NTNU – Trondheim
Norwegian University of
Science and Technology

Evaluating power of Value-at-Risk backtests

Torgeir Røynstrand
Nils Petter Nordbø
Vidar Kristoffer Strat

Industrial Economics and Technology Management
Submission date: June 2012
Supervisor: Sjur Westgaard, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Preface

This Master's thesis concludes the authors' Master of Science degrees in Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU), in the Spring of 2012. The thesis is written by Nils Petter Nordbø, Torgeir Røynstrand and Vidar Kristoffer Strat, under the guidance of Dr. Sjur Westgaard.

The thesis evaluates the power of the most recognized Value-at-Risk backtests. It is written as an academic research paper in article format and is intended for publishing.

Value-at-Risk, as a risk measure, has received great attention, and is widely debated. Our thesis will neither discuss weaknesses or advantages with this risk measure, nor offer any comparison against other measures, such as Expected Shortfall. The objective of the thesis is to contribute to the understanding of Value-at-Risk backtesting, show the limitations of today's practice and identify best practice.

First and foremost, we would like to thank our supervisor, Dr. Sjur Westgaard.

We thank Dr. Carol Alexander (Henley Business School, Reading), Dr. Peter Christoffersen (McGill University, Montreal), Dr. Denis Pelletier (North Carolina State University, Raleigh), Dr. Simone Manganelli (European Central Bank, Frankfurt), Dr. Oliver Linton (London School of Economics, London) and Dr. Mette Langaas (NTNU, Trondheim) for help on technical details along the way.

We would also like to thank Carsten Buch Sivertsen (Junior Partner, McKinsey & Company, Oslo), Dr. Stein-Erik Fleten (NTNU, Trondheim), Dr. Peter Molnar (NTNU, Trondheim), Daniel Haugstvedt (NTNU, Trondheim), Fritts Causby, Svein Arne Nordbø, Endre Røynstrand, Torgeir Røynstrand Sr., Jarand Røynstrand, Lisa St. Dennis, Kristen Strat and Torgeir Strat for read through and constructive criticism.

Trondheim, June 5, 2012

Norwegian Abstract

Value-at-Risk-modeller (VaR-modeller) gir kvantil-prognoser for fremtidige avkastninger. Dersom et realisert tap er større enn, eller lik, sin tilsvarende VaR-prognose, får vi et brudd. En VaR-modell er vanligvis validert ved å vurdere realiserte bruddsekvenser. Det er utviklet flere statistiske tester for dette formålet, kalt backtester. Denne artikkelen presenterer en omfattende styrkestudie av de mest anerkjente backtestene. Vi simulerer avkastningsserier og estimerer VaR-prognoser, slik at de resulterende bruddsekvensene ikke tilfredsstiller nullhypotesen til backtestene. Deretter benytter vi backtestene på disse sekvensene og undersøker deres evne til å forkaste feilaktig spesifiserte VaR-modeller. Den betingede dekningstesten Geometric, av Berkowitz et al. (2011), presterer best. Det trengs et minimum av datapunkter for å gjøre inferens med tilfredsstillende styrke. Et utvalg på 250 datapunkter, som er minstekravet satt av Basel Commitee on Banking Supervision (2011), vil ikke være tilstrekkelig. Den vanlig implementasjonen av den populære Dynamic Quantile backtesten, av Engle og Manganelli (2004), har for høy forkastningsrate for korrekt spesifiserte VaR-modeller.

Evaluating the Power of Value-at-Risk Backtests

Nils Petter Nordbø[a], Torgeir Røynstrand[a], Vidar Kristoffer Strat[a]

[a]*Department of Industrial Economics and Technology Management, Alfred Getz v. 1, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

Abstract

Value-at-Risk (VaR) models provide quantile forecasts for future returns. If a loss is greater than or equal to the corresponding VaR forecast, we have a breach. A VaR model is usually validated by considering realized breach sequences. Several statistical tests exist for this purpose, called backtests. This paper presents an extensive study of the statistical power for the most recognized backtests. We simulate returns and estimate VaR forecasts, resulting in breach sequences not satisfying the null hypothesis of the backtests. We apply the backtests on the data, and assess their ability to reject misspecified models. The Geometric conditional coverage test by Berkowitz et al. (2011) performs best. A minimum amount of observations is needed to make inference with satisfying power. A sample size of 250 data points, which is the minimum requirement set by the Basel Committe on Banking Supervision (2011), is not sufficient. The common implementation of the Dynamic Quantile test, by Engle and Manganelli (2004), has a too high rejection rate for correctly specified VaR models.

Keywords: Backtesting, Risk management, Value-at-Risk

1. Introduction

Value-at-Risk (VaR) gained increased popularity through the 1990s among financial institutions and later also non-financial firms. The successful introduction of RiskMetrics^{TM} by J. P. Morgan (1996), and the recognition of VaR as a regulatory tool[1], made it the standard risk measure. VaR is defined as the threshold value which loss will exceed with a given probability. A mathematical definition of VaR is given in Appendix A.

Currently used VaR models have several limitations, e.g. clustering of breaches, as illustrated by the frequent VaR breaches by financial firms during recent periods with high volatility. The most noticeable being the 2000–2001 dot-com bubble, the 2007–2011 financial crisis and the 2010 European sovereign debt crisis. Even though accurate VaR estimates do not prevent losses from happening, they can provide management with an understanding of current risks and assistance with the allocation of capital.

The extensive use of VaR, both for internal and external purposes, drives the demand for proper VaR models suitable for different types of markets. The models are evaluated by the accuracy of their forecasts. This is done by comparing each realized return to the corresponding VaR forecast. Whenever the loss exceeds the VaR, we have a breach or hit. This hit sequence should have a proportion of hits in line with the chosen target probability of the VaR model, and the hits should be inde-

[1]For the regulatory history, see Basel Committe on Banking Supervision (1996a,b, 2006, 2011)

pendent of past information. Statistical procedures testing for these properties are called backtests, and are quite numerous in the literature. This paper provides an extensive study of the power properties of the most recognized backtests.

Most backtests yield a test statistic that asymptotically follows a known probability distribution under the null hypothesis. However, the statistic does not necessarily follow this distribution for finite samples. Thus, we evaluate if the asymptotic distributions are appropriate to make inference on finite samples. We show that the test sizes for finite samples differ from the significance level of the tests. Using asymptotic critical values can thus lead to erroneous conclusions for finite samples used in practice. One should therefore always use finite sample distributions, as described by Dufour (2006), when backtesting.

The power study is introduced by assessing the power of the backtests to reject a hit process with a probability of breach different from the one tested for. We start by doing this from a solely theoretical point of view and then show the implication of estimation. We find evidence that the common implementation of the Dynamic Quantile test, by Engle and Manganelli (2004), has a too high rejection rate for correctly specified VaR models.

We expand the power study by including dependence in the breach series. We find that the Geometric conditional coverage test, by Berkowitz et al. (2011), performs best overall. To get a satisfactory power when using this test, we identify sample sizes of 1,000, 750 and 500 data points as lower limits when testing 1%, 5% and 10% VaR, respectively. This implies that backtesting using one year of data with daily observations, which is the minimum requirement set by Basel Committe on Banking Supervision (2011), will have too low power against misspecified VaR models.

The paper is organized as follows: Section 2 includes a review of the existing literature on backtesting and power studies, Section 3 gives a theoretical overview of the considered backtests, Section 4 shows the finite sample distribution methodology, Section 5 gives a detailed description of the experiments, shows the results and gives a discussion on these, and Section 6 concludes.

2. Relevant literature

When VaR became widespread in the late 1990s, a large literature on calculating interval forecasts already existed, as Chatfield (1993) summarizes. However, few tools for evaluating their performance were available. Kupiec (1995) points out the importance of assessing and quantifying the accuracy of VaR estimates, and develops a backtesting framework to test for correct number of breaches. Backtests only considering correct number of breaches, such as Kupiec (1995), are in the literature referred to as unconditional coverage tests. Christoffersen (1998) argues that the breaches would also need to be independent to validate a model. He formalizes an out-of-sample criterion which states that the probability of breach must be constant and equal to a desired level, conditional on all past information. In his paper, he introduces a simple test using a first-order Markov chain to test for violation of the criterion. Backtests considering only higher-order dynamics, such as clustered breaches, are referred to as independence tests. Backtests considering both properties are called conditional coverage tests.

In light of the recent financial crises, VaR models producing VaR estimates giving independent and correct number of breaches have been given increased focus. Conditional coverage tests are needed to validate such models, and several tests have been suggested to improve existing ones.

Christoffersen and Pelletier (2004) criticize the first-order Markov test of having too low power against general forms of dependence. They suggest a duration-based approach that

uses the time between breaches to test conditional coverage for an infinite number of lags. The first duration-based test suggested by Christoffersen and Pelletier (2004) fits the discrete durations to a continuous distribution, producing an erroneous null hypothesis.[2] Later, improvements have been suggested, by Haas (2005) and Berkowitz et al. (2011), using discrete distributions and correctly specified null hypothesis, yielding higher power.

The literature also provides regression-based tests such as the Dynamic Quantile test by Engle and Manganelli (2004). The Dynamic Quantile test utilizes the criterion by Christoffersen (1998) that the probability of breach must be independent of all past information, and allows us to test any variable in the information set. Unlike other tests, such as the first-order Markov test, past information is not restricted to be a binary variable representing whether last day was a breach or not. Christoffersen (1998) suggests a similar test in his paper, called the forecast efficiency test, using the J-test from the GMM framework by Hansen (1982). However, this test has not been given much attention in the literature.

Clements and Taylor (2003) criticize the Dynamic Quantile test, as the linear regression cannot be estimated efficiently with binary variables. They suggest a modified version of the Dynamic Quantile test by using a logistic transformation of the dependent variable.

With several backtests, there is a need for studies comparing and evaluating them. Christoffersen and Pelletier (2004) show that the duration-based approach gives higher power than the Markov test. They use an asymmetric GARCH(1,1)-t model with fixed parameters as underlying return process, and estimate VaR with a Historical Simulation VaR model. Using the same setup and parameters, Haas (2005) shows that his alternative parameterization of the duration-based test outperforms the continuous version by Christoffersen and Pelletier (2004). To our knowledge, only Berkowitz et al. (2011) have done a power study comparing a wide range of the most recognized backtests. They use an asymmetric GARCH(1,1)-t model with four sets of parameters as underlying return processes, and find the Dynamic Quantile test to perform best.

