CREATES Research Paper 2009-35

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

Peter Christoffersen, Jeremy Berkowitz and Denis Pelletier

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

Jeremy Berkowitz – University of Houston
jberkowitz@uh.edu

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

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

October 30, 2008

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 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 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.

1. Introduction

In the financial services industry, a primary concern of money managers is the on-going level of risk in their portfolios. For decades the textbook measure of portfolio risk was the standard deviation or "volatility". However, by the 1990's 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 percent 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. Under this system, banks are allowed to have their regulatory required capital based on the bank's own internal VaR forecasts.

While VaR began as way to measure risk, it is now also used as a management tool. A large bank has a fixed amount of capital which can be allocated by management to traders. In order 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 these reasons, both financial services firms themselves as well as Federal Reserve and FDIC Regulators have an enormous incentive to make sure 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.

In order 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. While of interest in its own right, the desk-

1

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, 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 simply use the square root of ten to scale the one-day ahead VaR.

Our umbrella framework for 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 (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/non-violations 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.

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 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 (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.

The paper proceeds as follows. In Section 2 we discuss the use of VaR as a managerial and operational tool within financial services firms. In Section 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 finite sample. In Section 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.

A. 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, a fixed-income desk, and a derivatives desk. 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 as 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 as 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 as 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.

B. VaR-based Portfolio Choice

VaR-based risk controls as described above form a passive use of VaR. That is, it does not inform the trading desk how to optimally trade when facing VaR risk limits nor does it 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 portfolio optimization constraint because it can encourage excessive risk-taking as 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, He and Issaenko (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.

C. 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).

While 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 technique.

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

We collected the actual daily profit and loss (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 while 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 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

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

where Ω_t is the risk manager's time-t information set. The VaR is the pth 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

---

[1] The normalization that we employ does not imply 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.

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 are 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 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 non-regulatory uses of VaR presented in Section 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

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

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

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

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. 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 test to have lower power than other tests considered in our study since it cannot capture time series dependence in the violations.

A. Autocorrelation Tests

Christoffersen (1998) notes that property (3) implies that any sequence of violations, $\{I_t\}$, should be an i.i.d. Bernoulli random variable with mean p. In order 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 $pr(I_{t+1} \,|\, I_t)$ are given by

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

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

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, *i.e.* 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

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

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,

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

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

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

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

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

In analogy with Engle and Manganelli (2004), we might consider the nth-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 β 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

(12) $D_i = t_i - t_{i-1}$

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

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

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

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

This allows us to rewrite (13) as

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

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

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

(15) $\lambda(D_i) = \dfrac{Pr(D_i = d)}{1 - Pr(D_i < d)}$

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

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

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

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

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,

$$Pr\left(D_i = d\right) = Pr\left(I_{t+1} = 0, I_{t+2} = 0, ..., I_{t+d} = 1\right)$$
$$= \left(1 - p_1\right)\left(1 - p_2\right)\cdots\left(1 - p_{d-1}\right)p_d$$

where

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

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

$$p_d = a d^{b-1}$$

with $b \leq 1$ 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 (separate realizations from the same process). 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. All 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, 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 from 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. For the Kupiec test, a multivariate version of the unconditional coverage test is developed in Perignon and Smith (2007).

15

5. Size and Power Properties

Given the large variety of backtesting procedures surveyed in Section 3, it is important to give risk managers 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) 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 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 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 Section 6.

<u>B. Finite Sample Power of the Tests</u>

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,

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

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-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 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 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. Overall, our GARCH estimates are similar to ones obtained by Perignon and Smith (2008) 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 (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

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

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 Section 3) or 5%. We look at one-day ahead VaR again as in Section 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 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 but the Geometric test is best for the larger sample sizes we examine. For business line 3 the power is again low everywhere except for the CaViaR test. For business line 4 the CaViaR is again best for small samples and the Geometric 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 non-parametric 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 as the dependence in the hit sequence is not of order 1 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 long the lines of the CaViaR method rather than a method based on the unconditional violation rate.

