Backtesting Value–at–Risk: A Duration–Based Approach

PETER CHRISTOFFERSEN
   McGill University, CIRANO, and CIREQ

DENIS PELLETIER
   *North Carolina State University*

ABSTRACT
Financial risk model evaluation or backtesting is a key part of the internal model's approach to market risk management as laid out by the Basle Committee on Banking Supervision. However, existing backtesting methods have relatively low power in realistic small sample settings. Our contribution is the exploration of new tools for backtesting based on the duration of days between the violations of the Value–at–Risk. Our Monte Carlo results show that in realistic situations, the new duration–based tests have considerably better power properties than the previously suggested tests.

KEYWORDS: GARCH, kurtosis, risk model evaluation

Financial risk model evaluation or *backtesting* is a key part of the internal model's approach to market risk management as laid out by the Basle Committee on Banking Supervision (1996). However, existing backtesting methods such as those developed in Christoffersen (1998) have relatively small power in realistic small sample settings. Methods suggested in Berkowitz (2001) fare better, but rely on information such as the shape of the left tail of the portfolio return distribution, which is often not available. By far the most common risk measure is Value-at-Risk (VaR), which is defined as a conditional quantile of the return distribution, and it says nothing about the shape of the tail to the left of the quantile.

We will refer to an event where the ex post portfolio loss exceeds the ex ante VaR measure as a *violation*. Of particular importance in backtesting is the clustering of violations. An institution's internal risk management team as well as external
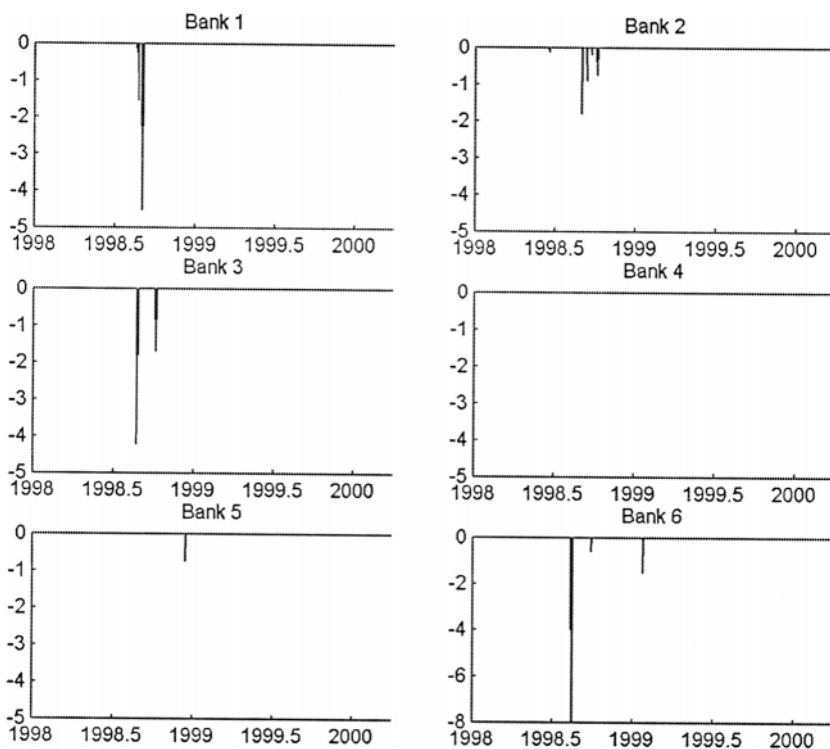
Figure 1 VaR exceedences from six major commercial banks Berkowitz and O'Brien (2002).

supervisors explicitly want to be able to detect clustering in violations. Large losses that occur in rapid succession are more likely to lead to disastrous events such as bankruptcy.

In the previous literature, due to the lack of real portfolio data, the evaluation of VaR techniques were largely based on artificial portfolios. Examples in this tradition include Beder (1995), Hendricks (1996), Kupiec (1995), Marshall and Siegel (1997), Pritsker (1997), and Christoffersen, Hahn and Inoue (2001). But recently Berkowitz and O'Brien (2002) have reported on the performance of actual VaR forecasts from six large (and anonymous) U.S. commercial banks.[1] Figure 1 reproduces a picture from their article that shows the VaR exceedences from the six banks reported in standard deviations of the portfolio returns. Even though the banks tend to be conservative — they have fewer than expected violations — the exceedences are large and appear to be clustered in time and across banks. The majority of violations appear to take place during the August 1998 Russia default and ensuing Long‑Term Capital Management (LTCM) debacle. From the perspective of a regulator worried about systemic risk, rejecting a particular

---

[1] Barone‑Adesi, Giannopoulos, and Vosper (2002) provides another example using real‑life portfolio returns.

bank's risk model due to the clustering of violations is particularly important if the violations also happen to be correlated across banks.

The detection of violation clustering is particularly important because of the widespread reliance on VaRs calculated from the so-called historical simulation (HS) technique. In the HS methodology, a sample of historical portfolio returns using current portfolio weights is first constructed. The VaR is then simply calculated as the *unconditional* quantile from the historical sample. The HS method thus largely ignores the last 20 years of academic research on conditional asset return models. Time variability is only captured through the rolling historical sample. In spite of forceful warnings, such as Pritsker (2001), the model-free nature of the HS technique is viewed as a great benefit by many practitioners. The widespread use of the HS technique motivates us to focus attention on backtesting VaRs calculated using this method.

While alternative methods for calculating portfolio measures such as the VaR have been investigated in for example Jorion (2000) and Christoffersen (2003), available methods for backtesting are still relatively few. Our contribution is thus the exploration of a new tool for backtesting based on the duration of days between the violations of the risk metric. The chief insight is that if the one-day-ahead VaR model is correctly specified for coverage rate, p, then, every day, the conditional expected duration until the next violation should be a constant $1/p$ days. We suggest various ways of testing this null hypothesis and we conduct a Monte Carlo analysis that compares the new tests to those currently available. Our results show that in many realistic situations, the duration-based tests have better power properties than the previously suggested tests. The size of the tests is easily controlled using the Monte Carlo testing approach of Dufour (2000). This procedure is described in detail below.

We hasten to add that the sort of omnibus backtesting procedures suggested here are meant as complements to, not substitutes for, the statistical diagnostic tests carried out on various aspects of the risk model in the model estimation stage. The tests suggested in this article can be viewed either as a final diagnostic for an internal model builder or alternatively as a feasible diagnostic for an external model evaluator for whom only limited, aggregate portfolio information is available.

The article is structured as follows: Section 1 outlines the previous first-order Markov tests. Section 2 suggests the new duration-based tests. Section 3 discusses details related to the implementation of the tests. Section 4 contains Monte Carlo evidence on the performance of the tests. Section 5 considers backtesting of tail density forecasts. Section 6 concludes.

1 EXTANT PROCEDURES FOR BACKTESTING VaR

Consider a time series of daily ex post portfolio returns, R_t, and a corresponding time series of ex ante Value-at-Risk forecasts, $VaR_t(p)$ with promised coverage rate p, such that ideally $\Pr_{t-1}(R_t < -VaR_t(p)) = p$. The negative sign arises from the convention of reporting the VaR as a positive number.