All power studies mentioned use a rolling window to estimate VaR. This study contributes by showing that the use of a rolling window affects the power of the backtests, mainly the Dynamic Quantile test. We estimate VaR using a Normal VaR model, instead of a Historical Simulation VaR model, as a parametric model gives a better foundation to evaluate backtests from a theoretical point of view. We also evaluate a larger set of alternative hypothesis than have been done in earlier power studies.

3. Backtests applied in the power study

This section provides an introduction to the backtests considered in this paper. All tests are model independent, i.e. they can be used without knowing the underlying VaR model.

A realized return less than the negative VaR causes a breach or hit. We define the hit function as

$$I_t = \begin{cases} 1 & \text{if } r_t < -\text{VaR}_t(p) \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where r_t is the return at day t and $\text{VaR}_t(p)$ is the VaR forecast for day t, with target probability p. As VaR_t is a forecast for day t, it is Ω_{t-1}-measurable, where Ω_{t-1} is the information set at day $t-1$. Christoffersen (1998) states that the hit sequence should satisfy

$$\Pr(I_t \mid \Omega_{t-1}) = p, \text{ for all } t \tag{2}$$

Equation (2) is the basis for three different kinds of tests: unconditional coverage (UC), independence (Ind) and conditional coverage (CC). Unconditional coverage tests assess

[2] Christoffersen and Pelletier (2004) account for the discreteness bias by using the Monte Carlo testing technique, as described by Dufour (2006).

whether the observed number of hits is in line with the expected number, $E\left[\sum_{t=1}^{n} I_t\right] = np$. Independence tests test whether the probability of breach is the same each day, regardless of previous outcomes, $\Pr(I_t \mid \Omega_{t-1}) = \Pr(I_{t+1} \mid \Omega_t)$ for all t. Conditional coverage tests test whether the probability of breach is a constant given value each day, $\Pr(I_t \mid \Omega_{t-1}) = p$, for all t. Note that the property in the conditional coverage tests equal Equation (2).

The power study will focus on conditional coverage tests, though we will for benchmarking purposes include the PF (proportion of failures) test by Kupiec (1995), which is an unconditional coverage test. We also include an overview of the independence tests for completeness.

3.1. Kupiec test for unconditional coverage

Kupiec (1995) notes that the probability of observing n_1 breaches, regardless of sample size n, should be binomially distributed. He derives the following likelihood

$$L\left(\pi; I_t, I_{t-1}, \dots\right) = \pi^{n_1}\left(1-\pi\right)^{n-n_1}, \qquad (3)$$

where π is the probability of breach. The maximum likelihood (ML) estimate is then $\hat{\pi} = n_1/n$. The log-likelihood statistic becomes

$$\mathrm{LR}_{\mathrm{UC}} = -2\ln\left[\frac{L\left(\pi; I_t, I_{t-1}, \dots\right)}{L\left(n_1/n; I_t, I_{t-1}\dots\right)}\right] \qquad (4)$$

We can test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} H_{0,\mathrm{UC}}&\colon\ \pi = p \\ H_{1,\mathrm{UC}}&\colon\ \pi \neq p \end{aligned}$$

Under the null hypothesis, $\mathrm{LR}_{\mathrm{UC}}$ will be asymptotically $\chi^2(1)$ distributed. We will refer to this test as PF.

3.2. Autocorrelation tests for independence and conditional coverage

3.2.1. First order Markov chain

Christoffersen (1998) notes that the hit sequence should be a Bernoulli process with mean p. He applies a first-order Markov process and estimates the one-step-ahead transition probabilities $\Pr(I_{t+1} \mid I_t)$, given by

$$\mathbf{\Pi}_1 = \begin{bmatrix} \pi_{0,0} & \pi_{0,1} \\ \pi_{1,0} & \pi_{1,1} \end{bmatrix} = \begin{bmatrix} 1-\pi_{0,1} & \pi_{0,1} \\ 1-\pi_{1,1} & \pi_{1,1} \end{bmatrix}, \qquad (5)$$

where $\pi_{i,j}$ is the transition $\Pr(I_{t+1} = j \mid I_t = i)$. The likelihood function of this process is approximately[3] given by

$$\begin{aligned} &L(\mathbf{\Pi}_1; I_1, I_2, \dots, I_T) \\ &\quad = (1-\pi_{0,1})^{n_{0,0}}\pi_{0,1}^{n_{0,1}}(1-\pi_{1,1})^{n_{1,0}}\pi_{1,1}^{n_{1,1}}, \end{aligned} \qquad (6)$$

where $n_{i,j}$ is the number of observations with value i followed by j. Maximizing Equation (6) gives the following ML estimates

$$\widehat{\mathbf{\Pi}}_1 = \begin{bmatrix} \frac{n_{0,0}}{n_{0,0}+n_{0,1}} & \frac{n_{0,1}}{n_{0,0}+n_{0,1}} \\ \frac{n_{1,0}}{n_{0,0}+n_{0,1}} & \frac{n_{1,1}}{n_{0,0}+n_{0,1}} \end{bmatrix} \qquad (7)$$

The log-likelihood statistic becomes

$$\mathrm{LR} = -2\ln\left[\frac{L\left(\mathbf{\Pi}_1; I_t, I_{t-1}, \dots\right)}{L\left(\widehat{\mathbf{\Pi}}_1; I_t, I_{t-1}, \dots\right)}\right] \qquad (8)$$

We test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} H_{0,\mathrm{Ind}}&\colon\ \pi_{1,0} = \pi_{1,1} \\ H_{1,\mathrm{Ind}}&\colon\ \pi_{1,0} \neq \pi_{1,1} \\ H_{0,CC}&\colon\ \pi_{1,0} = p \text{ and } \pi_{1,1} = p \\ H_{1,CC}&\colon\ \pi_{1,0} \neq p \text{ or } \pi_{1,1} \neq p \end{aligned}$$

Testing for independence gives the following ML estimates $\hat{\pi}_{1,0} = \hat{\pi}_{1,1} = (n_{0,1}+n_{1,1})/(n_{0,0}+n_{1,0}+n_{0,1}+n_{1,1})$.

Under the null hypothesis, LR will be asymptotically $\chi^2(1)$ distributed for independence and $\chi^2(2)$ distributed for conditional coverage. We will refer to the conditional coverage test as Markov.

[3] The first day should be censored, but are instead omitted. The exact likelihood can be found in Christoffersen (1996).

3.2.2. Dynamic Quantile with linear regression

Engle and Manganelli (2004) suggest a backtest based on explanatory variables available in the information set. Consider the following linear model

$$I_t - p = \alpha + \sum_{i=1}^{n} \beta_{1,i} I_{t-i} + \sum_{j=1}^{m} \beta_{2,j} g(\cdot) + u_t, \quad (9)$$

where p is the target probability, u_t is the error term, $\beta_{1,i}$ and $\beta_{2,j}$ are the regression coefficients, and $g(\cdot)$ is a function from the information set Ω_{t-1}[4]. We set $g(\cdot) = \text{VaR}_{t-j}$ and $n = m = 3$ in this paper.

Equation (9) can be estimated by ordinary linear regression. Independence and conditional coverage can be tested by the following Wald statistics

$$DQ_{\text{Ind}} = \frac{\widehat{\boldsymbol{\beta}}' \mathbf{R}' \left(\mathbf{R} \left[\mathbf{X}'\mathbf{X} \right]^{-1} \mathbf{R}' \right)^{-1} \mathbf{R} \widehat{\boldsymbol{\beta}}}{p(1-p),} \quad (10)$$

$$DQ_{\text{CC}} = \frac{\widehat{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{X} \widehat{\boldsymbol{\beta}}}{p(1-p)}, \quad (11)$$

where $\widehat{\boldsymbol{\beta}} = \left(\hat{\alpha}, \hat{\beta}_{1,1}, ..., \hat{\beta}_{1,n}, \hat{\beta}_{2,1}, ..., \hat{\beta}_{2,m} \right)'$, $\mathbf{R}$ is the $1 \times (n+m+1)$ matrix $(0, 1, ..., 1)$ and $\mathbf{X}$ is the matrix of ones in the first column and lagged hit and VaR sequences in the others.

The statistics will test the following hypothesis

$$\begin{aligned} &H_{0,\text{Ind}}\text{: } \beta_{1,i} = 0 \text{ and } \beta_{2,i} = 0 \\ &H_{1,\text{Ind}}\text{: } \beta_{1,i} \neq 0 \text{ or } \beta_{2,i} \neq 0 \\ &H_{0,CC}\text{: } \beta_{1,i} = 0,\ \beta_{2,i} = 0 \text{ and } \alpha = 0 \\ &H_{1,CC}\text{: } \beta_{1,i} \neq 0 \text{ or } \beta_{2,i} \neq 0 \text{ or } \alpha \neq 0 \end{aligned}$$

Under the null hypothesis, the test statistics will be asymptotically $\chi^2(n+m)$ distributed for independence and $\chi^2(n+m+1)$ distributed for conditional coverage. We will refer to the conditional coverage test as DQ.

[4] $g(\cdot)$ may be directly related to I_t (e.g. realized returns, VaR forecasts) or indirectly (e.g. realized returns on a benchmark portfolio, value of indicators).

3.2.3. Dynamic Quantile with logistic regression

Clements and Taylor (2003) point out that as I_t is binary, the linear specification in Equation (9) cannot be estimated efficiently. We can instead estimate a logit model

$$\Pr(I_t = 1) = \left(1 + e^{-X_t}\right)^{-1}, \quad (12)$$

where $X_t = \alpha + \sum_{i=1}^{n} \beta_{1,i} I_{t-i} + \sum_{j=1}^{m} \beta_{2,j} g(\cdot)$. We set $g(\cdot) = \text{VaR}_{t-j}$ and $n = m = 3$ in this paper.

Likelihood ratio tests can then be applied for testing the following hypothesis by adding constraints under the null hypothesis, for $i = 1, \ldots, n$ and $j = 1, \ldots, n$

$$\begin{aligned} &H_{0,\text{Ind}}\text{: } \beta_{1,i} = 0 \text{ and } \beta_{2,j} = 0 \\ &H_{1,\text{Ind}}\text{: } \beta_{1,i} \neq 0 \text{ or } \beta_{2,j} \neq 0 \\ &H_{0,CC}\text{: } \beta_{1,i} = 0,\ \beta_{2,j} = 0 \\ &\qquad \text{and } \alpha = \ln(p/(1-p)) \\ &H_{1,CC}\text{: } \beta_{1,i} \neq 0 \text{ or } \beta_{2,j} \neq 0 \\ &\qquad \text{or } \alpha \neq \ln(p/(1-p)) \end{aligned}$$

Under the null hypothesis, the log-likelihood ratio statistics will be asymptotically $\chi^2(n+m)$ distributed for independence and $\chi^2(n+m+1)$ distributed for conditional coverage. We will refer to the conditional test as DQLogit. The DQ and DQLogit test together will be referred to as the regression-based tests.

