Model Risk in Backtesting Risk Measures

Corinna Evers [*] Johannes Rohde[†]

April 2014

Abstract

Under the Basel II regulatory framework non-negligible statistical problems arise when backtesting risk measures. In this setting backtests often become infeasible due to a low number of violations leading to heavy size distortions. According to Escanciano and Olmo (2010, 2011) these problems persist when incorporating estimation and model risk by adjusting the asymptotic variance of the test statistics. In this paper, we analyze backtests based on hit and duration sequences in a univariate framework by running a simulation study in order to identify the problems of backtests that examine the adequacy of Value at Risk measures. One main finding indicates that backtests of all classes show heavy size distortions. These problems for the relevant Basel II set-up, however, cannot be alleviated by modifying backtests in a way that accounts for estimation risk or misspecification risk.

Keywords: Model risk, backtesting, Value at risk

JEL numbers: C12, C52, G32

[*] Hannover Rück SE, Karl-Wiechert-Allee 50, 30625 Hannover, Germany

[†] Institute of Statistics, School of Economics and Management, Leibniz University of Hannover, Germany

1 Introduction

Backtesting provides an instrument to analyze whether a model used for calculating risk measures is accurate. It is at the core of supervisory activity regarding the resilience of financial institutions in alleviating the impact of financial crisis as the accuracy of risk measures has implications for the solvency capital that financial institutions have to calculate.

BCBS [1996] regulations state that the calculation of a financial institutions' market capital requirement for preventing losses resulting from adverse market conditions be the maximum of either the 0.01% Value at Risk (VaR) or the average VaR reported during the previous 60 days multiplied by a factor depending on the sum of VaR violations during the reporting period (traffic-light approach). Thus, the accuracy of the VaR model is closely linked to the regulatory framework. An accurate VaR model satisfies two properties as defined by Kupiec [1995] and Christoffersen [1998].

Firstly, the unconditional coverage property, formally

$$Pr(I(\alpha) = 1) = \alpha, \tag{1}$$

where $\{I_t\}$ is the hit sequence indicating if a violation occurred or not, claims that the probability of violations during the reporting period equals the α level set for VaR calculation. The VaR model is deemed inaccurate in the sense of failing to be able to account for the incurred risk if the number of violations exceeds the number of expected losses. The risk model is too conservative if the VaR model yields less violations than to be expected.

A second claim is the independence of elements of the hit sequence. If the violations occur in a cluster, the financial institution might not be able to tackle the losses in contrast to a situation where the violations are spread out evenly over the reporting horizon. An accurate VaR model is therefore characterized by satisfying the property of unconditional coverage as well as the independence property,

$$I_t(\alpha) \overset{iid}{\sim} Ber(\alpha), \tag{2}$$

ie the hit sequence is identically and independently distributed with probability α.

Backtests are statistical tests designed for determining the accuracy of VaR models. While several tests have been proposed for each of the two properties, joint tests determine whether the VaR model is entirely accurate in the sense of fulfilling both (1) and (2). However, joint tests are not to be gauged as being universally preferable to mono-property tests as the ability to detect the violation of one of the two properties is decreasing (Campbell [2005]).

A type I error arises when an accurate model with a coverage of 99% is erroneously rejected. When the VaR model is inaccurate with lower coverage, eg 2% type II error is the probability that the inaccurate model is not rejected. If the power of the backtest is low, then the probability of classifying an inaccurate model as accurate (not rejecting the null) is comparatively high. Backtests should not be over- or undersized and possess high power. In a Monte Carlo study we analyze the problems of common backtest procedures. The main result of this paper will be that even when accounting for model risk, regulation sets restrictions to backtesting.

The paper is organized as follows: the next section describes relevant backtesting categories. It serves a starting point for further derivations of multivariate backtests which will be suggested as a mean to overcome problems resulting from supervisory restrictions. In the third chapter we conduct a Monte Carlo study and analyze the problems that arise when conducting univariate backtests in the course of regulation aspects. Finally, the last section provides a conclusion.

2 Overview of backtests

Backtests can be distinguished into frequency-based as well as size-based tests. While the former tests examine the sequence obtained from the exceedance of VaR above the realized profit and losses series, the latter tests are constructed from the size of the exceedance conditioned on the violations. As the regulatory framework is based upon the violations and not on their size, size-based tests are relatively rare to be found in the literature due to regulatory constraints (Lopez [1999]).