Define the hit sequence of VaR_t violations as

$$I_t = \begin{cases} 1, & \text{if } R_t < -VaR_t(p) \\ 0, & \text{else} \end{cases}. \tag{1}$$

Notice that the hit sequence appears to discard a large amount of information regarding the size of violations, etc. Recall, however, that the VaR forecast does not promise violations of a certain magnitude, but rather only their conditional frequency, that is, p. This is a major drawback of the VaR measure which we will discuss in Section 5.

Christoffersen (1998) tests the null hypothesis that

$$I_t \sim i.i.d. \; Bernoulli(p)$$

against the alternative that

$$I_t \sim i.i.d. \; Bernoulli(\pi)$$

and refers to this as the test of correct unconditional coverage (*uc*)

$$H_{0,uc} : \pi = p, \tag{2}$$

which is a test that on average the coverage is correct. The above test implicitly assumes that the hits are independent, an assumption which we now test explicitly. In order to test this hypothesis, an alternative is defined where the hit sequence follows a first-order Markov sequence with switching probability matrix

$$\Pi = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}, \tag{3}$$

where π_{ij} is the probability of an i on day $t-1$ being followed by a j on day t. The test of independence (*ind*) is then

$$H_{0,ind} : \pi_{01} = \pi_{11}. \tag{4}$$

Finally, the two tests can be combined in a test of conditional coverage (*cc*),

$$H_{0,cc} : \pi_{01} = \pi_{11} = p. \tag{5}$$

The idea behind the Markov alternative is that clustered violations represent a signal of risk model misspecification. Violation clustering is important as it implies repeated severe capital losses to the institution which together could result in bankruptcy.

Notice, however, that the Markov first-order alternative may have limited power against general forms of clustering. The first point of this article is to establish more general tests for clustering which nevertheless rely only on information in the hit sequence. Throughout the article we implicitly assume that the VaR is for a one-day horizon. To apply this backtesting framework to an horizon of more than one day, we would have to use nonoverlapping observations.[2]

---

[2] We implicitly assume that we observe the return process at least as frequently as we compute the VaR.

2 DURATION-BASED TESTS OF INDEPENDENCE

The above tests are reasonably good at catching misspecified risk models when the temporal dependence in the hit sequence is of a simple first-order Markov structure. However, we are interested in developing tests that have power against more general forms of dependence but which still rely on estimating only a few parameters.

The intuition behind the duration-based tests suggested below is that the clustering of violations will result in an excessive number of relatively short and relatively long no-hit durations, corresponding to market turbulence and market calm, respectively. Motivated by this intuition, we consider the duration of time (in days) between two VaR violations (i.e., the no-hit duration) as

$$D_i = t_i - t_{i-1}, \tag{6}$$

where t_i denotes the day of violation number i.[3]

Under the null hypothesis that the risk model is correctly specified, the no-hit duration should have no memory and a mean duration of $1/p$ days. To verify the no-memory property, note that under the null hypothesis we have the discrete probability distribution

$$\Pr(D = 1) = p$$
$$\Pr(D = 2) = (1 - p)p$$
$$\Pr(D = 2) = (1 - p)^2 p$$
$$\cdots$$
$$\Pr(D = d) = (1 - p)^{d-1} p.$$

A duration distribution is often best understood by its hazard function, which has the intuitive definition of the probability of a violation on day D after we have gone $D - 1$ days without a violation. The above probability distribution implies a flat discrete hazard function as the following derivation shows:

$$\lambda(d) = \frac{\Pr(D = d)}{1 - \sum_{j<d} \Pr(D = j)}$$
$$= \frac{(1 - p)^{d-1} p}{1 - \sum_{j=0}^{d-2} (1 - p)^j p}$$
$$= p.$$

The only memory-free (continuous)[4] random distribution is the exponential, thus we have that under the null the distribution of the no-hit durations should be

$$f_{\exp}(D; p) = p \exp(-pD). \tag{7}$$

---

[3] For a general introduction to duration modeling, see Kiefer (1988) and Gourieroux (2000).

[4] Notice that we use a continuous distribution even though we are counting time in days. This discreteness bias will be acounted for in the Monte Carlo tests. The exponential distribution can also be viewed as the continuous-time limit of the above discrete-time process. See Poirier (1995).

In order to establish a statistical test for independence we must specify a (parsimonious) alternative that allows for duration dependence. As a very simple case, consider the Weibull distribution where

$$f_W(D; a, b) = a^b b D^{b-1} \exp(-(aD)^b). \tag{8}$$

The Weibull distribution has the advantage that the hazard function has a closed-form representation, namely

$$\lambda_W(D) \equiv \frac{f_W(D)}{1 - F_W(D)} = a^b b D^{b-1}, \tag{9}$$

where the exponential distribution appears as a special case with a flat hazard, when $b = 1$. The Weibull will have a decreasing hazard function when $b < 1$, which corresponds to an excessive number of very short durations (very volatile periods) and an excessive number of very long durations (very tranquil periods). This could be evidence of misspecified volatility dynamics in the risk model.

Because of the bankruptcy threat from VaR violation clustering, the null hypothesis of independence is of particular interest. We therefore want to explicitly test the null hypothesis

$$H_{0, ind} : b = 1. \tag{10}$$

We could also use the gamma distribution under the alternative hypothesis. The probability density function (p.d.f.) in this case is

$$f_\Gamma(D; a, b) = \frac{a^b D^{b-1} \exp(-aD)}{\Gamma(b)}, \tag{11}$$

which also nests the exponential when $b = 1$. In this case we therefore also have the independence test null hypothesis as

$$H_{0, ind} : b = 1. \tag{12}$$

The gamma distribution does not have a closed-form solution for the hazard function, but the first two moments are b/a and b/a^2, respectively, so the notion of excess dispersion, which is defined as the variance over the squared expected value, is simply $1/b$. Note that the average duration in the exponential distribution is $1/p$ and the variance of durations is $1/p^2$, thus the notion of excess dispersion is one in the exponential distribution.

The above duration tests can potentially capture higher-order dependence in the hit sequence by simply testing the unconditional distribution of the durations. Dependence in the hit sequence may show up as an excess of relatively long no-hit durations (quiet periods) and an excess of relatively short no-hit durations, corresponding to violation clustering. However, in the above tests, any information in the ordering of the durations is completely lost. The information in the temporal ordering of no-hit durations could be captured using the framework of Engle and Russel's (1998) exponential autoregressive conditional duration (EACD)

model. In the EACD$(1, 0)$ model, the conditional expected duration takes the following form:

$$E_{i-1}[D_i] \equiv \psi_i = \omega + \alpha D_{i-1} \tag{13}$$

with $\alpha \in [0, 1)$. Assuming an underlying exponential density with mean equal to one, the conditional distribution of the duration is

$$f_{EACD}(D_i \,|\, \psi_i) = \frac{1}{\psi_i} \exp\left(-\frac{D_i}{\psi_i}\right). \tag{14}$$

The null of independent no-hit durations would then correspond to

$$H_{0,\,ind} : \alpha = 0. \tag{15}$$

Excess dispersion in the EACD$(1, 0)$ model is defined as

$$V[D_i]/E[D_i]^2 = \frac{1}{1 - 2\alpha^2}, \tag{16}$$

so that the ratio of the standard deviation to the mean duration is above one if $\alpha > 0$.

In our test specifications, the information set only contains past durations, but it could be extended to include all the conditioning information used to compute the VaR for example. This would translate into adding variables other than D_{i-1} into the right-hand side of Equation (13).