3.3. Duration-based tests for independence and conditional coverage

Christoffersen and Pelletier (2004) suggest a test based on the distance between each hit. The duration of time in days between two hits is defined as

$$D_i = t_i - t_{i-1}, \quad (13)$$

where t_i is the time of hit number i. If the backtesting criterion in Equation (2) is satisfied, hits will be i.i.d. Bernoulli random variables. Hence, the durations should be memoryless. They will then follow a geometric distri-

bution with the following probability distribution function (p.d.f.)

$$\Pr(D = d) = f_{\text{geo}}(d; p) = (1-p)^{d-1}p, \quad d \in \mathbb{N}^+, \tag{14}$$

where p is the unconditional probability of breach.

The discrete hazard function is given by

$$\begin{aligned}\lambda(d) &= \Pr(D = d \mid D \geq d) \\ &= \frac{\Pr(D = d)}{\Pr(D \geq d)} = \frac{f(d)}{S(d-1)},\end{aligned} \tag{15}$$

where $S(x) = \Pr(D > x) = 1 - F(x)$ denotes the survivor function and $F(x)$ the cumulative distribution function (c.d.f.). The memory-less property of the geometric distribution implies a constant hazard function.

Several parameterizations using the duration-based approach have been suggested. They should collapse to Equation (14) under the null hypothesis, but will have different sets of alternative hypothesis depending on the parameterization. Parameters are estimated using maximum likelihood. The likelihood calculation is straightforward, except for the first and the last duration. If the first observation is not a breach, the first duration will be the number of days until first breach. For example, if the first breach happens on day 5, then $D_1 = 5 - 1 = 4$. In this case, D_1 will be right censored, and we indicate this by setting $C_F = 1$. Thus, for this observation we use the survivor function instead of the p.d.f. to find the probability of observing a duration greater than D_1. The same method is applied to the last duration. If the last observation is not a breach, the last duration will be the number of days since last breach, and the last duration will be right censored. For example, if the last breach happens on day 98, and we have 100 days in the series, $D_N = 100 - 98 = 2$. We set $C_L = 1$ to use the probability from the survivor function instead of the p.d.f.

The log-likelihood functions of the duration-based tests can generally be written as

$$\begin{aligned}&\ln L(\boldsymbol{\theta}; D_1, D_2, \ldots) \\ &\quad = C_F \ln S(D_1; \boldsymbol{\theta}) + (1 - C_F) \ln f(D_1; \boldsymbol{\theta}) \\ &\qquad + \sum_{i=2}^{N-1} \ln f(D_i; \boldsymbol{\theta}) + C_L \ln S(D_N; \boldsymbol{\theta}) \\ &\qquad + (1 - C_L) \ln f(D_N; \boldsymbol{\theta}),\end{aligned} \tag{16}$$

where $\boldsymbol{\theta}$ is a vector of parameters contained in the space of possible parameters, $\boldsymbol{\theta} \in \Theta$. The p.d.f. and the survivor function depend on the chosen parameterization. For each test we will define the p.d.f., while the survivor function can be derived from it.

The log-likelihood ratio statistics for the tests are given by

$$\begin{aligned}\text{LR} = &2 \ln L(\boldsymbol{\theta}^*; D_1, D_2, \ldots), \\ &- 2 \ln L(\boldsymbol{\theta}_0; D_1, D_2, \ldots) \\ &\boldsymbol{\theta}^* \in \Theta, \quad \boldsymbol{\theta}_0 \in \Theta_0\end{aligned} \tag{17}$$

where $\boldsymbol{\theta}^*$ is the vector of parameters maximizing the log-likelihood function, and $\boldsymbol{\theta}_0$ is the vector of parameters maximizing the log-likelihood function under the null hypothesis.

We will not define Θ and Θ_0 for each test explicitly. Θ will be the set of possible parameters restricted by the p.d.f., while Θ_0 will be the set of parameters satisfying the constraints for both the p.d.f. and the null hypothesis.

3.3.1. Continuous Weibull distribution

Christoffersen and Pelletier (2004) fit the discrete durations to a continuous distribution, which gives a slightly misspecified null hypothesis. Under this null hypothesis, the durations will be memory-less and follow an exponential distribution

$$f_{\text{exp}}(d; p) = pe^{-pd}, \quad d \in \mathbb{N}^+ \tag{18}$$

In order to establish a statistical test for independence, they fit the durations to the Weibull distribution

$$f_{\text{CW}}(d; a, b) = a^b b d^{b-1} e^{-(ad)^b}, \quad d \in \mathbb{N}^+ \tag{19}$$

$$\boldsymbol{\theta} = (a, b), \quad a > 0, \quad b > 0 \tag{20}$$

where a is the probability of breach, and b determines the memory of the process. This distribution has the characteristic that when $b = 1$, it equals the exponential distribution, thus it is memory-free. We can test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} H_{0,\text{Ind}}&: b = 1 \\ H_{1,\text{Ind}}&: b \neq 1 \\ H_{0,\text{CC}}&: a = p \text{ and } b = 1 \\ H_{1,\text{CC}}&: a \neq p \text{ or } b \neq 1 \end{aligned}$$

Under the null hypothesis, LR will be asymptotically $\chi^2(1)$ distributed for independence and $\chi^2(2)$ distributed for conditional coverage. However, it is important to recognize that the null hypothesis in Equation (18) is misspecified for discrete durations and will not be equal to the geometric distribution in Equation (14). With geometric distributed durations, LR will not converge to any asymptotic distribution. In order to test the hypothesis that durations follow a geometric distribution, an empirical distribution must be obtained to account for the bias. We will refer to the conditional coverage test as WeibullCon.

3.3.2. Discrete Weibull distribution

Nakagawa and Osaki (1975) define the discrete Weibull density as

$$f_{\text{DW}}(d; q, b) = q^{(d-1)^b} - q^{d^b}, \quad d \in \mathbb{N}^+ \tag{21}$$

$$\boldsymbol{\theta} = (q, b), \quad q \in (0, 1), \quad b > 0 \tag{22}$$

where q is the probability of at least one non-hit observation before a hit occurs, while b determines the memory of the process. With $b = 1$ we get the geometric distribution and we can test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} H_{0,\text{Ind}}&: b = 1 \\ H_{1,\text{Ind}}&: b \neq 1 \\ H_{0,\text{CC}}&: q = 1 - p \text{ and } b = 1 \\ H_{1,\text{CC}}&: q \neq 1 - p \text{ or } b \neq 1 \end{aligned}$$

Under the null hypothesis, LR will be asymptotically $\chi^2(1)$ distributed for independence and $\chi^2(2)$ distributed for conditional coverage. We will refer to the conditional coverage test as WeibullDisc.

3.3.3. Discrete Weibull distribution with alternative parameterization

Haas (2005) suggests the following parameterization, as it will enhance numerical stability.

$$f_{\text{DWH}}(d; a, b) = e^{-a^b(d-1)^b} - e^{-(ad)^b}, \quad d \in \mathbb{N}^+ \tag{23}$$

$$\boldsymbol{\theta} = (a, b), \quad a > 0, \quad b > 0 \tag{24}$$

where a determines the probability of a hit and b the memory of the process. Under the null hypothesis the durations follow a geometric distribution. We can test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} H_{0,\text{Ind}}&: b = 1 \\ H_{1,\text{Ind}}&: b \neq 1 \\ H_{0,\text{CC}}&: a = -\ln(1 - p) \text{ and } b = 1 \\ H_{1,\text{CC}}&: a \neq -\ln(1 - p) \text{ or } b \neq 1 \end{aligned}$$

Under the null hypothesis, LR will be asymptotically $\chi^2(1)$ distributed for independence and $\chi^2(2)$ distributed for conditional coverage. We will refer to the conditional coverage test as WeibullHaas.

3.3.4. Geometric distribution with time-varying hazard rate

Berkowitz et al. (2011) define the hazard function directly, and derive the p.d.f. using the conditional probability. The hazard function and the p.d.f. are defined as

$$\lambda_{\text{DG}}(d; a, b) = ad^{b-1}, \quad d \in \mathbb{N}^+ \tag{25}$$

$$f_{\text{DG}}(d; a, b) = \lambda_{\text{DG}}(d; a, b) \prod_{i=1}^{d-1} (1 - \lambda_{\text{DG}}(i; a, b)) \tag{26}$$

$$\boldsymbol{\theta} = (a, b), \quad a \in (0, 1), \quad b \leq 1 \tag{27}$$

where a determines the probability of a hit and b determines the memory of the process. Under the null hypothesis the hazard function is constant and the p.d.f. collapses to the geometric distribution. We can test the following hypothesis by adding constraints under the null hypothesis

$$\begin{aligned} &H_{0,\text{Ind}}\text{: } b = 1 \\ &H_{1,\text{Ind}}\text{: } b \neq 1 \\ &H_{0,\text{CC}}\text{: } b = 1 \text{ and } a = p \\ &H_{1,\text{CC}}\text{: } b \neq 1 \text{ or } a \neq p \end{aligned}$$

It is important to recognize that the true parameter b, which is restricted to be 1 or less, lies on the boundary of the parameter space when testing $b = 1$. LR will still converge asymptotically to a distribution, but it will not follow a usual χ^2 distribution. The asymptotic distributions are given by Self and Liang (1987).[5] The distributions of LR will asymptotically be a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$ for independence, and a 50:50 mixture of $\chi^2(1)$ and $\chi^2(2)$ for conditional coverage. We will refer to the conditional coverage test as Geometric.

4. Inferring with finite samples

Most backtests yield a test statistic asymptotically following a known probability distribution. However, the test statistic does not necessarily follow this distribution for finite samples. The unknown distribution can be approximated by simulating data yielding the test statistic under the null hypothesis. Using the approximated distributions, we can make accurate inferences for newly calculated test statistics, as described by Dufour (2006).

