting method is currently recommended under the Basel Accord, the evidence presented here strongly suggests a possible direction for improvements to future regulatory schemes. Regulators may benefit from adopting an approach along the lines of the CaViaR method rather than a method based on the unconditional violation rate.

### Acknowledgments

The second author acknowledges financial support from CREATES, FQRSC, IFM2, and SSRHC, and the third author acknowledges financial support from the North Carolina State University Enterprise Risk Management Initiative and the Edwin Gill Research Grant.

### References

Alexander, G. J., A. M. Baptista. 2004. A comparison of VaR and CVaR constraints on portfolio selection with the mean-variance model. *Management Sci.* **50**(9) 1261–1273.

Andersen, T., T. Bollerslev, P. Christoffersen, F. X. Diebold. 2006. Volatility and correlation forecasting. G. Elliott, C. W. J. Granger, A. Timmermann, eds. *Handbook of Economic Forecasting*. North-Holland, Amsterdam, 778–878.

Bartlett, M. S. 1955. *An Introduction to Stochastic Processes*. Cambridge University Press, Cambridge, UK.

Basak, S., A. Shapiro. 2001. Value-at-risk based risk management: Optimal policies and asset prices. *Rev. Financial Stud.* **14**(2) 371–405.

Basel Committee on Banking Supervision. 1996. *Amendment to the Capital Accord to Incorporate Market Risks*. Bank for International Settlements, Basel, Switzerland.

Basel Committee on Banking Supervision. 2004. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Bank for International Settlements, Basel, Switzerland.

Berkowitz, J., J. O'Brien. 2002. How accurate are the value-at-risk models at commercial banks? *J. Finance* **57**(3) 1093–1111.

Blanco, C., S. Blomstrom. 1999. VaR applications: Setting VaR-based limits. Working paper, Financial Engineering Associates, Inc., Berkeley, CA.

Cameron, A. C., P. K. Trivedi. 2005. *Microeconometrics—Methods and Applications*. Cambridge University Press, New York.

Campbell, J. Y., R. J. Shiller. 1987. Cointegration and tests of present value models. *J. Political Econom.* **95**(5) 1062–1088.

Christoffersen, P. F. 1998. Evaluating interval forecasts. *Internat. Econom. Rev.* **39**(4) 841–862.

Christoffersen, P. F., D. Pelletier. 2004. Backtesting value-at-risk: A duration-based approach. *J. Financial Econometrics* **2**(1) 84–108.

Cuoco, D., H. He, S. Isaenko. 2008. Optimal dynamic trading strategies with risk limits. *Oper. Res.* **56**(2) 358–368.

Dufour, J.-M. 2006. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *J. Econometrics* **133**(2) 443–477.

Durlauf, S. N. 1991. Spectral based testing of the martingale hypothesis. *J. Econometrics* **50** 355–376.

Engle, R. F., S. Manganelli. 2004. CAViaR: Conditional autoregressive value-at-risk by regression quantiles. *J. Bus. Econom. Statist.* **22**(4) 367–381.

Engle, R. F., J. Russel. 1998. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**(5) 1127–1162.

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

Haas, M. 2005. Improved duration-based backtesting of value-at-risk. *J. Risk* **8**(2) 17–38.

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

Kiefer, N. 1988. Economic duration data and hazard functions. *J. Econom. Literature* **26**(2) 646–679.

Kupiec, P. 1995. Techniques for verifying the accuracy of risk measurement models. *J. Derivatives* **3** 73–84.

Lopez, J. A. 1999. Regulatory evaluation of value-at-risk models. *J. Risk* **1**(2) 37–64.

Lütkepohl, H. 1993. *Introduction to Multiple Time Series Analysis*, 2nd ed. Springer-Verlag, Berlin.

Mincer, J., V. Zarnowitz. 1969. The evaluation of economic forecasts. J. Mincer, ed. *Economic Forecasts and Expectations*. National Bureau of Economic Research, Cambridge, MA, 3–46.

Paramasamy, S. 1992. On the multivariate Kolmogorov-Smirnov distribution. *Statist. Probab. Lett.* **15**(2) 149–155.

Perignon, C., D. R. Smith. 2007. Which value-at-risk method works best for bank trading revenues? Working paper, Simon Fraser University, Burnaby, British Columbia, Canada.

Perignon, C., D. Smith. 2008. The level and quality of value-at-risk disclosure by commercial banks. Working paper, Simon Fraser University, Burnaby, British Columbia, Canada.

Perignon, C., Z. Deng, Z. Wang. 2007. Do banks overstate their value-at-risk? Working paper, Simon Fraser University, Burnaby, British Columbia, Canada.

Taylor, J. W. 2005. Generating volatility forecasts from value at risk estimates. *Management Sci.* **51**(5) 712–725.