3 TEST IMPLEMENTATION

We will first discuss the specific implementation of the hit sequence tests suggested above. Later we will simulate observations from a realistic portfolio return process and calculate risk measures from the popular HS risk model, which in turn provides us with hit sequences for testing.

3.1 Implementing the Markov Tests

The likelihood function for a sample of T i.i.d. observations from a Bernoulli variable, I_t, with known probability p is written as

$$L(I, p) = p^{T_1}(1 - p)^{T - T_1}. \tag{17}$$

where T_1 is the number of ones in the sample. The likelihood function for an i.i.d. Bernoulli with unknown probability parameter, π_1, to be estimated is

$$L(I, \pi_1) = \pi_1^{T_1}(1 - \pi_1)^{T - T_1}. \tag{18}$$

The maximum-likelihood (ML) estimate of π_1 is

$$\hat{\pi}_1 = T_1/T \tag{19}$$

and we can thus write a likelihood ratio test of unconditional coverage as

$$LR_{uc} = 2(\ln L(I, \hat{\pi}_1) - \ln L(I, p)). \tag{20}$$

For the independence test, the likelihood under the alternative hypothesis is

$$L(I, \pi_{01}, \pi_{11}) = (1 - \pi_{01})^{T_0 - T_{01}} \pi_{01}^{T_{01}} (1 - \pi_{11})^{T_1 - T_{11}} \pi_{11}^{T_{11}}, \tag{21}$$

where T_{ij} denotes the number of observations with a j following an i. The ML estimates are

$$\hat{\pi}_{01} = T_{01}/T_0 \tag{22}$$

$$\hat{\pi}_{11} = T_{11}/T_1 \tag{23}$$

and the independence test statistic is

$$LR_{ind} = 2(\ln L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - \ln L(I, \hat{\pi}_1)). \tag{24}$$

Finally, the test of conditional coverage is written as

$$LR_{cc} = 2(\ln L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - \ln L(I, p)). \tag{25}$$

We note that all the tests are carried out conditioning on the first observation. The tests are asymptotically distributed as chi-square with degree of freedom one for the *uc* and *ind* tests and two for the *cc* test. But we will rely on finite sample *p*-values below.

Finally, as a practical matter, if the sample at hand has $T_{11} = 0$, which can easily happen in small samples and with small coverage rates, then we calculate the first-order Markov likelihood as

$$L(I, \pi_{01}, \pi_{11}) = (1 - \pi_{01})^{T_0 - T_{01}} \pi_{01}^{T_{01}} \tag{26}$$

and carry out the tests as above.

3.2   Implementing the Weibull and EACD Tests

In order to implement our tests based on the duration between violations, we first need to transform the hit sequence into a duration series D_i. While doing this transformation we also create the series C_i to indicate if a duration is censored ($C_i = 1$) or not ($C_i = 0$). Except for the first and last duration, the procedure is straightforward; we just count the number of days between each violation and set $C_i = 0$. 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. If instead the hit sequence starts with a one, then D_1 is simply the number of days until the second hit and $C_1 = 0$.

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 duration is right-censored. In the same manner, if the last observation of the hit sequence is a one, then $D_{N(T)} = t_{N(T)} - t_{N(T)-1}$ and $C_{N(T)} = 0$.

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

$$\ln L(D; \Theta) = C_1 \ln S(D_1) + (1 - C_1)\ln f(D_1) + \sum_{i=2}^{N(T)-1} \ln(f(D_i)) \tag{27}$$

$$+ C_{N(T)} \ln S(D_{N(T)}) + (1 - C_{N(T)})\ln f(D_{N(T)}). \tag{28}$$

Once the durations are computed and the truncations taken care of then the likelihood ratio tests can be calculated in a straightforward fashion. The only added complication is that the ML estimates are no longer available in closed form, they must be found using numerical optimization.[5] For the unrestricted EACD likelihood, this implies maximizing simultaneously over two parameters, α and ω. For the unrestricted Weibull likelihood, we only have to numerically maximize it over one parameter, since for a given value of b, the first-order condition with respect to a has an explicit solution:[6]

$$\hat{a} = \left(\frac{N(T) - C_1 - C_{N(T)}}{\sum_{i=1}^{N(T)} D_i^b} \right)^{1/b}. \tag{29}$$

3.3 Finite Sample Inference

While the large-sample distributions of the likelihood ratio tests we have suggested above are well known,[7] they may not lead to reliable inference in realistic risk management settings. The nominal sample sizes can be reasonably large, say two to four years of daily data, but the scarcity of violations of, for example, the 1% VaR renders the effective sample size small. In this section we therefore introduce the Dufour (2000) Monte Carlo testing technique.

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, $i = 1, \ldots, N$. We denote by LR_0 the test 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 in our application. We thus benefit from the advantage of not having nuisance parameters under the null hypothesis.

We next rank LR_i, $i = 0, \ldots, N$ in nondecreasing order and obtain the Monte Carlo p-value $\hat{p}_N(LR_0)$, where

$$\hat{p}_N(LR_0) = \frac{N\hat{G}_N(LR_0) + 1}{N + 1} \tag{30}$$

---

[5] We have also investigated Lagrange multiplier (LM) tests which require less numerical optimization than do likelihood ratio (LR tests). However, in finite sample simulations we found that the power in the LM tests were lower than in the LR tests; thus, we only report LR results below.

[6] For numerical stability, we recommend working with a^b instead of a, since b can take values close to zero.

[7] Testing $\alpha = 0$ in the EACD(1, 0) model presents a potential difficulty asymptotically in that it is on the boundary of the parameter space. However, the Monte Carlo (MC) method we apply is valid even in this case. See Andrews (2001) for details.

with

$$\hat{G}_N(LR_0) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(LR_i > LR_0), \qquad (31)$$

where $\mathbf{1}(*)$ takes on the value one if $*$ is true and the value zero otherwise.

When working with binary sequences, the test values can only take a countable number of distinct values. Therefore we need a rule to break ties between the test value obtained from the sample and those obtained from Monte Carlo simulation under the null hypothesis. The tiebreaking procedure is as follows: For each test statistic, LR_i, $i = 0, \ldots, N$, we draw an independent realization of a uniform distribution on the [0; 1] interval. Denote these draws by U_i, $i = 0, \ldots, N$. The Monte Carlo p-value is now given by

$$\tilde{p}_N(LR_0) = \frac{N\tilde{G}_N(LR_0) + 1}{N + 1} \qquad (32)$$

with

$$\tilde{G}_N(LR_0) = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(LR_i \leq LR_0) + \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(LR_i = LR_0)\mathbf{1}(U_i \geq U_0). \qquad (33)$$

There are two additional advantages of using a simulation procedure. The first is that possible systematic biases arising from the use of continuous distributions to study discrete processes are accounted for. 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. On the other hand, bootstrap procedures could be inconsistent in this case.

4 BACKTESTING VaRs FROM HISTORICAL SIMULATION

We now assess the power of the proposed duration tests in the context of a Monte Carlo study. Consider a portfolio where the returns are drawn from a GARCH(1,1)-$t(d)$ model with an asymmetric leverage effect, that is,

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

where the innovation z_{t+1}s are drawn independently from a Student's $t(d)$ distribution. Notice that the innovations have been rescaled to ensure that the conditional variance of return will be σ_{t+1}^2.