Assuming the test statistic, S, has an unknown distribution under H_0. Let $\mathbf{S} = (S_1, \ldots, S_N)$ be a sample of N i.i.d. random variables with the same distribution as S. The cumulative distribution function of S, $F(S_0)$, is approximated by,

$$\hat{F}_N(S_0; \mathbf{S}) = \frac{1}{N} \sum_{i=1}^{N} I(S_i \leq S_0), \tag{28}$$

where $I(C)$ is the indicator function of condition C such that,

$$I(C) = \begin{cases} 1 & \text{if condition } C \text{ holds} \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

There is a positive probability of a tie between the calculated statistic and a point in $\mathbf{S}$. To handle ties we define a vector of random variables, $\mathbf{U} = (U_1, \ldots, U_N)$. These are i.i.d. random variables from $U(0, 1)$. Each U_i is associated with S_i in the following manner.

$$\mathbf{Z_i} = (S_i, U_i) \tag{30}$$

$$\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_N}) \tag{31}$$

This is used to compare two statistics, S_i and S_j where $S_i = S_j$. We compare them according to the following definition,

$$\mathbf{Z_i} \leq \mathbf{Z_j} \Leftrightarrow \{S_i < S_j \text{ or } (S_i = S_j,\, U_i \leq U_j)\} \tag{32}$$

The p-value for a newly calculated statistic with its associated random variable, $\mathbf{Z_0} = (S_0, U_0)$, is given by the following formula,

$$\tilde{p}_N(\mathbf{Z_0}; \mathbf{Z}) = \frac{N \times \tilde{G}_N(\mathbf{Z_0}; \mathbf{Z}) + 1}{N + 1}, \tag{33}$$

where,

$$\begin{aligned} \tilde{G}_N(\mathbf{Z_0}; \mathbf{Z}) &= \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{Z_i} \geq \mathbf{Z_0}) \\ &= 1 - \hat{F}_N(S_0; \mathbf{S}) \\ &\quad + \frac{1}{N} \sum_{i=1}^{N} I(S_i = S_0)\, I(U_i \geq U_0) \end{aligned} \tag{34}$$

We use $N = 50{,}000$ when approximating the finite sample distributions. Distributions are approximated for each specification of lookback, sample size and probability of breach.

[5] $H_{0,\text{Ind}}$ and $H_{0,\text{CC}}$ correspond to case 5 and 6 in Self and Liang (1987), respectively.

5. Test size and power properties

Christoffersen (1998) states a criterion, which a breach process should satisfy. We repeat this criterion here, and will from now on refer to it as the conditional coverage criterion.

$$\Pr(I_t \mid \Omega_{t-1}) = p \text{ for all } t \tag{35}$$

From the conditional coverage criterion, we define a weaker criterion ensuring independence of hits, which we refer to as the independence criterion

$$\Pr(I_i \mid \Omega_{i-1}) = \Pr(I_{j-1} \mid \Omega_{j-2}) \text{ for all } i, j \tag{36}$$

This section starts by assessing the test size of the backtests. We simulate data satisfying the conditional coverage criterion, and calculate rejection rates by applying asymptotic critical values.

The power study is then commenced by backtesting simulated return and VaR series resulting in a breach sequence satisfying the independence criterion. The probability of breach differs from the one tested under the null.

We then show how the different backtests react to estimation of VaR using a rolling window. The Normal VaR model is applied on simulated return series from the standard normal distribution and the Student-t distributions, and the ability of the backtests to reject the model is shown.

We conclude with a study of conditional coverage power, backtesting hit sequences also possessing dependence.

All experiments are performed with 1%, 5% and 10% VaR models, sample sizes of 100, 250, 500, 750, 1000, 1,250, and 1,500, and lookback periods of 250 and 1,000 days. We will only present a representative selection in this paper. The complete results will be referred to as the extensive study.[6]

[6]The complete results can be obtained by contacting the authors.

5.1. Test sizes using asymptotic critical values

We simulate return and VaR series resulting in hit sequences satisfying the conditional coverage criterion, with an underlying probability of breach, π. To examine the test size, the data is backtested with the null hypothesis satisfied. We vary the underlying probability of breach and backtest accordingly. The simulation process is further detailed in Appendix B.

If an asymptotic distribution is accurate for the considered sample size, the rejection rates should be close to the significance level of the test, which we set to 5%. Table 1 shows the test size of the tests using asymptotic critical values. The first column reports the sample size of the hit sequence.

The top panel shows the results for 1% VaR. For small samples PF performs poorly and Geometric is generally undersized. Both perform well for large ones. WeibullCon is slightly oversized. Markov and DQLogit are clearly undersized for finite samples, while DQ and WeibullHaas are oversized. WeibullDisc is undersized for small sample sizes and oversized for large ones.

The panel in the middle shows the results for 5% VaR. PF, DQLogit and Geometric are close to the significance level of 5% for most sample sizes. DQ, WeibullDisc and WeibullHaas are oversized for small sample sizes, but perform well for large ones. WeibullCon is generally oversized, while Markov is undersized for small sample sizes and oversized for large ones.

The results in the bottom panel are for 10% VaR. PF, Markov, DQ, DQLogit and Geometric are close to the significance level of 5% for most sample sizes. WeibullDisc and WeibullHaas tend to be oversized for small sample sizes, but are close to significance level for large ones. WeibullCon is clearly oversized for all sample sizes.

We conclude that the test size, generally, differs from the significance level for finite sam-

Table 1: Test size at 5% significance level using asymptotic critical values

Sample	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
				$\pi = 1\%$				
100	0.018	0.018	0.303	0.051	0.057	0.045	0.249	0.030
250	0.095	0.012	0.134	0.021	0.062	0.025	0.155	0.020
500	0.071	0.010	0.148	0.017	0.082	0.044	0.106	0.030
750	0.039	0.021	0.126	0.019	0.093	0.080	0.101	0.073
1000	0.055	0.027	0.102	0.018	0.085	0.084	0.086	0.053
1250	0.066	0.026	0.087	0.021	0.077	0.081	0.081	0.055
1500	0.056	0.033	0.076	0.022	0.069	0.074	0.072	0.057
				$\pi = 5\%$				
100	0.063	0.022	0.086	0.039	0.084	0.062	0.108	0.030
250	0.060	0.042	0.065	0.042	0.078	0.084	0.083	0.053
500	0.053	0.040	0.058	0.052	0.072	0.064	0.062	0.044
750	0.054	0.047	0.057	0.057	0.075	0.059	0.058	0.044
1000	0.051	0.054	0.055	0.060	0.079	0.057	0.055	0.045
1250	0.045	0.078	0.052	0.061	0.083	0.057	0.053	0.046
1500	0.052	0.063	0.054	0.060	0.092	0.052	0.054	0.046
				$\pi = 10\%$				
100	0.044	0.041	0.063	0.056	0.095	0.084	0.084	0.049
250	0.057	0.057	0.053	0.064	0.097	0.061	0.062	0.044
500	0.054	0.049	0.051	0.060	0.139	0.056	0.055	0.045
750	0.053	0.053	0.051	0.055	0.194	0.053	0.053	0.044
1000	0.045	0.053	0.050	0.054	0.248	0.053	0.052	0.046
1250	0.049	0.050	0.051	0.055	0.308	0.051	0.051	0.045
1500	0.048	0.052	0.049	0.051	0.363	0.052	0.054	0.048

Note: Return and VaR series are simulated resulting in hit sequences satisfying the conditional coverage criterion, with a breach probability π. The data is backtested with the probability of breach set to π under the null. Rejection rates are calculated over 50,000 successful[7] Monte Carlo trials. Sample is the sample size of the simulated data. PF is an unconditional coverage test. Markov is a first-order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

ples.[8] Using asymptotic critical values can, as a consequence, give very misleading results. For instance will PF with 250 days sample size and 1% VaR, which is commonly used, reject almost twice its significance level. When computing power we will therefore apply the Monte Carlo testing technique, as described in Section 4.

[7]Samples where the backtests fail to calculate a test statistic are re-simulated. The results are therefore conditional on the test statistics being feasible. A note on the feasibility ratios of the tests can be found in Appendix C.

[8]The test sizes will converge to 5% for large samples for all tests, except WeibullCon, which has a misspecified null hypothesis. Note that the convergence will not be monotonic, as seen from Table 1.

Table 2: Power to reject wrong unconditional probability of breach at 5% significance level using finite sample distributions

π	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
0.025	0.991	0.971	0.658	0.875	0.967	0.979	0.979	0.987
0.030	0.908	0.838	0.325	0.625	0.808	0.854	0.856	0.892
0.035	0.664	0.547	0.122	0.333	0.505	0.578	0.586	0.639
0.040	0.343	0.258	0.048	0.148	0.228	0.276	0.289	0.322
0.045	0.123	0.096	0.036	0.069	0.087	0.106	0.113	0.122
0.050	0.050	0.053	0.048	0.050	0.048	0.047	0.050	0.053
0.055	0.101	0.096	0.101	0.068	0.079	0.072	0.077	0.086
0.060	0.272	0.230	0.213	0.132	0.186	0.182	0.187	0.217
0.065	0.516	0.446	0.392	0.262	0.358	0.377	0.395	0.446
0.070	0.747	0.684	0.597	0.450	0.595	0.613	0.631	0.683
0.075	0.897	0.851	0.777	0.653	0.792	0.809	0.819	0.863

Note: Return and VaR series are simulated resulting in hit sequences satisfying the independence criterion. The underlying probability of breach, π, ranges from 0.025 to 0.075. The data is backtested with the probability of breach set to 0.05 under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. The sample size is 1,000 and the lookback period for the estimation is 250 days. PF is an unconditional coverage test. Markov is a first-order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

5.2. Testing unconditional coverage using finite sample distributions

We simulate return and VaR series resulting in hit sequences satisfying the independence criterion. The underlying probability of breach, π, ranges from 2.5% to 7.5%. The data generation is further detailed in Appendix B. To examine the power against deviating proportion of hits, the data is backtested with the probability of breach set to 5% under the null.

Table 2 shows the power at 5% significance level using finite sample distributions. The first column shows the underlying probability of breach. PF performs best, as expected. PF is designed to test unconditional coverage, and thus will not have to sacrifice any power to test non-existing dependence in hits. Markov combines an independence test and an unconditional coverage test, where the latter is equivalent to PF. It performs well. The duration-based tests, with the exception of WeibullCon, perform almost as well as PF. The regression-based tests perform noticeably worse than the other tests. The extensive results state that the internal ranking of the tests is the same for all sample sizes and all values of π, though the power varies.

5.3. Implications of estimating VaR using a rolling lookback window

The most common technique to construct VaR estimates is by using a rolling lookback window.[9] However, this technique introduces bias in the backtests as the VaR estimates will be serially dependent because of overlapping samples. It is possible to avoid serial dependence by using non-overlapping samples, but with 250 lookback days one will be able to create only one estimate for every year of data. Despite its disadvantages, rolling sample window is to our knowledge the best method for practical purposes. We will continue the power

[9]This technique is suggested by Alexander (2008a) and is widely used by practitioners, as discussed by Pérignon and Smith (2010).