C. 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 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. We do not report results for the Kupiec test since 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 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 $\hat{p}_N(LR_0)$, where

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

and

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

The indicator function 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 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, $LR_i, i = 0,1,...,N$, we draw an independent realization of a uniform distribution on the $[0;1]$ interval. Denote these draws by $U_i, i = 0,1,...,N$. We obtain the Monte Carlo p-value by replacing $\hat{\mathrm{G}}_N(\mathrm{LR}_0)$ with

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

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 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. 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 3 out of 4 business lines. For business line 1, the VaR based on a 250-day Historical simulation approach appears to works well. The same cannot be said for the other three business lines. Our backtests indicate that the VaR models for lines 2 through 4 are statistically inaccurate and may need modification.

Our dataset 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 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 Section 5 and we report the R^2 for the following two regression of absolute P/L's on the 1% one-day ahead VaR obtained with either (i) Historical Simulation with a 250-day rolling window (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 data and Historical Simulation. But these R^2 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 use 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. While 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 non-normality 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. Since 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.

References

Alexander, G. and A. Baptista, (2004), "A Comparison of VaR and CVaR Constraints on Portfolio Selection with the Mean-Variance Model," *Management Science*, 50, 1261–1273.

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.

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

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

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

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

Blanco, C. and S. Blomstrom (1999), "VaR Applications: Setting VaR-based Limits," Working Paper, Financial Engineering Associates, Inc.

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

Campbell, S. D. (2007), "A Review of Backtesting and Backtesting Procedures," *Journal of Risk*, 9, 1-17.

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

Cuoco, D. H. He, S. Isaenko (2008), "Optimal Dynamic Trading Strategies with Risk Limits," *Operations Research*, 56, 358-368.

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.

Elsinger, H, A. Lehar, and M. Summer (2006), "Risk Assessment for Banking Systems," *Management Science*, 52, 1301-1314.

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.

Jorion, P. (2006). *Value-at-Risk: the New Benchmark for Managing Financial Risk*, Third Edition, McGraw-Hill: Chicago.

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.

Lopez, J.A. (1999), "Regulatory Evaluation of Value-at-Risk Models," *Journal of Risk*, 1, 37-64.

Lopez, J.A. and C. Walter (2001), "Evaluating Covariance Matrix Forecasts in a Value-at-Risk Framework," *Journal of Risk*, 3, 69-98.

Lucas, A. (2000), "A Note on Optimal Estimation from a Risk Management Perspective under Possibly Misspecified Tail Behavior," *Journal of Business and Economic Statistics*, 18(1), 31-39.

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

Mincer, J., and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts" in J. Mincer (Ed.), *Economic Forecasts and Expectations*, Cambridge: National Bureau of Economic Research, 3-46.

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.

Perignon, C. and D.R. Smith (2007), "Which Value-at-Risk Method Works Best for Bank Trading Revenues?" Working Paper, Simon Fraser University.

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

Taylor, J.W. (2005), "Generating Volatility Forecasts from Value at Risk Estimates," *Management Science*, 51, 712–725.