In the simulations below we choose the following parameterization:

$\alpha = 0.1$
$\theta = 0.5$
$\beta = 0.85$
$\omega = 3.9683e - 6$
$d = 8,$

where ω is set to target an annual standard deviation of 0.20. The parameters imply a daily volatility persistence of 0.975, a mean of zero, a conditional skewness of zero, and a conditional (excess) kurtosis of 1.5. This particular data generating process (DGP) is constructed to form a realistic representation of an equity portfolio return distribution.[8]

The risk measurement method under study is the popular HS technique. It takes the VaR on a certain day to be simply the unconditional quantile of the past T_e daily observations. Specifically

$$VaR_{t+1}^p = -percentile(\{R_\tau\}_{\tau=t-T_e+1}^t, 100p).$$

From the return sample and the above VaR, we are implicitly assuming that $1 is invested each day. Equivalently, the VaR can be interpreted as being calculated as a percentage of the portfolio value.

In practice, the sample size is often determined by practical considerations such as the amount of effort involved in valuing the current portfolio holdings using past prices on the underlying securities. For the purposes of this Monte Carlo experiment, we set $T_e = 250$ or $T_e = 500$, corresponding to roughly one or two years of trading days.

In practice the VaR coverage rate, p, is typically chosen to be either 1% or 5%, and below we assess the power to reject the HS model using either of those rates. Figure 2 shows a return sample path from the above GARCH-$t(d)$ process along with the 1% and 5% VaRs from the HS model (with $T_e = 500$). Notice the peculiar step-shaped VaRs resulting from the HS method. Notice also the infrequent changes in the 1% VaR.[9]

The 1% VaR exceedances from the return sample path are shown in Figure 3, reported in daily standard deviations of returns. The simulated data in Figure 3 can thus be compared with the real-life data in Figure 1, which was taken from Berkowitz and O'Brien (2002). Notice that the simulated data shares the stylized features with the real-life data in Figure 1.[10]

Before calculating actual finite sample power in the suggested tests we want to give a sense of the appropriateness of the duration-dependence alternative. To this end we simulate one very long realization (five million observations) of the GARCH return process and calculate 1% and 5% VaRs from HS with a rolling set of 500 in-sample returns. The zero-one hit sequence is then calculated from the ex post daily returns and the ex ante VaRs, and the sequence of durations between violations is calculated from the hit sequence. From this duration sequence we fit a Weibull distribution and calculate the hazard function from it. We also estimate nonparametrically the empirical hazard function of the simulated durations via

---

[8] The parameter values are similar to estimates of this GARCH model on daily S&P 500 returns (not reported here) and to estimates on daily FX returns published in Bollerslev (1987).

[9] When $T_e = 250$ and $p = 1\%$, the VaR is calculated as the simple average between the second and third lowest return.

[10] Note that we have simulated 1000 observations in Figure 3, while Figure 1 contains between 550 and 750 observations per bank.
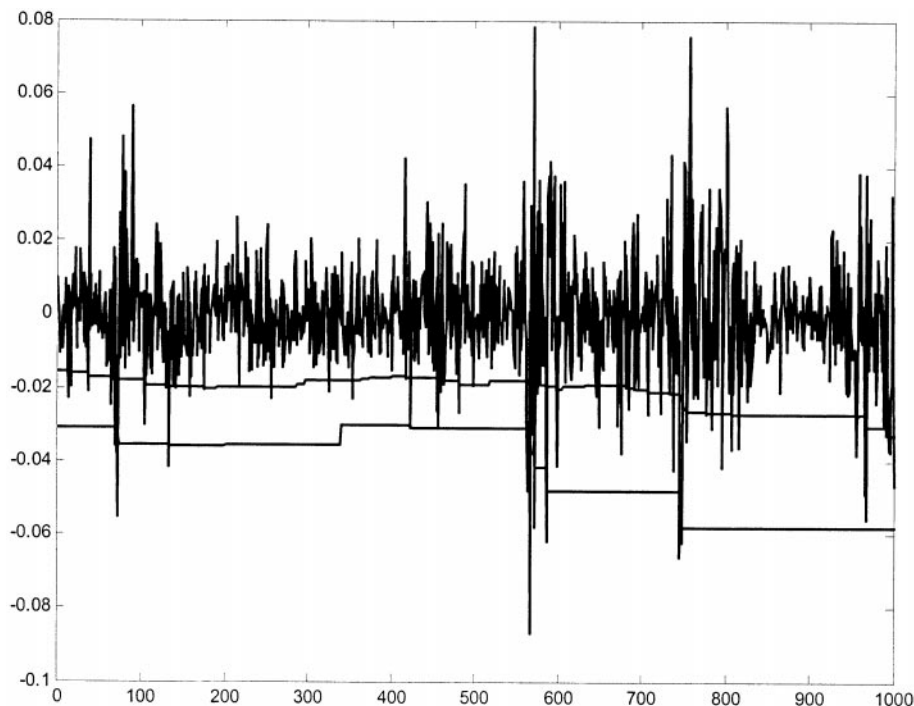
Figure 2 GARCH-$t(d)$ simulated portfolio returns with 1% and 5% VaR from HS with $T_e = 500$.

the Kaplan-Meier product-limit estimator of the survival function [see Kiefer (1988)]. These Weibull and empirical hazards are estimated over intervals of 10 days, so if there is a probability p of getting a hit at each day then the probability that a given duration will last 10 days or less is

$$\sum_{i=1}^{10} \Pr(D = i) = \sum_{i=1}^{10} (1-p)^{i-1} p$$
$$= 1 - (1-p)^{10}.$$

For p equal to 1% and 5% we get a constant hazard of 0.0956 and 0.4013, respectively, over a 10-day interval.

We see in Figure 4 that the hazards are distinctly downward sloping, which corresponds to positive duration dependence. The relevant flat hazard corresponding to i.i.d. violations is superimposed for comparison. Figure 4 also shows that the GARCH and the Weibull hazards are reasonably close together, which suggests that the Weibull distribution offers a useful alternative hypothesis in this type of test.

Figure 5 shows the duration dependence via simple histograms of the duration between the violations from the HS VaRs. The top panel again shows the 1% VaR and the bottom panel shows the 5% VaR.
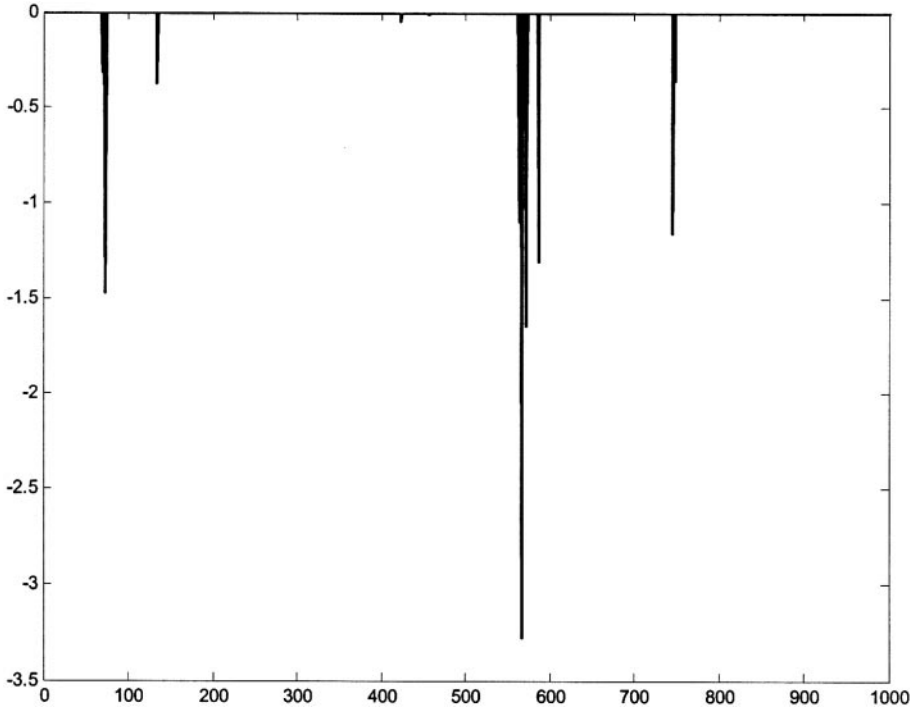
Figure 3 GARCH-$t(d)$ simulated portfolio returns with exceedences of 1% VaRs from HS with $T_e = 500$. Reported in standard deviations of returns.