Table 3: Rejection rates at 5% significance level using finite sample distributions and estimating VaR with a rolling window

L	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
Normal VaR model								
50	0.020	0.041	0.696	0.672	0.046	0.030	0.031	0.018
100	0.009	0.028	0.341	0.328	0.030	0.023	0.024	0.018
250	0.009	0.026	0.147	0.146	0.029	0.024	0.025	0.021
500	0.016	0.029	0.101	0.100	0.031	0.028	0.030	0.026
1,000	0.029	0.039	0.090	0.091	0.038	0.036	0.038	0.036
Historical Simulation VaR model								
50	0.072	0.082	0.971	0.959	0.088	0.055	0.059	0.049
100	0.003	0.031	0.636	0.599	0.029	0.021	0.022	0.015
250	0.001	0.022	0.252	0.248	0.022	0.020	0.021	0.017
500	0.003	0.021	0.146	0.148	0.024	0.020	0.021	0.018
1,000	0.023	0.035	0.125	0.124	0.036	0.034	0.036	0.033

Note: Normally distributed returns are simulated. The VaR series are estimated with a 5% 1) Normal VaR model and 2) Historical Simulation VaR model. The models use a rolling window to estimate VaR. The data is backtested with the probability of breach set to 5% under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. The sample size is 1,000 and L is the size of the lookback window. PF is an unconditional coverage test. Markov is a first-order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

study with this technique and discuss the results in light of the implications.

5.3.1. Implications of estimation using correctly specified VaR model

We simulate i.i.d. returns from the standard normal distribution

$$r_t \sim N(0,1) \tag{37}$$

We then estimate VaR using a Normal VaR model

$$\widehat{\text{VaR}}_t(p) = -\bar{r_t} - \Phi^{-1}(p)\, s_t, \tag{38}$$

where Φ^{-1} is the inverse c.d.f. of the standard normal distribution, and $\bar{r}_t$ and s_t are the estimated mean and standard deviation using a rolling window. The estimators are given in Appendix D. p is the target probability, set to 5%.

The data is backtested with the probability of breach set to 5% under the null. If the estimated parameters in Equation (38) are replaced with population parameters, we will get a breach process satisfying the conditional coverage criterion in Equation (35), with an unconditional probability of breach equal to p. As we then are under the null hypothesis, the rejection rates of the tests will be the same as the significance level of the tests. However, as we estimate the parameters, this criterion is no longer satisfied.

The rejection rates for the tests are shown in Table 3. Results using a Historical Simulation VaR model[10] are presented for comparison. Table 3 shows that the rejection rates are different from the significance level of the tests due to estimation.

[10] Alexander (2008a) gives details on the Historical Simulation VaR model and its properties.

Estimating the Normal VaR model will make the unconditional probability of breach depend on the size of the lookback window, L. We have

$$\begin{aligned}\Pr(r_t < -\widehat{\mathrm{VaR}}_t(p)) &= \Pr(I_t = 1) \\ &= F_{L-1}\left(\frac{\Phi^{-1}(p)}{\sqrt{1+\frac{1}{L}}}\right),\end{aligned} \tag{39}$$

where Φ^{-1} is the inverse c.d.f. of the standard normal distribution and F_{L-1} is the c.d.f. of the Student-t distribution with $L-1$ degrees of freedom. The complete derivation can be found in Appendix E.

An unconditional probability of breach different from p will increase the rejection rate of all backtests. For appropriate large values of L, this implication is negligible.[11] Therefore, we assume the unconditional probability of breach, and the variance of I_t, to be the same whether the parameters in the VaR model are estimated or not.

Estimating VaR using a rolling window, with overlapping samples, will give three important implications.

First, dependence in the VaR estimates due to rolling window will give autocovariance in the hit sequence[12]. The variance of the total number of breaches is given by

$$\mathrm{var}\left(\sum_{t=1}^{N} I_t\right) = N\mathrm{var}(I_t) + \sum_{i\neq j} cov(I_i, I_j) \tag{40}$$

where $cov(I_i, I_j)$ is the covariance between I_i and I_j.

Appendix G shows that $\sum_{i\neq j} cov(I_i, I_j) < 0$. This implies lower variance for the total number of breaches. The isolated effect is that a larger proportion of the test statistics will fall within the non-rejection region of the test, and the rejection rates will decrease. This is directly observed for the PF test as it only considers the total number of breaches.

Second, autocovariance in the hit sequence also gives the following conditional probability of breach

$$\Pr(I_t = 1 \mid I_{t-l} = 1) = \frac{cov(I_t, I_{t-l})}{p} + p \tag{41}$$

$$\Pr(I_t = 1 \mid I_{t-l} = 0) = -\frac{cov(I_t, I_{t-l})}{1-p} + p, \tag{42}$$

where l is the time lag.[13] As $\mathrm{cov}(I_t, I_{t-l}) \neq 0$, the conditional probability of breach differs from the unconditional probability.

This bias in the conditional probability will increase the rejection rate of all conditional coverage tests, as they test for independence. To assess the magnitude, we can look at the rejection rate of the Markov and the duration-based tests. These tests will be affected by both the lowered variance and the biased conditional probability. We know that the latter will increase the rejection rate. The lowered variance will, on the other hand, decrease the rejection rates. From the results in Table 3, we observe rejection rates lower than the significance level, indicating that the bias resulting from lower variance is the strongest.

Third, there is covariance between the hit function and the lagged VaR estimates. This can be observed in Figure F.3(b).

This will increase the rejection rate when testing for dependence between hits and lagged VaR estimates. The magnitude can be assessed by looking at DQ and DQLogit, which use lagged VaR estimates as regressors. These tests will be affected by the implications from lowered variance, the conditional probability, and the covariance between hits and lagged VaR estimates. The first implication leads to decreased rejection rates, while the two other will

[11] For smaller lookback windows we can multiply the VaR estimate with a correction constant to correct the bias.

[12] This can be observed from Figure F.3(a).

[13] The derivation of Equation (41) and Equation (42) can be found in Appendix H.

increase it. As Markov is a special case of the DQ test, we know that the combined effect of the two first implications should be a decrease in the rejection rate. From the results in Table 3, we observe rejection rates higher than the significance level, indicating that the bias from the non-zero covariance between hits and lagged VaR estimates is very strong.

When we have a correctly specified VaR model producing VaR estimates very close to the true VaR, we do not want to reject the null hypothesis for practical purposes. Hence, we state that a correctly specified VaR model is our null hypothesis when estimation risk is negligible. The following discussion is based on this assumption.

As seen from Table 3, the probability of committing a type I error is higher than the significance level for DQ and DQLogit, and lower for the others. Ideally, we like to use a test procedure for which the type I error probability is small. A rejection rate under the null hypothesis lower than the significance level is not a serious problem, as a rejection for an undersized test will strengthen our belief that we are under an alternative hypothesis. However, rejecting the null hypothesis more than the significance level, as DQ and DQLogit do, is a serious problem, as we have no upper bound for the type I error probability. As a consequence making a type I error is no longer highly unlikely.

5.3.2. Implications of estimation using incorrectly specified VaR model

We simulate return series from the following leptokurtic model

$$r_t = t\left(v\right)\sqrt{\frac{v-2}{v}}, \tag{43}$$

where $t\left(v\right)$ is a Student-t distributed random variable with v degrees of freedom. r_t is constructed such that the variance is held constant, $var\left(r_t\right) = 1$, for $v > 2$. We estimate VaR using the 5% Normal VaR model as in Equation (38). The data is backtested with the probability of breach set to 5% under the null. Using the Normal VaR model with population parameters for the mean and the standard deviation, denoted $\text{VaR}_t(p)$, will give the following probability of breach

$$\begin{aligned}\pi^* &= \Pr(r_t < -\text{VaR}_t(p)) \\ &= F_v\left(\Phi^{-1}\left(p\right)\sqrt{\frac{v}{v-2}}\right),\end{aligned} \tag{44}$$

where F_v is the c.d.f. of a Student-t distribution with v degrees of freedom. We vary v to get $\pi^* = 0.025$, 0.030, 0.035, 0.040 and 0.045. The values of v providing π^* are found by Equation (44) and are listed in Table I.8.

Table 4 shows the rejection rates. When we introduce leptokurtosis, to deviate further from the null hypothesis, we observe that the power increases for all tests as expected. Comparing the results in Table 2 and Table 4, we observe that the rejection rates are lower when estimating VaR, with the exception of DQ and DQLogit for values of π^* close to the tested value, 5%.

Estimation using a rolling window will make the probability of committing type II error higher for all tests, except for DQ and DQLogit for π^* close to 5%.

5.4. Testing conditional coverage using finite sample distributions

We simulate return series from the following GARCH(1, 1)-t model, by Bollerslev (1986)

$$r_t = t\left(v\right)\sigma_t\sqrt{\frac{v-2}{v}} \tag{45}$$

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{46}$$

where $t\left(v\right)$ is a Student-t random variable with v degrees of freedom, α is the error parameter, β is the persistence parameter and ω is a constant. We set $\alpha + \beta = 0.99$ and $\omega = 0.01$. This setup gives stationarity, which ensures a finite and positive unconditional variance, and

Table 4: Power to reject wrong unconditional probability of breach at 5% significance level using finite sample distributions and estimating VaR with a rolling window

π^*	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
0.025	0.685	0.561	0.358	0.597	0.526	0.591	0.614	0.675
0.030	0.544	0.435	0.267	0.489	0.378	0.468	0.469	0.547
0.035	0.359	0.289	0.202	0.377	0.247	0.325	0.315	0.362
0.040	0.187	0.142	0.142	0.247	0.113	0.156	0.163	0.192
0.045	0.049	0.043	0.119	0.158	0.037	0.049	0.054	0.061

Note: Student-t distributed returns are simulated. The VaR series are estimated with a 5% Normal VaR model with a rolling window. The data satisfies the independence criterion, with π^* ranging from 0.025 to 0.045 by varying the degrees of freedom as described in Equation (44). The data is backtested with the probability of breach set to 5% under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. The sample size is 1,000 and the lookback period for the estimation is 250 days. PF is an unconditional coverage test. Markov is a first-order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