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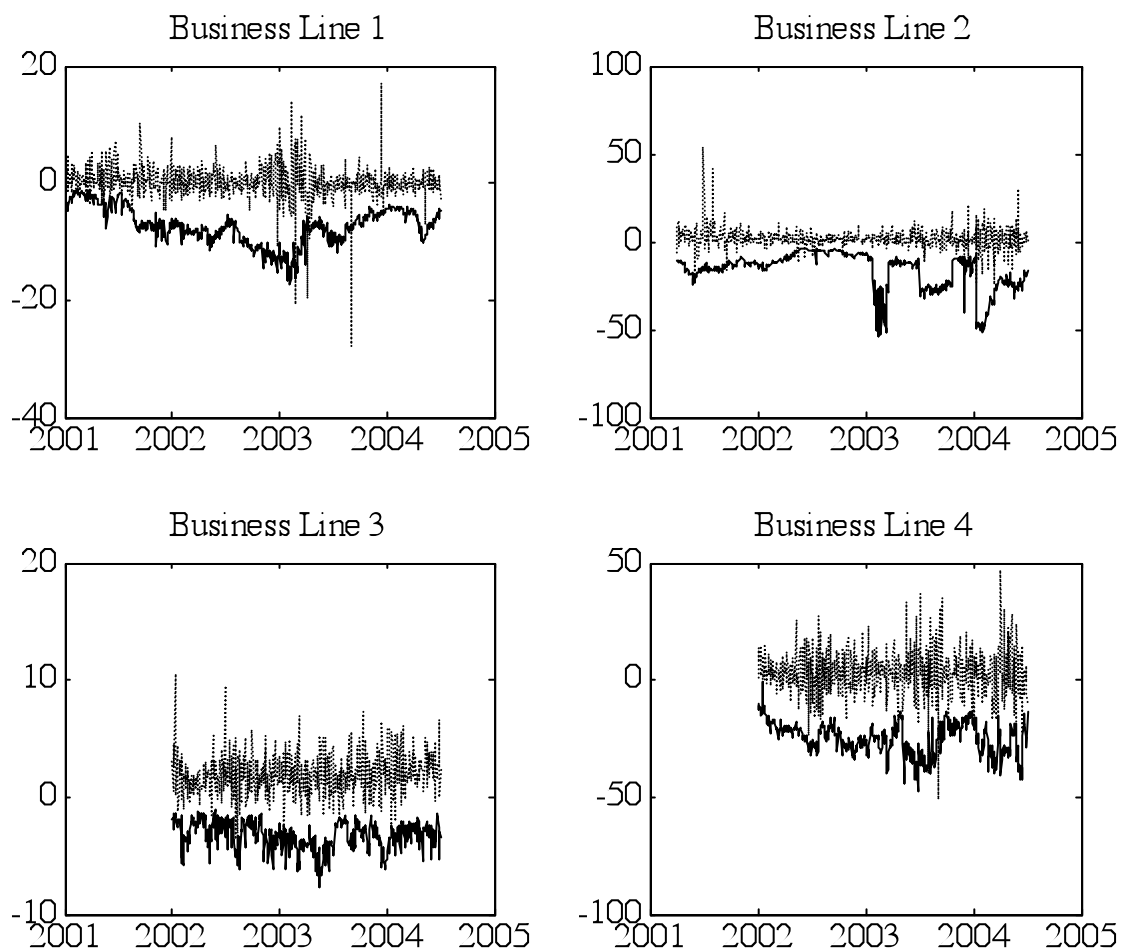
Notes to Figure: We plot the P/Ls (dashed lines) and 1-day, 1% VaRs (solid lines) from the four business lines.