The most basic backtests for testing the unconditional coverage property, the time until first failure (TUFF) test and its generalization, the proportion of failures (POF) test, were suggested by Kupiec [1995]. As shown in Kupiec [1995] the simplicity of the TUFF test ignores the total number of failures since the start of monitoring, the POF test should always be run to verify potential loss estimates in place or in addition. In contrast to the TUFF framework, where only the elapsed time until the first failure is considered, the POF uses the entire information. To this (and all further analyses) consider a hit sequence $\{I_t\}_{t=1}^{n}$ of size n, where $\forall t : I_t \in \{0, 1\}$, n_1 denotes the number of hits (ie $I_t = 1$) and $n_0 = n - n_1$ (ie $n_0 = \sharp(I_t = 0)$). The probability of observing n_1 hits in a sample of size n is given by the the probability function of the binomial distribution,

$$Pr(\sharp(I_t = 1) = n_1) = \binom{n}{n_1} (1 - \alpha)^{n_0} \, \alpha^{n_1}.$$

For the null hypothesis of the POF test, $H_0 : \alpha = \hat{\Pi}$ with $\hat{\Pi} = \frac{n_1}{n}$, the associated test is a Likelihood Ratio (LR) test and its test statistics is given by

$$K = -2 \log \left(L(\alpha)/L(\hat{\Pi}) \right)$$

where α denotes the failure probability under the null and $L(\cdot)$ is the corresponding Likelihood function.

However, if the sample size is relatively small, both tests appear to have poor ability to distinguish between the underlying failure probability in the null hypothesis and failure probabilities that are slightly higher (see Kupiec [1995]). Thus, these frameworks might not be adequate for the analysis of the accuracy of VaR estimates covering only one trading year. Furthermore, a frequently arising problem consists in the non-existence of violations during the reporting period. This issue becomes most important when VaR models with a small failure probability are evaluated. In these cases the Kupiec tests are not computable.

When testing the *iid* hypothesis of the hit sequence the autocorrelation of the sequence itself or the equidistance of the time span between consecutive violations is examined. These tests require complete specification of the alternative hypotheses in the sense that the way how violation clusters occur has to be specified exactly. Autocorrelation-based tests can be constructed by testing the autocorrelation structure in the hit sequence itself, $\{I_t\}$, or in the demeaned sequence, $\{I_t - \alpha\}$, which forms a sequence of martingale difference summands (Berkowitz et al. [2009]).

The test by Christoffersen [1998] was the first test of this kind. The basic idea behind this LR-type test consists in the following comparison: If there is no dependence between two consecutive observations, then the probability of monitoring no violation on the day after a violation took place should be equal to the probability of monitoring no violation when no violation was observed on the day before, too.

As in Kupiec [1995] the LR framework is used and built on Markov chains. The independence of the observations of the hit sequence is tested under the null against the alternative of a first-order Markov chain where the stochastic matrix

$$\Pi_1 = \begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}$$

represents the transition matrix and $\pi_{i,j} = P(I_t = j | I_{t-1} = i), i, j \in \{0, 1\}$ the transition probabilities. Let n_{ij} be the number of observations with value i and previous value j. Then the likelihood function for the hit sequence $\{I_t\}$ yields

$$L(\Pi_1) := L(\Pi_1; \{I_t\}) = \pi_{00}^{n_{00}} \pi_{01}^{n_{01}} \pi_{10}^{n_{10}} \pi_{11}^{n_{11}}$$

This is the likelihood under validness of the alternative model while the likelihood for the null model can be computed by considering the stochastic matrix

$$\Pi_2 = \begin{pmatrix} 1 - \pi_2 & \pi_2 \\ 1 - \pi_2 & \pi_2 \end{pmatrix}.$$

Employing this model under the null it is easy to see that the independence of the hit sequence is tested by this means since the rows have all the same entries. Under the null previous observations do not influence the probability of monitoring a violation. Matrix entries π_2 represent the probability of a violation and according to this the number of observations are aggregated over index j as the past value j has no influence on the present value i, $\pi_2 = \frac{n_{01}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$. Thus,

$$L(\Pi_2) := L(\Pi_2; \{I_t\}) = (1-\pi_2)^{(n_{00}+n_{10})}\pi_2^{n_{01}+n_{11}}$$

indicates the likelihood function under the null model.