Data and other resource constraints often force risk managers to backtest their models on relatively short backtesting samples. We therefore conduct our power experiment with samples sizes from 250 to 1500 days in increments of 250 days. Thus our backtesting samples correspond to approximately one through six years of daily returns.

Below we simulate GARCH returns and calculate HS VaRs and the various tests in 5000 Monte Carlo replications. We present three types of results. We first present the raw power results, which are simply calculated as the frequency of rejections of the null hypothesis in the simulation samples for which we can perform the tests. In order to compute the p-values of the tests we simulate $N = 9999$ hit sequence samples under the null hypothesis that the sequences are distributed i.i.d. Bernoulli(p).

In the simulations we reject the samples for which we cannot compute the tests. For example, to compute the independence test with the Markov model, we need at least one violation, otherwise the LR test is equal to zero when we calculate the likelihood from Equation (26). Similarly, we need at least one noncensored duration and an additional possibly censored duration to perform the Weibull[11] and EACD

---

[11] The likelihood of the Weibull distribution can be unbounded when we have only one uncensored observation. When this happens we discard the sample.
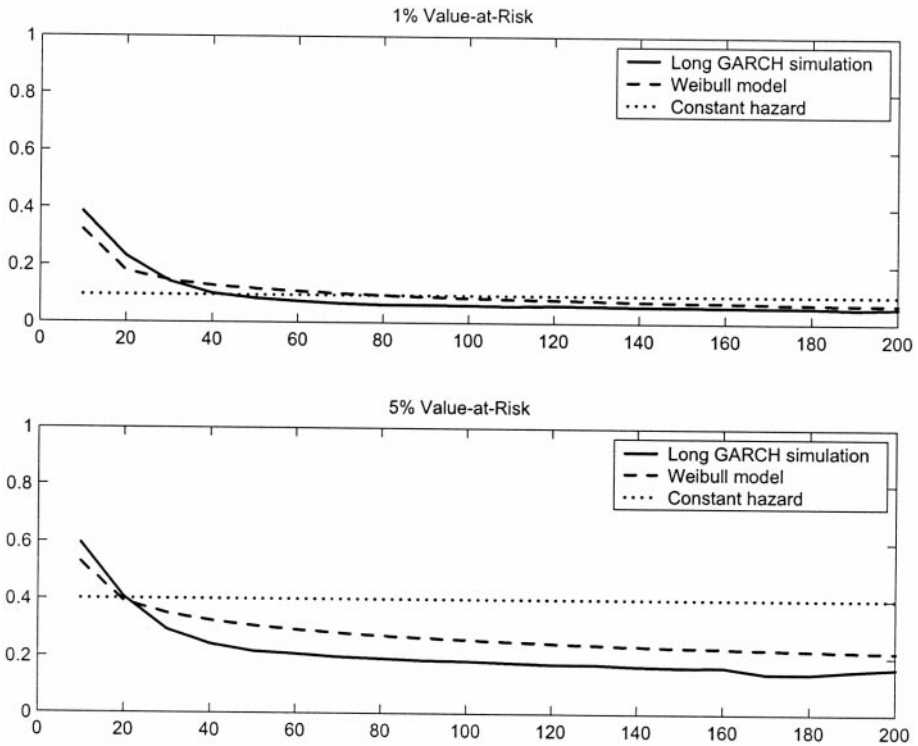
Figure 4 Data-based and Weibull-based hazard functions of durations between VaR violations. HS risk model on GARCH-$t(d)$ portfolio returns with $T_e = 500$.

independence tests. Of course, this constitutes a nontrivial sample selection rule for the smallest sample sizes and the 1% VaR coverage rate in particular. We therefore also present the sample selection frequency, that is, the fraction of simulated samples for which we can compute each test. Finally, we report effective power, which corresponds to multiplying the raw power by the sample selection frequency.

The results of the Monte Carlo simulations are presented in Tables 1–6. We report the empirical rejection frequencies (power) for the Markov, Weibull, and EACD independence tests for various significance test levels, VaR coverage rates, and backtesting sample sizes. Table 1 reports power for an HS risk model with $T_e = 500$ observations in the rolling estimation samples. Table 2 gives the sample selection frequencies, that is, the fraction of samples drawn that were possible to use for calculating the tests. Table 3 reports effective power, which is simply the power entries from Table 1 multiplied by the relevant sample selection frequency in Table 2. Tables 4–6 show the results when the rolling samples for VaR calculation contain $T_e = 250$ observations. Notice that we focus solely on the independence tests here because the HS risk models under study are correctly specified unconditionally.
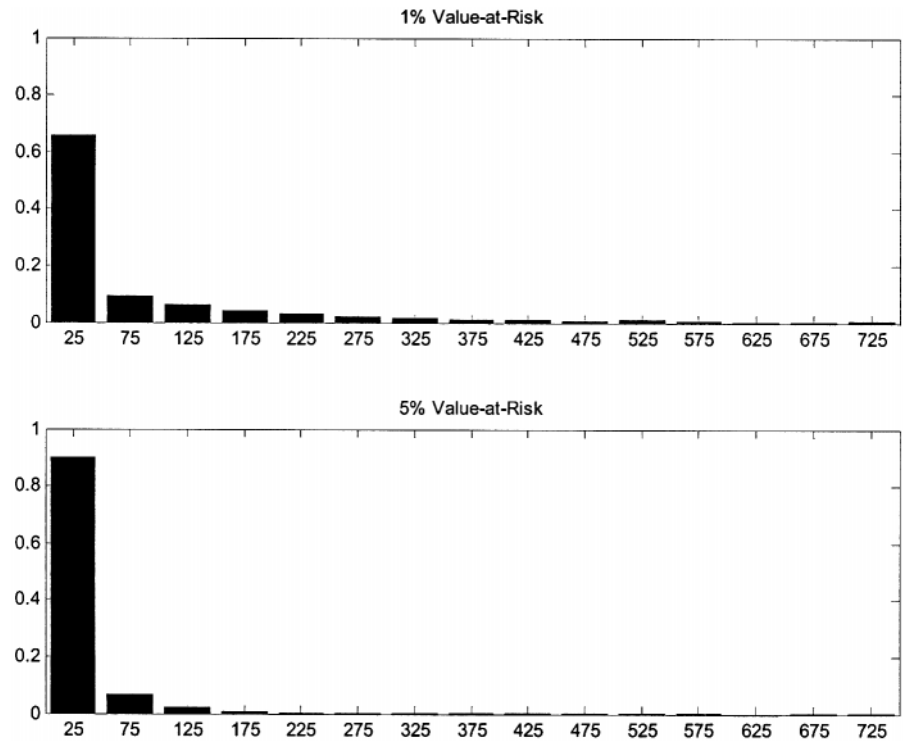
Figure 5 Histograms of duration between VaR violations. HS risk model on GARCH-$t(d)$ portfolio returns with $T_e = 500$.