Figure 2: Absolute Demeaned P/Ls for Four Business Lines
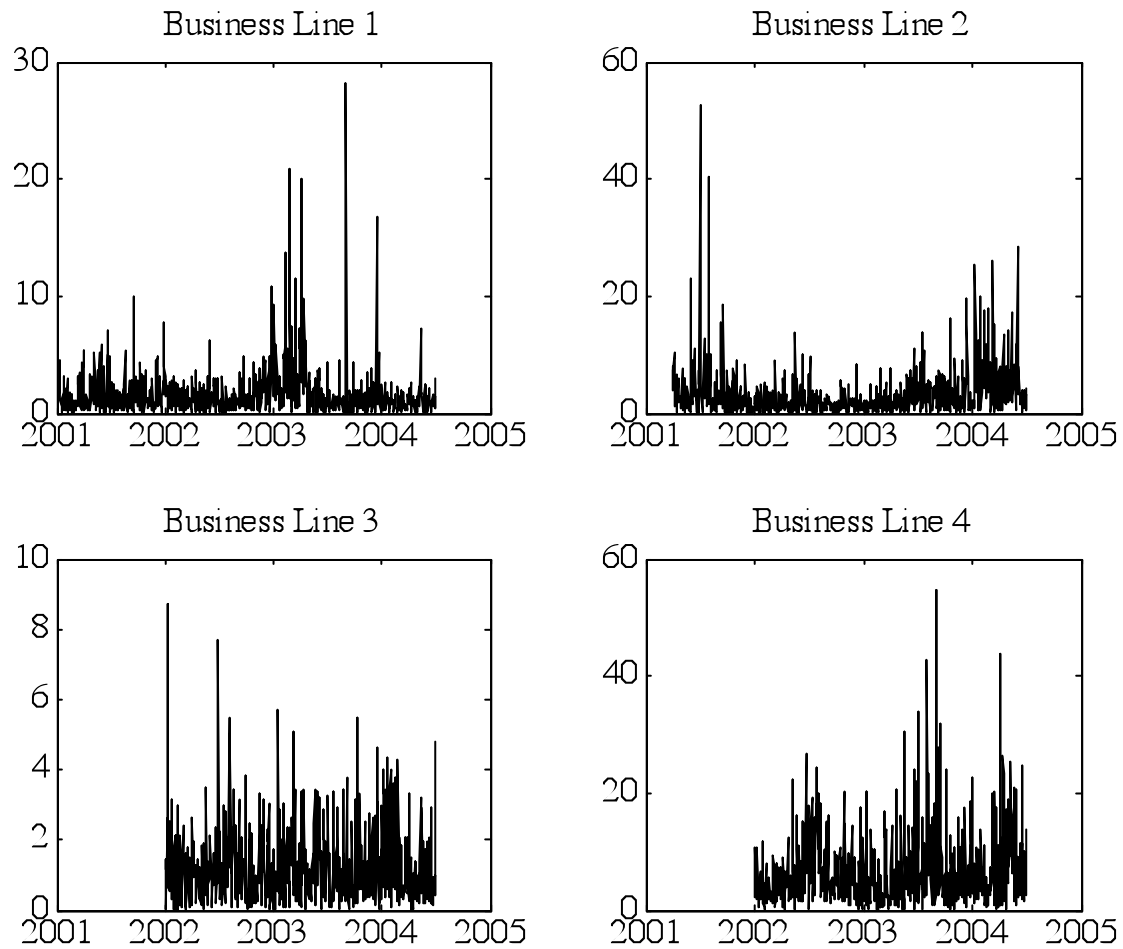
Notes to Figure: We subtract the sample mean from each of the four P/Ls in Figure 1 and plot the absolute value of these demeaned P/Ls.

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	Kupiec
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
1000	0.076	0.147	0.076	0.157	0.119	0.067	0.094	0.125	0.117
1250	0.102	0.146	0.055	0.128	0.111	0.075	0.112	0.140	0.116
1500	0.101	0.131	0.064	0.127	0.095	0.071	0.112	0.137	0.121

5 % VaR

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.089	0.097	0.209	0.142	0.072	0.113	0.095	0.095	0.112
1250	0.092	0.093	0.161	0.149	0.061	0.121	0.096	0.098	0.104
1500	0.087	0.098	0.152	0.160	0.063	0.114	0.095	0.097	0.095

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. Kupiec is a test of unconditional coverage. 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	Kupiec
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
1000	0.371	0.519	0.168	0.182	0.342	0.618	0.508	0.473	0.025
1250	0.434	0.564	0.187	0.228	0.370	0.682	0.542	0.503	0.025
1500	0.446	0.603	0.202	0.244	0.410	0.737	0.585	0.564	0.023

Business Line 2

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.281	0.361	0.173	0.218	0.421	0.532	0.390	0.323	0.070
1250	0.281	0.400	0.160	0.265	0.475	0.580	0.381	0.327	0.090
1500	0.280	0.423	0.160	0.304	0.507	0.617	0.426	0.371	0.085

Business Line 3

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.106	0.146	0.036	0.051	0.044	0.526	0.137	0.121	0.011
1250	0.131	0.127	0.039	0.049	0.047	0.611	0.137	0.130	0.009
1500	0.147	0.126	0.031	0.042	0.038	0.686	0.150	0.147	0.005

Business Line 4

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.333	0.419	0.202	0.267	0.503	0.574	0.449	0.375	0.085
1250	0.317	0.458	0.196	0.343	0.544	0.612	0.455	0.392	0.112
1500	0.329	0.510	0.202	0.389	0.597	0.655	0.488	0.427	0.114

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. Kupiec is a test of unconditional coverage. 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	Kupiec
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
1000	0.484	0.696	0.251	0.270	0.679	0.692	0.586	0.607	0.019
1250	0.543	0.762	0.294	0.325	0.756	0.761	0.665	0.675	0.013
1500	0.593	0.815	0.328	0.379	0.819	0.851	0.720	0.722	0.010