Using $L(\Pi_1)$ and $L(\Pi_2)$ the LR test statistic for the Christoffersen test of independence is given by

$$LR.IND = -2 \log\left(\frac{L(\Pi_1)}{L(\Pi_2)}\right)$$

which is χ^2 distributed with one degree of freedom. Note that the Christoffersen [1998] test provides no possibility for testing conditional coverage as LR.IND does not depend on the true coverage probability α. A joint test for both testing the independence and the conditional coverage property as well is provided below.

A problem which arises when using this backtest is that the Christoffersen test of independence only examines for dependence between two consecutive observations. Campbell [2005] notes the possibility that the probability of monitoring a violation today is not influenced by yesterday's observation but indeed could be influenced by prior observations.

Next to the test for proving independence of observations of the hit sequence Christoffersen [1998] introduced a test of unconditional coverage, testing $E[I_t] = \alpha$ against its alternative $E[I_t] \neq \alpha$. The joint test of conditional coverage and independence by Christoffersen [1998] combines those tests to examine whether both properties of a VaR measure are jointly fulfilled.

The basic idea is as simple as for the independence test: First, if the unconditional coverage property is fulfilled, then $\frac{n_{00}+n_{10}}{n_{00}+n_{01}+n_{10}+n_{11}} = \alpha$ must hold implying that the proportion of observed violations matches with the hit probability α. Furthermore, as stated previously, the probability of a non-violation following a previous hit equals the probability of a non-violation following a previous non-violation, i.e. $\frac{n_{00}}{n_{00}+n_{01}} = \frac{n_{10}}{n_{10}+n_{11}}$, when the independence property is on hand. Combining this, if the VaR measure fulfils the independence property, these probabilities should match the total proportion of non-violations. Thus, provided the unconditional property is valid, this leads to

$$\frac{n_{00}}{n_{00}+n_{10}} = \frac{n_{10}}{n_{10}+n_{11}} = \frac{n_{00}+n_{01}}{n_{00}+n_{01}+n_{10}+n_{11}} = \alpha$$

which denotes the tested hypothesis under the null. In terms of the LR framework the likelihood of the null of the unconditional coverage test is tested here against the alternative of the independence test, forming a test of conditional coverage in effect. Thus, the test statistics results in

$$LR.CC = -2 \log \left(\frac{L(\alpha)}{L(\Pi_1)} \right).$$

Christoffersen [1998] shows that the limiting distribution of the joint test is $\chi^2(2)$. However, even if running a joint test might seem always preferable over running the unconditional coverage test and the independence test separately, one has to note that joint tests dismiss VaR measures that violate only one property. As a result, the joint test may detect the violation of either unconditional coverage or independence in less cases than a test which covers only one of these properties. According to Campbell [2005] the employment of a test which comprises only a sole property might be preferable when prior information about the VaR measure is available.

Escanciano and Olmo [2010] provide a test of unconditional coverage as well as a test of conditional coverage. Their analysis bases on a Monte Carlo study, where the unconditional and the conditional coverage tests are compared to a corrected version of these tests. The corrected versions account for the impact of estimation risk arising when forecasts are carried out. All tests are based on the demeaned hit sequence $\{I_t - \alpha\}$.

The test of unconditional coverage is derived from the validity of $E[I_t] = \alpha$ under the null model. Its test statistics is presented by

$$S_P = \frac{1}{\sqrt{n}} \sum_{t+R=1}^{P} (I_t - \alpha)$$

and is predicated on the unconditional coverage tests by Kupiec [1995] and Christoffersen [1998]. It can easily be checked that $\frac{1}{\sigma} S_P$ converges against a standard normal distribution, where $\sigma = \sqrt{\alpha\,(1-\alpha)}$ is nothing else than the standard deviation of the binomial distribution for I_t. This holds as S_P marks the standardized version of $\{I_t\}$ with

$$\frac{1}{\sigma\, P^{-\frac{1}{2}}}\, S_P = \frac{\frac{1}{P}\sum_{t+R=1}^{P}(I_t - \alpha)}{\sigma\, P^{-\frac{1}{2}}} = \frac{1}{\sqrt{P}\,\sigma} \sum_{t+R=1}^{P} (I_t - \alpha) \longrightarrow N(0;1).$$