The results are quite striking. The main result in Table 1 is that for inference samples of 750 days and above, the Weibull tests are always more powerful than the Markov and EACD tests in rejecting the HS risk models. This result holds across inference sample sizes, VaR coverage rates, and significance levels chosen. The differences in power are sometimes very large. For example, in Table 1, using a 1% significance level, the 5% VaR in a sample of 1250 observations has a Weibull rejection frequency of 69.2% and a Markov rejection frequency of only 39.5%. The Weibull test clearly appears to pick up dependence in the hit violations, which is ignored by the Markov test.

For an inference sample size of 500, the ranking of tests depends on the inference sample size, VaR coverage rate, and significance level in question. Typically either the Markov or the EACD test performs the best.

For an inference sample size of 250, the power is typically very low in any of the three tests. This is a serious issue, as the backtesting guide for market risk capital requirements uses a sample size of one year when assessing model adequacy.[12] The EACD test is often the most powerful in the case of 250 inference

---

[12] We thank an anonymous referee for pointing out this important issue.

Table 1 Power of independence tests: HS VaR calculated on 500 GARCH(1,1)-$t(d)$ returns.

	Significance level: 1%				Significance level: 5%				Significance level: 10%		
	Coverage rate: 1%				Coverage rate: 1%				Coverage rate: 1%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.060	0.018	0.150	250	0.263	0.104	0.234	250	0.330	0.195	0.278
500	0.105	0.114	0.164	500	0.307	0.267	0.250	500	0.370	0.369	0.303
750	0.157	0.236	0.167	750	0.290	0.415	0.251	750	0.435	0.536	0.311
1000	0.224	0.378	0.159	1000	0.360	0.546	0.253	1000	0.523	0.648	0.303
1250	0.266	0.484	0.145	1250	0.382	0.674	0.237	1250	0.514	0.758	0.291
1500	0.308	0.596	0.132	1500	0.427	0.752	0.222	1500	0.543	0.820	0.271
	Coverage rate: 5%				Coverage rate: 5%				Coverage rate: 5%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.107	0.052	0.159	250	0.205	0.152	0.273	250	0.257	0.235	0.342
500	0.215	0.238	0.324	500	0.296	0.403	0.440	500	0.351	0.509	0.504
750	0.271	0.413	0.389	750	0.367	0.607	0.501	750	0.429	0.706	0.563
1000	0.339	0.546	0.440	1000	0.443	0.734	0.555	1000	0.533	0.810	0.615
1250	0.395	0.692	0.493	1250	0.530	0.833	0.601	1250	0.654	0.895	0.661
1500	0.434	0.750	0.514	1500	0.627	0.882	0.638	1500	0.735	0.927	0.700

Table 2 Sample selection frequency: HS VaR calculated on 500 GARCH(1,1)-$t(d)$ returns.

| | Coverage rate: 1% | | | | Coverage rate: 5% | | |
| | Test | | | | Test | | |
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.778	0.589	0.598	250	0.987	0.972	0.974
500	0.956	0.891	0.896	500	1.000	1.000	0.999
750	0.998	0.987	0.986	750	1.000	1.000	1.000
1000	1.000	0.999	0.997	1000	1.000	1.000	1.000
1250	1.000	1.000	1.000	1250	1.000	1.000	1.000
1500	1.000	1.000	1.000	1500	1.000	1.000	1.000

observations, which is curious because the performance of the EACD test is quite sporadic for larger sample sizes. Generally the EACD test appears to do quite well at smaller sample sizes, but relatively poorly at larger sample sizes. We suspect that the nonlinear estimate of the α parameter is poorly behaved in this application.

Table 2 shows the sample selection frequencies corresponding to the power calculations in Table 1. As expected, the sample rejection issue is the most serious for inference samples of 250 observations. For inference samples of 500 or more, virtually no samples are rejected.

Table 3 reports the effective power, calculated as the power in Table 1 multiplied by the relevant sample selection frequency in Table 2. Comparing Tables 1 and 3, it is clear that the test that has the highest power in any given case in Table 1, also has the highest power in Table 3. But the levels of power are of course lower in Table 3 compared with Table 1, but only dramatically so for inference samples of 250 observations.

Table 4 shows the power calculations for the case when the VaR is calculated on 250 in-sample observations rather than 500, as was the case in Tables 1–3. The overall picture from Table 1 emerges again: The Weibull test is always best for inference samples of 750 observations or more. For samples of 500 observations, the rankings vary case by case, and for 250 observations, the power is generally very low.

Table 5 reports the sample selection frequencies corresponding to Table 4. In this case the sample selection frequencies are even higher than in Table 2. For a VaR coverage rate of 5% the rejection frequencies are negligible for all sample sizes.

Table 6 shows the effective power from Table 4. Again we simply multiply the power in Table 4 by the sample selection frequency in Table 5. Notice again that the most powerful test in Table 4 is also the most powerful test in Table 6. Notice also that for most entries the power numbers in Table 6 are very similar to those in Table 4.

Table 3 Effective power of independence tests: HS VaR calculated on 500 GARCH(1,1)-$t(d)$ returns.

	Significance level: 1%				Significance level: 5%				Significance level: 10%		
	Coverage rate: 1%				Coverage rate: 1%				Coverage rate: 1%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.047	0.011	0.090	250	0.205	0.061	0.140	250	0.257	0.115	0.167
500	0.100	0.101	0.147	500	0.294	0.238	0.224	500	0.353	0.329	0.271
750	0.156	0.233	0.164	750	0.289	0.410	0.247	750	0.434	0.529	0.307
1000	0.224	0.378	0.158	1000	0.360	0.545	0.253	1000	0.522	0.647	0.302
1250	0.266	0.484	0.145	1250	0.382	0.674	0.237	1250	0.514	0.758	0.291
1500	0.308	0.596	0.132	1500	0.427	0.752	0.222	1500	0.543	0.820	0.271
	Coverage rate: 5%				Coverage rate: 5%				Coverage rate: 5%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.105	0.051	0.154	250	0.203	0.148	0.266	250	0.253	0.229	0.333
500	0.215	0.238	0.324	500	0.296	0.403	0.440	500	0.351	0.509	0.504
750	0.271	0.413	0.389	750	0.367	0.607	0.501	750	0.429	0.706	0.563
1000	0.339	0.546	0.440	1000	0.443	0.734	0.555	1000	0.533	0.810	0.615
1250	0.395	0.692	0.493	1250	0.530	0.833	0.601	1250	0.654	0.895	0.661
1500	0.434	0.750	0.514	1500	0.627	0.882	0.638	1500	0.735	0.927	0.700

Table 4 Power of independence tests: HS VaR calculated on 250 GARCH(1,1)-$t(d)$ returns.