Business Line 2

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.419	0.658	0.276	0.487	0.784	0.702	0.558	0.549	0.164
1250	0.466	0.721	0.310	0.548	0.842	0.741	0.625	0.609	0.140
1500	0.504	0.781	0.335	0.606	0.900	0.819	0.681	0.655	0.140

Business Line 3

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.113	0.123	0.062	0.031	0.052	0.511	0.102	0.107	0.002
1250	0.114	0.125	0.069	0.029	0.044	0.583	0.113	0.122	0.001
1500	0.107	0.125	0.065	0.033	0.053	0.713	0.117	0.116	0.001

Business Line 4

Sample	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	KS	CVM	Kupiec
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
1000	0.443	0.705	0.278	0.540	0.838	0.729	0.606	0.592	0.150
1250	0.501	0.769	0.317	0.613	0.888	0.768	0.667	0.658	0.125
1500	0.548	0.824	0.361	0.677	0.936	0.837	0.734	0.713	0.132

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. Kupiec is a test of unconditional coverage. 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	VIX	KS	CVM	CavMult	Kupiec
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

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	VIX	KS	CVM	CavMult	Kupiec
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

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	VIX	KS	CVM	CavMult	Kupiec
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

	LB(1)	LB(5)	Markov	Weibull	Geometric	Caviar	VIX	KS	CVM	CavMult	Kupiec
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 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. Kupiec is a test of unconditional coverage. 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.

Table 8: Forecasting Portfolio Risk

Business Line 1	R^2
Real data	0.0360
Simulated data and HS VaR	0.0015
Simulated data and true VaR	0.0812

Business Line 2	R^2
Real data	0.0694
Simulated data and HS VaR	0.0438
Simulated data and true VaR	0.1415

Business Line 3	R^2
Real data	0.0148
Simulated data and HS VaR	0.0009
Simulated data and true VaR	0.0080

Business Line 4	R^2
Real data	0.0191
Simulated data and HS VaR	0.0185
Simulated data and true VaR	0.1211

Notes to Table: 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 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.

Research Papers
2009

CREATES
Center for Research in Econometric
Analysis of Time Series

2009-21: Ole E. Barndorff-Nielsen, José Manuel Corcuera and Mark Podolskij: Multipower Variation for Brownian Semistationary Processes

2009-22: Giuseppe Cavaliere, Anders Rahbek and A.M.Robert Taylor: Co-integration Rank Testing under Conditional Heteroskedasticity

2009-23: Michael Frömmel and Robinson Kruse: Interest rate convergence in the EMS prior to European Monetary Union

2009-24: Dominique Guégan: A Meta-Distribution for Non-Stationary Samples

2009-25: Ole E. Barndorff-Nielsen and Almut E. D. Veraart: Stochastic volatility of volatility in continuous time

2009-26: Tim Bollerslev and Viktor Todorov: Tails, Fears and Risk Premia

2009-27: Kim Christensen, Roel Oomen and Mark Podolskij: Realised Quantile-Based Estimation of the Integrated Variance

2009-28: Takamitsu Kurita, Heino Bohn Nielsen and Anders Rahbek: An I(2) Cointegration Model with Piecewise Linear Trends: Likelihood Analysis and Application

2009-29: Martin M. Andreasen: Stochastic Volatility and DSGE Models

2009-30: Eduardo Rossi and Paolo Santucci de Magistris: Long Memory and Tail dependence in Trading Volume and Volatility

2009-31: Eduardo Rossi and Paolo Santucci de Magistris: A No Arbitrage Fractional Cointegration Analysis Of The Range Based Volatility

2009-32: Alessandro Palandri: The Effects of Interest Rate Movements on Assets' Conditional Second Moments

2009-33: Peter Christoffersen, Redouane Elkamhi, Bruno Feunou and Kris Jacobs: Option Valuation with Conditional Heteroskedasticity and Non-Normality

2009-34: Peter Christoffersen, Steven Heston and Kris Jacobs: The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work so Well

2009-35: Peter Christoffersen, Jeremy Berkowitz and Denis Pelletier: Evaluating Value-at-Risk Models with Desk-Level Data