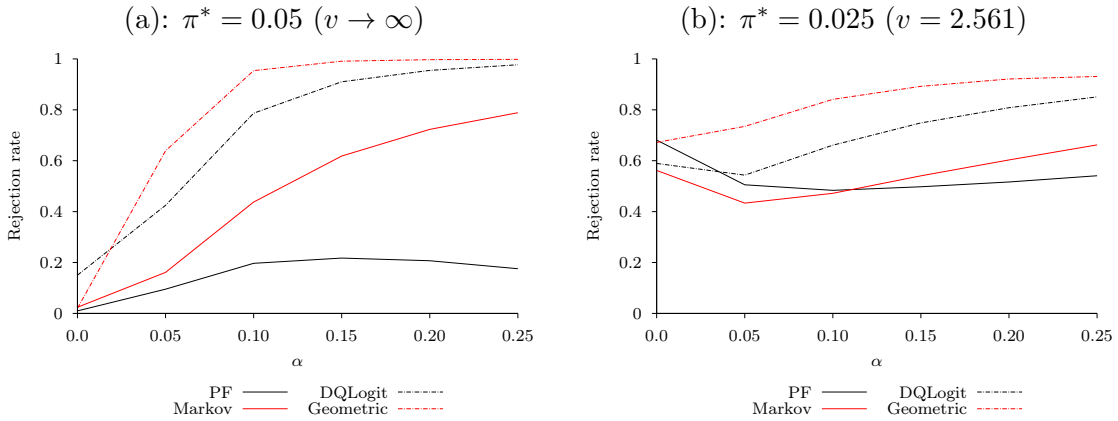

Figure 1: Power to reject wrong conditional probability of breach at 5% significance level using finite sample distributions and estimating VaR with a rolling window. GARCH$(1,1)$-t distributed returns are simulated. The VaR series are estimated with a 5% Normal VaR model. π^* is the probability of breach given a 5% Normal VaR model, with population parameters, applied on Student-t returns with v degrees of freedom, as described in Equation (44). The data is backtested with the probability of breach is set to 5% under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. α is the error parameter in the GARCH model. The sample size is 1,000 days and the lookback period for the estimation is 250 days. Details on each test are given in Section 3.

reflects typical financial markets.[14] VaR is still estimated using a Normal 5% VaR model with rolling window. α is varied from 0.00 to 0.25. To examine the power against wrong proportion and clustering of hits, the data is backtested with the probability of breach set to 5% under the null. Table 5 shows the power at 5% significance level using finite sample distributions. The first column shows the error parameter, α.

Figure 1 illustrates the results from Table 5 showing the Geometric, DQLogit, PF and Markov test.

[14]Examples are equity markets and exchange rates as explained by Alexander (2008b) and Taylor, S. J. (2005), respectively.

Table 5: Power to reject wrong conditional probability of breach at 5% significance level using finite sample distributions and estimating VaR with a rolling window

α	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
				$\pi^* = 0.03$ ($v = 2.818$)				
0.00	0.548	0.439	0.269	0.494	0.389	0.472	0.483	0.556
0.05	0.354	0.312	0.407	0.485	0.381	0.595	0.595	0.696
0.10	0.327	0.391	0.628	0.644	0.573	0.777	0.784	0.852
0.15	0.349	0.488	0.740	0.762	0.681	0.858	0.868	0.909
0.20	0.382	0.571	0.817	0.825	0.739	0.901	0.903	0.936
0.25	0.412	0.641	0.858	0.877	0.772	0.922	0.925	0.952
				$\pi^* = 0.04$ ($v = 3.938$)				
0.00	0.180	0.137	0.141	0.252	0.116	0.153	0.165	0.194
0.05	0.121	0.165	0.426	0.408	0.274	0.534	0.535	0.643
0.10	0.123	0.326	0.715	0.687	0.612	0.834	0.841	0.903
0.15	0.134	0.471	0.844	0.818	0.771	0.930	0.925	0.958
0.20	0.149	0.566	0.902	0.883	0.837	0.959	0.958	0.977
0.25	0.172	0.650	0.931	0.922	0.877	0.970	0.973	0.984
				$\pi^* = 0.05$ ($v \to \infty$)				
0.00	0.010	0.024	0.151	0.150	0.028	0.023	0.026	0.020
0.05	0.095	0.161	0.493	0.425	0.235	0.517	0.536	0.639
0.10	0.197	0.438	0.834	0.787	0.755	0.915	0.920	0.954
0.15	0.217	0.618	0.932	0.910	0.909	0.982	0.980	0.991
0.20	0.207	0.723	0.968	0.955	0.956	0.994	0.992	0.997
0.25	0.176	0.789	0.984	0.977	0.974	0.996	0.997	0.998

Note: GARCH(1, 1)-t distributed returns are simulated. The VaR series are estimated with a 5% Normal VaR model. π^* is the probability of breach given a 5% Normal VaR model, with population parameters, applied on Student-t returns with v degrees of freedom, as described in Equation (44). The data is backtested with the probability of breach set to 5% under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. α is the error parameter in the GARCH model. The sample size is 1,000 days and the lookback period for the estimation is 250 days. PF is an unconditional coverage test. Markov is a first order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

Figure 1 (a) shows the power with $\pi^* = 0.05$ ($v \to \infty$). The Geometric test performs best with steepest power curve and highest power for high degrees of heteroscedasticity. The DQLogit test has higher power for low degrees of heteroscedasticity. This is due to the problems regarding parameter estimation with a rolling window, as discussed in Section 5.3.1. As expected, the PF test has the lowest power.

Figure 1 (b) shows the power with $\pi^* = 0.025$, ($v = 2.561$). Again the Geometric test performs best with highest power for all specifications. The PF test performs well for the case of no heteroscedasticity, but does not capture dependence in the hit sequence, when $\alpha > 0$.

The same findings are seen in the extensive results for all sample sizes larger than 100 data points and all VaR target probabilities, though the power varies.

5.4.1. Implications of sample size

As expected, sample size and power are clearly connected. Figure 2 shows the power with $\pi^* = 0.025$ ($v = 2.561$) for different sample sizes. These settings yield high degrees of heteroscedasticity and leptokurtosis in the underlying returns. The power increases steadily as the sample size increases from 100 to 1,500 data points. The power is below 0.60 with 250 data points for all tests considered. Testing under such conditions would result in a probability of type II error of 0.40. Thus for a clearly misspecified model, we would still fail to reject it 40% of the time.

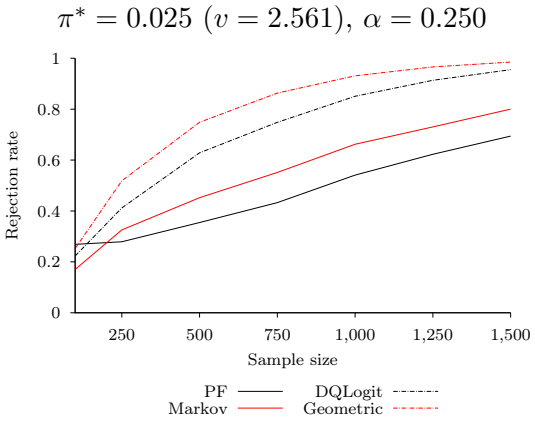

Figure 2: Power at 5% significance level using finite sample distributions estimating VaR using a rolling window. GARCH(1, 1)-t distributed returns are simulated. The VaR series are estimated with a 5% Normal VaR model. π^* is the probability of breach given a 5% Normal VaR model, with population parameters, applied on Student-t returns with v degrees of freedom, as described in Equation (44). The data is backtested with the probability of breach is set to 5% under the null. Rejection rates are calculated over 20,000 successful Monte Carlo trials. α is the error parameter in the GARCH model. The lookback period for the estimation is 250 days. Details on each test are given in Section 3.

Table 6 is a practitioner's table which indicates how large a sample size is needed to make use of the Geometric backtest. The table gives sample sizes which yield power above 0.70 for reasonable introduction of heteroscedasticity and leptokurtosis.

Table 6: Sample size recommendation for backtesting with the Geometric test

VaR target probability	Min. sample size
1%	1,000
5%	750
10%	500

Note: The table shows the minimum sample size recommended, given the VaR target probability, for backtesting using the Geometric test. Details on each test are given in Section 3. The sample size recommendations yield power above 0.70 for reasonable introduction of heteroscedasticity and leptokurtosis.

6. Conclusion

This paper evaluates, through an extensive power study, the performance of the most recognized Value-at-Risk backtests. We provide four key findings: i) Asymptotic critical values should not be used when backtesting finite-sample data. ii) The common implementation of the Dynamic Quantile test, by Engle and Manganelli (2004), has a too high rejection rate for correctly specified VaR models. iii) The Geometric test, by Berkowitz et al. (2011), has the highest power overall. iv) Backtesting 100 or 250 data points will not be sufficient due to low power.

We first assess the test size of all backtests. The results show that none of the tests follow the asymptotic distribution for sample sizes of 100 up to 1,500. Some tests have up to twice the expected rejection rate for certain specifications. Using asymptotic distributions when backtesting can thus give misleading results. One should therefore use finite sample distributions from Monte Carlo simulations to calculate p-values, as described in Section 4.

We then review the power of the backtests to reject hit processes with a probability of breach different from the one tested for. We start by doing this from a solely theoretical point of view and then show the implication of estimation. We reveal several novel findings. The most important is that the common implementation of the Dynamic Quantile test, by Engle

and Manganelli (2004), has a too high rejection rate for correctly specified VaR models, when estimating VaR with a rolling window. We illustrate this with both the Normal and the Historical Simulation VaR model. The latter model is used by Berkowitz et al. (2011) when they conclude that the Dynamic Quantile test has best power properties.

The results from the power study, considering varying breach probability and clustering of breaches, show that the Geometric backtest performs best overall.

The power strongly depends on the amount of backtesting data, where more data gives higher power. We identify minimum sample sizes, when applying finite sample distributions to the Geometric backtest. These sample sizes are 1,000, 750 and 500 when testing for 1%, 5% and 10% VaR, respectively. The other backtests, which have less power, will need even more data. Sample sizes of 100 and 250 data points for backtesting yield low power. This is an important finding in light of regulatory demands, by Basel Committe on Banking Supervision (2011), stating that backtesting can be done using a minimum of one year with daily data.

Appendix A. Definition of Value-at-Risk

Given some target probability, p, the Value-at-risk of the portfolio is given by the smallest number, l, such that the probability that the loss exceeds l is not larger than p. Mathematically, if r_t is the random variable representing the upcoming return at time t, then $\mathrm{VaR}_t(p)$ is given by,

$$\mathrm{VaR}_t(p) = -\inf\{x \in \mathbb{R} : \Pr(r_t > x) \leq 1 - p\} \tag{A.1}$$

For example, if a portfolio has a 5% daily VaR of 12%, there is 5% probability to experience a return less than or equal to -12% the next day given that the portfolio does not trade. It is also expected that the portfolio will suffer a return less than or equal to -12% 1 out of 20 days if the VaR holds a constant target probability.