	Significance level: 1%				Significance level: 5%				Significance level: 10%		
	Coverage rate: 1%				Coverage rate: 1%				Coverage rate: 1%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.059	0.005	0.114	250	0.217	0.072	0.195	250	0.285	0.166	0.251
500	0.079	0.053	0.098	500	0.278	0.196	0.183	500	0.336	0.304	0.236
750	0.108	0.133	0.069	750	0.254	0.313	0.132	750	0.401	0.437	0.182
1000	0.153	0.222	0.045	1000	0.290	0.406	0.105	1000	0.467	0.535	0.149
1250	0.203	0.310	0.035	1250	0.305	0.536	0.084	1250	0.463	0.645	0.123
1500	0.230	0.420	0.029	1500	0.321	0.634	0.070	1500	0.459	0.736	0.101
	Coverage rate: 5%				Coverage rate: 5%				Coverage rate: 5%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.115	0.068	0.189	250	0.210	0.169	0.311	250	0.266	0.250	0.380
500	0.212	0.247	0.288	500	0.295	0.421	0.408	500	0.354	0.530	0.475
750	0.244	0.388	0.346	750	0.346	0.603	0.456	750	0.419	0.700	0.517
1000	0.299	0.500	0.345	1000	0.413	0.707	0.480	1000	0.507	0.790	0.553
1250	0.344	0.622	0.394	1250	0.499	0.796	0.497	1250	0.631	0.862	0.569
1500	0.385	0.695	0.393	1500	0.582	0.849	0.537	1500	0.688	0.896	0.606

Table 5 Sample selection frequency: HS VaR calculated on 250 GARCH(1,1)-$t(d)$ returns.

	Coverage rate: 1%				Coverage rate: 5%		
	Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.877	0.695	0.706	250	0.997	0.993	0.993
500	0.994	0.975	0.976	500	1.000	1.000	1.000
750	1.000	0.999	0.999	750	1.000	1.000	1.000
1000	1.000	1.000	1.000	1000	1.000	1.000	1.000
1250	1.000	1.000	1.000	1250	1.000	1.000	1.000
1500	1.000	1.000	1.000	1500	1.000	1.000	1.000

Comparing numbers across Tables 1 and 4 and across Tables 3 and 6, we note that the HS VaR with $T_e = 500$ rolling sample observations often has a higher rejection frequency than the HS VaR with $T_e = 250$ rolling sample observations. This result is interesting because practitioners often work very hard to expand their databases, enabling them to increase their rolling estimation sample period. Our results suggest that such efforts may be misguided because lengthening the size of the rolling sample does not necessarily eliminate the distributional problems with HS.

5   BACKTESTING TAIL DENSITY FORECASTS

The choice of VaR as a portfolio risk measure can be criticized on several fronts. Most importantly, the quantile nature of the VaR implies that the shape of the return distribution to the left of the VaR is ignored. Particularly in portfolios with highly nonlinear distributions, such as those including options, this shortcoming can be crucial. Theoreticians have criticized the VaR measure both from a utility-theoretic perspective [Artzner et al. (1999)] and from a dynamic trading perspective [Basak and Shapiro (2000)]. Although some of these criticisms have recently been challenged [Cuoco, He, and Issaenko (2001)], it is safe to say that risk managers ought to be interested in knowing the entire distribution of returns, and in particular the left tail. Backtesting distributions rather than VaRs then becomes important.

Consider the standard density forecast evaluation approach [see, e.g., Diebold, Gunther, and Tay (1998)] of calculating the uniform transform variable

$$U_t = F_t(R_t),$$

where $F_t(*)$ is the a priori density forecast for time t. The null hypothesis that the density forecast is optimal corresponds to

$$U_t \sim i.i.d. \ Uniform(0,1).$$

Table 6 Effective power of independence tests: HS VaR calculated on 250 GARCH(1,1)-$t(d)$ returns.

	Significance level: 1%				Significance level: 5%				Significance level: 10%		
	Coverage rate: 1%				Coverage rate: 1%				Coverage rate: 1%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.051	0.004	0.080	250	0.190	0.050	0.138	250	0.250	0.116	0.177
500	0.078	0.052	0.095	500	0.276	0.191	0.179	500	0.334	0.296	0.230
750	0.108	0.133	0.069	750	0.254	0.313	0.132	750	0.401	0.436	0.181
1000	0.153	0.222	0.045	1000	0.290	0.406	0.105	1000	0.467	0.535	0.149
1250	0.203	0.310	0.035	1250	0.305	0.536	0.084	1250	0.463	0.645	0.123
1500	0.230	0.420	0.029	1500	0.321	0.634	0.070	1500	0.459	0.736	0.101
	Coverage rate: 5%				Coverage rate: 5%				Coverage rate: 5%		
	Test				Test				Test		
Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD	Sample size	Markov	Weibull	EACD
250	0.115	0.068	0.187	250	0.209	0.167	0.308	250	0.266	0.248	0.378
500	0.212	0.247	0.288	500	0.295	0.421	0.408	500	0.354	0.530	0.475
750	0.244	0.388	0.346	750	0.346	0.603	0.456	750	0.419	0.700	0.517
1000	0.299	0.500	0.345	1000	0.413	0.707	0.480	1000	0.507	0.790	0.553
1250	0.344	0.622	0.394	1250	0.499	0.796	0.497	1250	0.631	0.862	0.569
1500	0.385	0.695	0.393	1500	0.582	0.849	0.537	1500	0.688	0.896	0.606

Berkowitz (2001) argues that the bounded support of the uniform variable renders standard inference difficult. One is forced to rely on nonparametric tests, which have notoriously poor small sample properties. He suggests a simple transformation using the inverse normal cumulative density function (c.d.f.)

$$Z_t = \Phi^{-1}(U_t),$$

after which the hypothesis

$$Z_t \sim i.i.d. \ Normal(0, 1)$$

can easily be tested.

Berkowitz further argues that confining attention to the left tail of the distribution has particular merit in the backtesting of risk models where the left tail contains the largest losses that are most likely to impose bankruptcy risk. He defines the censored variable

$$Z_t^* = \begin{cases} Z_t, & \text{if } R_t < VaR_t \\ \Phi^{-1}(VaR_t), & \text{else} \end{cases}$$

and tests the null that

$$Z_t^* \sim Censored \ Normal(0, 1, VaR_t).$$

We note first that Berkowitz (2001) only tests the unconditional distribution of Z_t^*. The information in the potential clustering of the VaR exceedences is ignored.

Second, note that the censored variable complication is not needed. If we want to test that the transforms of the $100p$ percent largest losses are themselves uniform, then we can simply multiply the subset of the uniform by $1/p$, apply the transformation, and test for standard normality again.[13] That is,

$$U_i^{**} = \begin{cases} U_t/p, \text{if } R_t < VaR_t \\ \text{Else not defined.} \end{cases}$$

We then have that

$$Z_i^{**} = \Phi^{-1}(U_i^{**}) \sim i.i.d. \ Normal(0, 1).$$

Note that due to censoring there is no notion of time in the sequence Z_i^{**}. We might want to make a joint analysis of both Z_i^{**} and the duration between violations D_i. To do this we would like to write a joint density for these two processes under the alternative. We know that under the null hypothesis the risk model is correctly specified, the Z_i^{**} should be i.i.d. $N(0, 1)$, D_i should be i.i.d. exponential with mean $1/p$, and the processes should be independent. The question is how to write a joint density for these two processes as the alternative hypothesis knowing that, for example, the marginal p.d.f. of D_i is a Weibull and some other p.d.f. for Z_i^{**}. Copulas provide a useful tool for doing so.