When adjusting σ for estimation risk it can be shown that the term of the estimated standard deviation gets the form

$$\sigma_{corr} = \Big(\alpha\,(1-\alpha) + \pi\, \hat{A}\, \hat{V}\, \hat{A}'\Big)^{-\frac{1}{2}}$$

when the applied forecast scheme is set fixed and the underlying DGP is a GARCH process of order (1,1). Note that Escanciano and Olmo [2010] also provide adjusted tests for rolling and recursive forecast schemes. For $\pi \hat{A}\hat{V}\hat{A}' = 0$ the impact of estimation risk is asymptotically irrelevant.

The parameter $\pi = \lim_{n\to\infty} \frac{P}{R}$ denotes the relation between the length P of the out-of-sample series and the first R observations which are used to estimate the process parameters. It is quiet intuitive that for a large value of R in relation to P (and, thus, a relatively long in-sample series) the influence of estimation risk becomes negligibly small. The matrix V is of dimension (3×3) and contains the variances and covariances of the data generating process, while A denotes a (3×1)-vector containing the first derivations of the DGP wrt the GARCH parameters, respectively, and $\hat{A}$ and $\hat{V}$ denote consistent estimators for A and V. For a detailed derivation of A and V see Appendix. The resulting test statistics

$$\tilde{S}_P = \frac{1}{\sqrt{n}\,\sigma_{corr}} \sum_{t=1}^{n} (I_t - \alpha)$$

follows an $N(0;1)$ distribution for $n \to \infty$.

The leadoff duration-based backtesting approach was proposed by Christoffersen and Pelletier [2004] with the motivation to overcome the pitfall of small power of backtests existing by then in small sample sizes and to uncover not only first order Markov dependencies as given by the independence test by Christoffersen [1998]. This approach is justified by the authors by the existence of no-hit periods which are either relatively short by reason of high market volatility or relatively long when the market is calmed down. For this, we define $d_i = t_i - t_{i-1}, i = 1, \ldots, I$ as the duration between the hit number $i-1$ and i occurring at dates t_{i-1} and t_i $(t \in \{1, \ldots, n\})$, respectively.

To construct the test that emanates from the independence of the durations and thus, from a correct specified VaR model, a memoryless probability distribution is needed to model the durations. The only continuous distribution which accounts for a constant failure probability α is given by the exponential distribution with the density

$$f^{Exp}(d) = \alpha \,\exp(-\alpha\, d).$$

Note that the corresponding hazard function for the exponential distribution is $\lambda^{Exp}(d) = \alpha$ which can be interpreted as the probability of observing a violation at date d after the last hit took place under the condition of having waited for $d-1$ days is constantly α and independent from d, ie memoryless. Thus, the null of independence checks whether the durations d_i come from an exponential distribution with likelihood function

$$\ln L(\alpha) = n \ln(\alpha) - \alpha \bar{d}.$$

For the alternative model a duration distribution with a non-constant hazard rate is required. The simplest case represents the Weibull distribution with density

$$f^W(d) = \alpha^b \, b \, d^{b-1} \, \exp(-(\alpha \, d)^b)$$

where $b \in \mathbb{R}_{>0}$ is a shape parameter. Note that the exponential distribution is nested by the Weibull distribution for $b = 1$. The hazard rate can easily be obtained by

$$\lambda^W(d) = \alpha^b \, b \, d^{b-1}.$$

For $b < 1$ the Weibull hazard rate is decreasing. Transferred to financial market analysis a decreasing λ^W indicates the tendency of the market to more extreme durations, i.e. periods of relatively short or relatively long duration. The log-likelihood function under the alternative is then given by

$$\ln L(\alpha; k) = \ln \lambda + \ln k + (k-1) \sum_i \ln d_i - \lambda \sum_i d_i^k.$$

Thereby, the pair of hypotheses can be reformulated in terms of the shape parameter b by $H_0 : b = 1$ versus $H_1 : b \neq 1$.