Appendix B. Generation of return and VaR series under the null hypothesis

Consider a sequence of VaR forecasts, $\mathrm{VaR}_t(p)$, and hits I_t, given by

$$I_t = \begin{cases} 1 & \text{if } r_t < -\mathrm{VaR}_t(p) \\ 0 & \text{otherwise,} \end{cases} \tag{B.1}$$

Christoffersen (1998) states that a sequence of VaR forecasts is efficient with respect to the information set if the following criterion is satisfied.

$$\Pr(I_t \mid \Omega_{t-1}) = p, \text{ for all } t, \tag{B.2}$$

where Ω_{t-1} is the information set at time $t-1$, and p is the probability of breach.

To simulate stochastic return and VaR series satisfying this criterion, we draw random numbers from two normal distributions.

$$\mu_t \sim N(0, 1) \tag{B.3}$$
$$r_t \sim N(\mu_t, 1) \tag{B.4}$$

where μ_t is $t-1$ measurable and is the mean of the return distribution at time t. We assume that we know that returns are normally distributed with population standard deviation 1. As μ_t is measurable at time $t-1$, the VaR calculation is straightforward.

$$\mathrm{VaR}_t(p) = -\mu_t - \Phi^{-1}(p), \tag{B.5}$$

such that Equation (B.2) is satisfied, that is

$$\begin{aligned} \Pr(I_t \mid \Omega_{t-1}) &= \Pr(r_t < -\mathrm{VaR}_t(p) \mid \Omega_{t-1}) \\ &= \Pr\left(r_t - \mu_t < \Phi^{-1}(p) \mid \Omega_{t-1}\right) \\ &= \Phi(\Phi^{-1}(p)) \\ &= p, \end{aligned} \tag{B.6}$$

where Φ is the c.d.f. of the standard normal distribution.

Table C.7: Feasibility ratios at 5% significance level with 1% unconditional probability of breach

Sample	PF	Markov	DQ	DQLogit	Cont	Disc	Haas	Geo
100	1.000	0.627	0.251	0.252	0.202	0.185	0.248	0.262
250	1.000	0.922	0.705	0.705	0.622	0.564	0.695	0.710
500	1.000	0.994	0.963	0.961	0.933	0.882	0.954	0.956
750	1.000	0.999	0.995	0.995	0.989	0.974	0.994	0.995
1000	1.000	1.000	1.000	0.999	0.999	0.993	1.000	1.000
1250	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
1500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: Return and VaR series are simulated resulting in hit sequences satisfying the conditional coverage criterion. The underlying probability of breach is 1%. The data is backtested with the probability of breach set to 1% under the null. Sample is the sample size of the simulated data. PF is an unconditional coverage test. Markov is a first-order Markov test. DQ and DQLogit are regression-based tests. Cont, Disc, Haas and Geo are duration-based tests referred to as WeibullCon, WeibullDisc, WeibullHaas and Geometric. Details on each test are given in Section 3.

Appendix C. Feasibility ratios

All backtests, except the PF test, have certain minimum requirements in order to calculate a test statistic. E.g. if the return and VaR series result in zero breaches, only the PF test is able to compute a test statistic. All other backtests will be infeasible. Larger sample size and higher probability of breach increase the feasibility ratios of the tests.

Table C.7 shows the feasibility ratios of the tests for different sample sizes with a probability of breach set to 1%. We see that the feasibility ratio is low for sample sizes of 100 and 250, and close to one for sample sizes larger or equal to 500. In our study, the feasibility ratios are close to one for all recommended sample sizes.

Appendix D. Normal VaR estimation

Consider a return process with i.i.d. normal variables with mean μ and standard deviation σ.

$$r_t \sim N\left(\mu, \sigma^2\right) \tag{D.1}$$

The theoretical VaR, given a target breach probability p, is then given by

$$\mathrm{VaR}_t(p) = -\mu - \Phi^{-1}(p)\sigma, \tag{D.2}$$

where Φ^{-1} is the inverse c.d.f. of the standard normal distribution.

With unknown mean and standard deviation, the parameters must be estimated. We calculate the sample mean, $\bar{r}_t$, using the L previous observations.

$$\bar{r}_t = \frac{1}{L}\sum_{i=t-L}^{t-1} r_i \tag{D.3}$$

We calculate the sample variance, s_t^2, using the same observations.

$$s_t^2 = \frac{1}{L-1}\sum_{i=t-L}^{t-1} (r_i - \bar{r}_t)^2 \tag{D.4}$$

The estimated VaR forecast at time t is then given by

$$\widehat{\mathrm{VaR}}_t(p) = -\bar{r}_t - \Phi^{-1}(p)s_t \tag{D.5}$$

Appendix E. Unconditional probability of breach when using Normal VaR estimation

Consider a return process with i.i.d. normal variables

$$r_t \sim N\left(\mu, \sigma^2\right), \tag{E.1}$$

and VaR estimated as described in Appendix D.

The unconditional probability of breach using the VaR estimate will be

$$\begin{aligned}\Pr(r_t < -\widehat{\mathrm{VaR}}_t(p)) &= \Pr(r_t < \bar{r}_t + \Phi^{-1}(p)s_t) \\ &= \Pr\left(\frac{r_t - \bar{r}_t}{s_t} < \Phi^{-1}(p)\right)\end{aligned} \tag{E.2}$$

Knowing that r_t and $\bar{r}_t$ are independent and normally distributed, $y_t = r_t - \bar{r}_t$ will be distributed as

$$y_t = r_t - \bar{r}_t \sim N\left(0, \sigma^2 + \frac{\sigma^2}{L}\right), \tag{E.3}$$

or

$$y_t = z_t\sqrt{\sigma^2 + \frac{\sigma^2}{L}}, \tag{E.4}$$

where z_t are i.i.d. standard normal variables and L is the lookback period.

The sample variance is a chi-squared distributed random variable with $L-1$ degrees of freedom, multiplied with a constant

$$s_t^2 = \frac{\sigma^2}{L-1}V \tag{E.5}$$

$$V \sim \chi^2_{L-1} \tag{E.6}$$

y_t and s_t are independent. The ratio of y_t and s_t will be

$$\frac{r_t - \bar{r}_t}{s_t} = \frac{z_t\sqrt{\sigma^2 + \frac{\sigma^2}{L}}}{\sqrt{\frac{\sigma^2}{L-1}V}} = z_t\sqrt{\frac{L-1}{V}}\sqrt{1 + \frac{1}{L}} \tag{E.7}$$

We substitute the random variables in Equation (E.7) with a Student-t distributed random variable which is defined as

$$t(v) = z\sqrt{\frac{v}{V}}, \tag{E.8}$$

where V is a chi-squared distributed random variable with v degrees of freedom, and z is a standard normal distributed random variable. We then get

$$\frac{r_t - \bar{r}_t}{s_t} = t(L-1)\sqrt{1 + \frac{1}{L}}, \tag{E.9}$$

and calculate the probability given in Equation (E.2)

$$\begin{aligned}&\Pr\left(t(L-1)\sqrt{1 + \frac{1}{L}} < \Phi^{-1}(p)\right) \\ &\quad = \Pr\left(t(L-1) < \frac{\Phi^{-1}(p)}{\sqrt{1 + \frac{1}{L}}}\right) \\ &\quad = F_{L-1}\left(\frac{\Phi^{-1}(p)}{\sqrt{1 + \frac{1}{L}}}\right),\end{aligned} \tag{E.10}$$

where F_{L-1} is the c.d.f. of a Student-t distribution with $L-1$ degrees of freedom.

The unconditional probability of breach is

$$\Pr(r_t < -\widehat{\mathrm{VaR}}_t(p)) = F_{L-1}\left(\frac{\Phi^{-1}(p)}{\sqrt{1 + \frac{1}{L}}}\right), \tag{E.11}$$

and will approach p as L approaches infinity.

Appendix F. Linear dependences due to Normal VaR estimation

Consider a return process generated from the standard normal distribution

$$r_t \sim N(0,1) \tag{F.1}$$

We apply the Normal VaR model in Equation (D.5), assuming that returns are normally distributed with unknown mean and standard deviation.

s_t and $\bar{r}_t$ are serially dependent, as their estimators use overlapping data. s_t will have $L-1$ common observations with s_{t-1}, where L is the lookback window as defined in Appendix D. The same applies to $\bar{r}_t$.

As s_t and $\bar{r}_t$ are serially dependent, VaR estimates, $\widehat{\mathrm{VaR}}_t(p)$, will be serially dependent as well.

The hit function, I_t, is given by

$$I_t = \begin{cases} 1 & \text{if } r_t < -\widehat{\mathrm{VaR}}_t(p) \\ 0 & \text{otherwise,} \end{cases} \tag{F.2}$$

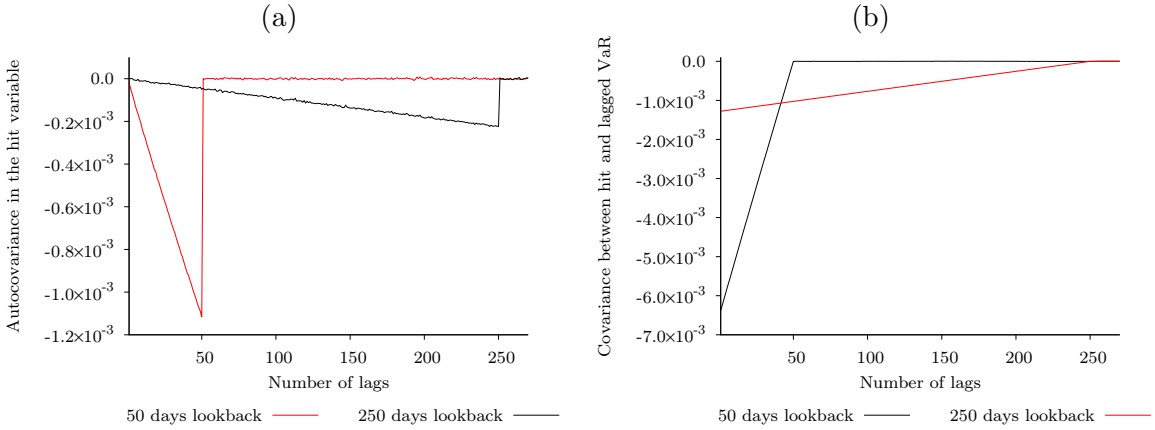

Figure F.3: 1,000,000,000 standard normal distributed returns are simulated. The VaR series is estimated with a 10% Normal VaR model. We then estimate i) the autocovariance in the hit series and ii) the covariance between hits and lagged VaR estimates, with 50 days and 250 days lookback windows.

and will be dependent on lagged VaR estimates and lagged values of itself.

We simulate a long data series of returns and apply the Normal VaR model. We then estimate i) the autocovariance in the hit series and ii) the covariance between hits and lagged VaR estimates. The results are given in Figure F.3.