---

[13] We are grateful to Nour Meddahi for pointing this out.

A (bivariate) copula is a function C from $[0; 1] \times [0; 1]$ to $[0; 1]$ with the following properties:

1. For every u, v in $[0; 1]$,

    $C(u, 0) = 0 = C(0, v)$

    and

    $C(u, 1) = u \quad \text{and} \quad C(1, v) = v.$

2. For every u_1, u_2, v_1, v_2 in $[0; 1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$

    $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$

In order to explain how copulas can be used we apply Sklar's theorem [Nelsen (1998)], which states: Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all x, y in $\mathbb{R}$,

$H(x, y) = C(F(x), G(y)).$

If F and G are continuous then C is unique. Conversely, if C is a copula and F and G are distribution functions, then H is a joint distribution function with marginal densities F and G.

So if we have two densities under the alternative (e.g., $f(D_i)$ and $g(Z_i^{**})$), then we can easily construct a joint density by applying a copula. Suppose the considered bivariate copula $C(u, v; \theta)$ is a function of a unique parameter θ and that we have $C(u, v; \theta_0) = uv$ and $C(u, v; \theta) \neq uv$ for $\theta \neq \theta_0$. This gives us a basis for a test because $C(F(x), G(y); \theta_0) = F(x)G(y)$ means that x and y are independent.

An example of such a copula is the Ali-Mikhail-Haq family of copulas where

$$C(u, v; \theta) = \frac{uv}{1 - \theta(1 - u)(1 - v)}; \quad \theta \in [-1, 1]$$

and we have $C(u, v; \theta) = uv$ if $\theta = 0$. A possible alternative hypothesis could be that D_i is i.i.d. Weibull(a, b), Z_i^{**} is i.i.d. $N(\mu, \sigma^2)$, and $C(u, v; \theta)$ is from the Ali-Mikhail-Haq family of copulas. We could then test

$H_0 : a = p, b = 1, \mu = 0, \sigma = 1, \theta = 0$

$H_1 :$ at least one of these equalities does not hold

in a likelihood ratio framework similar to the one considered for the VaR tests above. Another useful approach could be the graphical procedure proposed by Fermanian and Scaillet (2003). We plan to pursue the implementation of this procedure in future work.

6   CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

We have presented a new set of procedures for backtesting risk models. The chief insight is that if the one-day VaR model is correctly specified for coverage rate, p,

then, every day, the conditional expected duration until the next violation should be a constant $1/p$ days. We suggest various ways of testing this null hypothesis and we conduct a Monte Carlo analysis that compares the new tests to those currently available. Our results show that in many of the situations we consider, the duration-based tests have much better power properties than the previously suggested tests. The size of the tests is easily controlled through finite-sample p-values, which we calculate using Monte Carlo simulation.

The majority of financial institutions use VaR as a risk measure, and many calculate VaR using the so-called HS approach. While the main focus of our article has thus been backtesting VaRs from HS, we also suggest extensions to density and density tail backtesting.

The immediate potential extensions to our Monte Carlo results are several. First, it may be interesting to calculate the power of the tests with different GARCH specifications using, for example, Engle and Lee (1999) and Hansen (1994). Second, we could consider structural breaks in the underlying return models, such as those investigated by Andreou and Ghysels (2002). Finally, Hamilton and Jorda (2002) recently introduced a class of dynamic hazard models. Exploring these for the purpose of backtesting could be interesting.

We could also consider more complicated portfolios, including options and other derivatives. Examining the duration patterns from misspecified risk models in this case could suggest other alternative hypotheses than the ones suggested here. We leave these extensions for future work. Finally, we stress that the current regulator practice of requiring backtesting on samples of only 250 daily observations is likely to prove futile, as the power to reject misspecified risk models is very low in this case.

REFERENCES

Andreou, E., and E. Ghysels. (2002). "Quality Control for Value at Risk: Monitoring Disruptions in the Distribution of Risk Exposure." Working paper, University of North Carolina.

Andrews, D. (2001). "Testing When a Parameter is on the Boundary of the Maintained Hypothesis." *Econometrica 69*, 683–734.

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

Barone-Adesi, G. K., K. Giannopoulos, and L. Vosper. (2002). "Backtesting Derivative Portfolios with Filtered Historical Stimulation (FHS)." *European Financial Management 8*, 31–58.

Basak, S., and A. Shapiro. (2000). "Value at Risk Based Risk Management: Optimal Policies and Asset Prices." *Review of Financial Studies 14*, 371–405.

Basle Committee on Banking Supervision. (1996). "Amendment to the Capital Accord to Incorporate Market Risks." Basle, Switzerland: BIS.

Beder, T. (1995). "VaR: Seductive but Dangerous." *Financial Analysts Journal 51*, 5, 12–24.

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

Berkowitz, J. and J. O'Brien. (2002). "How Accurate are the Value-at-Risk Models at Commercial Banks" *Journal of Finance 57*, 1093–1112.

Bollerslev, T. (1987). "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics 69*, 542–547.

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

Christoffersen, P. (2003). *Elements of Financial Risk Management*. San Diego, CA: Academic Press.

Christoffersen, P., J. Hahn, and A. Inoue. (2001). "Testing and Comparing Value-at-Risk Measures." *Journal of Empirical Finance 8*, 325–342.

Cuoco, D., H. He, and S. Issaenko. (2001). "Optimal Dynamic Trading Strategies with Risk Limits." Working paper, Yale University.

Diebold, F.X., T. Gunther, and A. Tay. (1998). "Evaluating Density Forecasts, with Applications to Financial Risk Management." *International Economic Review 39*, 863–883.

Dufour, J.-M. (2000). "Monte Carlo Tests with Nuisance Parameters. A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics." Working paper, Université de Montréal.

Engle, R., and G. J. Lee. (1999). "A Permanent and Transitory Component Model of Stock Return Volatility." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*. Oxford: Oxford University Press.

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

Fermanian, J.-D., and O. Scaillet. (2003). "Nonparametric Estimation of Copulas for Time Series." *Journal of Risk 5*, 25–54.

Gourieroux, C. (2000). *Econometrics of Qualitative Dependent Variables*, Paul B. Klassen[trans.]. Cambridge: Cambridge University Press.

Hamilton, J., and O. Jorda. (2002). "A Model of the Federal Funds Rate Target." *Journal of Political Economy 110*, 1135–1167.

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

Hendricks, D. (1996). "Evaluation of Value-at-Risk Models Using Historical Data." In *Economic Policy Review*, Federal Reserve Bank of New York, April, 39–69.

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.

Jorion, P. (2000). *Value-at-Risk: The New Benchmark for Controlling Financial Risk*. Chicago: McGraw-Hill.

Marshall, C., and M. Siegel. (1997). "Value at Risk: Implementing a Risk Measurement Standard." *Journal of Derivatives 4*, 91–110.

Nelsen, R. (1998). "An Introduction to Copulas." In *Lectures Notes in Statistics 139*. Berlin: Springer Verlag.

Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Pritsker, M. (1997). "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time." *Journal of Financial Services Research 12*, 201–241.

Pritsker, M. (2001). "The Hidden Dangers of Historical Simulation." Working Paper 2001-27, Board of Governors of the Federal Reserve System.