We see that the magnitude of the autocovariance in the hit sequence increases with the time lag. When the lag is larger than the lookback window, the autocovariance is zero. The magnitude of the covariance between hits and lagged VaR estimates decreases with the time lag. For lags larger than the lookback window, the covariances are zero.

Appendix G. Lower variance of the total number of breaches due to Normal VaR estimation

Consider a VaR sequence obtained from a Normal VaR model. The variance of the total number of breaches, will then be lower with parameters estimated, using overlapping data, than with population parameters. We have

$$\mathrm{var}\left(\sum_{t=1}^{N} I_t\right) = N\mathrm{var}(I_t) + 2\sum_{i<j} cov(I_i, I_j) \tag{G.1}$$

where $\mathrm{var}(I_t) = p^*(1-p^*)$. p^* is the unconditional probability of breach with estimation, given by Equation (E.11).

In Appendix F we showed that $cov(I_i, I_j) = 0$ for $|i-j| > L$, where L is the lookback window as defined in Appendix D. As the hit process is stationary, the autocovariance only depends on the time-shift, l. Hence, we have $cov(I_i, I_{i+l}) = cov(I_{i+\tau}, I_{i+l+\tau}) = \gamma_l$, where γ_l is the autocovariance function.

We simplify Equation (G.1) to take these facts into account

$$\mathrm{var}\left(\sum_{t=1}^{N} I_t\right) = N\mathrm{var}(I_t) + 2\sum_{l=1}^{L}(N-l)\gamma_l \tag{G.2}$$

If we assume that p^* is equal to the unconditional probability of breach without estimation, $\mathrm{var}(I_t)$ will be the same with estimation as without.

Without estimation, $\gamma_l = 0$ for all $l > 0$, and the variance of the sum of hits will simply be $N\mathrm{var}(I_t)$. However, with estimation, the last sum in Equation (G.2) will be negative, as shown in Appendix F, resulting in lower variance.

Lower variance due to estimation is easily observed in simulations. Figure G.4 shows the c.d.f. for a binomial random variable, repre-

senting total number of breaches without estimation, and the c.d.f. with Normal VaR estimation using a rolling window. From the graph we see that Normal VaR estimation gives lower variance.

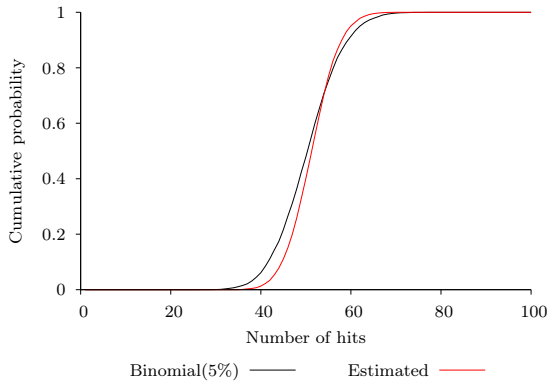

Figure G.4: The cumulative distribution function for a 1,000 trial Binomial(5%) random variable and the cumulative distribution function for the sum of 1,000 hits from normal distributed returns with a 5% estimated Normal VaR. The lookback is 250 days. The distributions are estimated with 10,000 simulations.

Appendix H. Conditional probability of breach with autocovariance in the hit sequence

Consider a hit sequence with binary variables, I_t, where the probability of breach is given by

$$\Pr(I_t = 1) = p \tag{H.1}$$

Note that by using the law of total expectations, we have

$$\begin{aligned} E[I_t] &= E[I_t \mid I_t = 1]\Pr(I_t = 1) \\ &\quad + E[I_t \mid I_t = 0]\Pr(I_t = 0) = \Pr(I_t = 1) \\ &= p \end{aligned} \tag{H.2}$$

We assume there exist dependence in the sequence such that hits are autocorrelated. Thus, they have an autocovariance, γ_l, different from zero, where l is the time lag.

$$\begin{aligned} \operatorname{cov}(I_t, I_{t-l}) &= E\left[(I_t - E[I_t])(I_{t-l} - E[I_{t-l}])\right] \\ &= E[I_t I_{t-l}] - E[I_t]E[I_{t-l}] \\ &= E[I_t I_{t-l}] - p^2 = \gamma_l \end{aligned} \tag{H.3}$$

If the variables are autocorrelated, the conditional probability of breach is given by

$$\begin{aligned} \Pr(I_t = 1 \mid I_{t-l} = 1) &= \frac{\Pr(I_t = 1 \cap I_{t-l} = 1)}{\Pr(I_{t-l} = 1)} \\ &= \frac{E[I_t I_{t-l}]}{E[I_{t-l}]} \end{aligned} \tag{H.4}$$

$$\begin{aligned} \Pr(I_t = 1 \mid I_{t-l} = 0) &= \frac{\Pr(I_t = 1 \cap I_{t-l} = 0)}{\Pr(I_{t-l} = 0)} \\ &= \frac{E[I_t] - E[I_t I_{t-l}]}{1 - E[I_{t-l}]} \end{aligned} \tag{H.5}$$

We substitute $E[I_t I_{t-l}]$ with the relation from Equation H.3 and $E[I_{t-l}]$ with p from Equation H.2 and get

$$\Pr(I_t = 1 \mid I_{t-l} = 1) = \frac{\gamma_l + p^2}{p} = \frac{\gamma_l}{p} + p \tag{H.6}$$

$$\begin{aligned} \Pr(I_t = 1 \mid I_{t-l} = 0) &= \frac{p - \gamma_l - p^2}{1 - p} \\ &= \frac{p(1-p) - \gamma_l}{1 - p} \\ &= -\frac{\gamma_l}{1 - p} + p \end{aligned} \tag{H.7}$$

$\Pr(I_t = 0 \mid I_{t-l} = 1)$ and $\Pr(I_t = 0 \mid I_{t-l} = 0)$ follow from the law of total probability

$$\begin{aligned} &\Pr(I_t = 0 \mid I_{t-l} = 1) \\ &\quad = 1 - \Pr(I_t = 1 \mid I_{t-l} = 1) \\ &\quad = -\frac{\gamma_l}{p} + (1 - p) \end{aligned} \tag{H.8}$$

$$\begin{aligned} &\Pr(I_t = 0 \mid I_{t-l} = 0) \\ &\quad = 1 - \Pr(I_t = 1 \mid I_{t-l} = 0) \\ &\quad = \frac{\gamma_l}{1 - p} + (1 - p) \end{aligned} \tag{H.9}$$

Appendix I. Probability of breach using a Normal VaR model with Student-t distributed returns

Consider a return process

$$r_t = t(v)\sqrt{\frac{v-2}{v}}, \tag{I.1}$$

where $t(v)$ is a Student-t distributed random variable with v degrees of freedom.

We assume that we know the population mean, $\mu = 0$, and the population standard deviation, $\sigma = 1$, of the return process, but do not know the underlying distribution. If we calculate VaR making the mistake of assuming normally distributed returns as in Equation (D.2), we will get an underlying probability of breach, π, different from the target probability, p, used in the Normal VaR model.

Table I.8: Degrees of freedom given Student-t underlying returns and Normal VaR forecasts with population parameters

			$p = 0.01$		
π	0.015	0.014	0.013	0.012	0.011
v	4.977	7.522	10.920	17.340	36.178
			$p = 0.05$		
π	0.025	0.030	0.035	0.040	0.045
v	2.561	2.818	3.218	3.938	5.789
			$p = 0.10$		
π	0.050	0.060	0.070	0.080	0.090
v	2.764	3.156	3.807	5.100	8.944

Note: Degrees of freedom, v, are found by solving Equation (I.2) numerically for given π and p.

To find v given p and π, we solve the following equation numerically.

$$\begin{aligned}\pi &= \Pr(r_t < -\mathrm{VaR}_t(p)) \\ &= F_v\left(\Phi^{-1}(p)\sqrt{\frac{v}{v-2}}\right),\end{aligned} \tag{I.2}$$

where F_v is the c.d.f. of a Student-t distribution with v degrees of freedom, and Φ^{-1} is the inverse c.d.f. of the standard normal distribution. Table I.8 lists the values of v such that we get π for a given p.

References

Alexander, C., 2008a. Market Risk Analysis. Vol. IV. Wiley.

Alexander, C., 2008b. Market Risk Analysis. Vol. II. Wiley.

Basel Committe on Banking Supervision, 1996a. Amendment to the Capital Accord to Incorporate Market Risks. Bank for International Settlements, Basel.

Basel Committe on Banking Supervision, 1996b. Supervisory Framework For The Use of "Backtesting" in Conjunction With The Internal Models Approach to Market Risk Capital Requirements. Bank for International Settlements, Basel.

Basel Committe on Banking Supervision, 2006. International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements, Basel.

Basel Committe on Banking Supervision, 2011. Basel III: A global regulatory framework for more resilient banks and banking systems. Bank for International Settlements, Basel.

Berkowitz, J., Christoffersen, P. F., Pelletier, D., 2011. Evaluating Value-at-Risk models with desk-level data. Management Science 57 (12), 2213–2227.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307–327.

Chatfield, C., 1993. Calculating interval forecasts. Journal of Business and Economic Statistics 11 (2), 121–135.

Christoffersen, P. F., 1996. Essays on forecasting in economics. Ph.D. thesis, Univeristy of Pennsylvania.

Christoffersen, P. F., 1998. Evaluating interval forecasts. International Economic Review 39 (4), 841–862.

Christoffersen, P. F., Pelletier, D., 2004. Backtesting Value-at-Risk: A duration-based approach. Journal of Financial Econometrics 2 (1), 84–108.

Clements, M. P., Taylor, N., 2003. Evaluating interval forecasts of high-frequency financial data. Journal of Applied Econometrics 18 (4), 445–456.

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

Engle, R. F., Manganelli, S., 2004. CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. Journal of Business and Economic Statistics 22 (4), 367–381.

Haas, M., 2005. Improved duration-based backtesting of Value-at-Risk. Journal of Risk 8 (2), 17–38.

Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

J. P. Morgan, 1996. RiskMetrics, 4th Edition. New York.

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

Nakagawa, T., Osaki, S., 1975. The discrete Weibull distribution. IEEE Transactions on Reliability R-24 (5), 300–301.

Pérignon, C., Smith, D. R., 2010. The level and quality of Value-at-Risk disclosure by commercial banks. Journal of Banking and Finance 34 (2), 362–377.

Self, S. G., Liang, K.-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82 (398), 605–610.

Taylor, S. J., 2005. Asset price dynamics, volatility, and prediction. Princeton University Press.