The null of independence can be tested by a Likelihood ratio test by evaluation of

$$LR_{Dur} = -2 \, \frac{\ln L(\alpha)}{\ln L(\alpha; b)}$$

which follows a χ^2 distribution with two degrees of freedom.

In order to conduct the test, it is necessary to transform the hit sequence $\{I_t\}$ into a duration sequence $\{d_i\}_{i=1}^{I}$. When enforcing the transformation it has to be kept into account that the first and last duration is possibly censored, ie the duration of the first no-hit period is longer than d_1 as there is no data available before. Of course, the only exception consists in the case that the first observation is already a hit. Likewise, the last duration could be longer than d_I if the last observation of $\{I_t\}$ displays no hit.

In the above spanned framework it is possible to model dependencies of higher order than the Markov-type test. However, this test contains no information about the exact order of dependence, but could only be captured by the EACD framework by Engle and Russell [1998].

Another test of independence that does not exploit the hit sequence directly, but the properties of the durations between consecutive hits, was recently proposed by Candelon et al. [2011]. The major motivation behind the construction of this test is to overcome the drawback of low power in realistic sample sizes when conducting backtests.

The idea behind this test is as follows: To each distribution which belongs to the Pearson family of distributions an orthonormal polynomial can be associated. Orthonormal polynomials build a sequence of polynomials in which each two polynomials are pairwise orthonormal under the L^2-inner product. Considering the duration sequence $\{d_i\}$ as being discrete, the orthonormal polynomial associated with the geometric distribution can be employed.

Define the number of employed polynomials h, the orthonormal polynomial associated to the memoryless geometric distribution follows the recursion

$$M_h = M_{j+1}(d;\alpha) = \frac{(1-\beta)(2j+1)+\beta(j-d+1)}{(j+1)\sqrt{(1-\beta)}}\, M_j(d;\alpha) - \frac{j}{j+1} M_{j-1}(d;\beta)$$

for any $j \in \mathbb{N}_0$, $\forall d \in \mathbb{N}_0$, $d := d_i \, \forall i \in \{1,\dots,I\}$ and initial values $M_{-1}(d;\alpha) = 0$, $M_0(d;\beta) = 1$. Using the method of moments to estimate the parameters of this polynomial regression efficient and consistent estimates can be obtained. Thus, under the null of conditional coverage the moment condition

$$H_0 : E[M_j(d;\alpha)] = 0$$

is tested. Here, the duration sequence follows a geometric distribution with hit probability α, meaning that there is no correlation between two consecutive hits as the geometric distribution provides the only memoryless discrete probability distribution.

In contrast to the duration-based test by Christoffersen and Pelletier [2004], this framework allows to test separately for unconditional coverage and the independence hypothesis. The reasoning is straightforward: As the expectation of a geometric distributed random variable with parameter α is equal to $\frac{1}{\alpha}$, it is easily shown that this is equivalent to the condition for the orthonormal polynomial of order $h = 1$ that is tested under H_0 of unconditional coverage:

$$\begin{aligned} E[M_1(d;\alpha)] &= E\left[\frac{1-\alpha d}{\sqrt{1-\alpha}}\right] \\ &= \frac{1-\alpha\frac{1}{\alpha}}{\sqrt{1-\alpha}} = 0 \qquad \text{for } E\left[d\right] = \frac{1}{\alpha} \end{aligned}$$

The usage of orthonormal polynomials enables to run the test within the GMM framework with known asymptotic covariance matrices. The test statistics employing the polynomial order h is

$$C^G_{CC}(h) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_j(d_i;\alpha)\right)' \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_j(d_i;\alpha)\right)$$

which follows a χ^2 limiting distribution with h degrees of freedom and $j = 1,\dots,h$. Note that for the special case of unconditional coverage and $h = 1$ the test statistics becomes

$$C^G_{CC}(1) = C^G_{UC} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_1(d_i;\alpha)\right)^2 .$$

When presuming that $\{d_t\}$ is continuous the tests are run with the same conditions adjusted for the exponential distribution and its corresponding orthonormal polynomials following the recursion

$$L_h = L_{j+1}(d;\alpha) = \frac{1}{n+1}\left[(2n+1-\alpha d)\, L_j(d;\alpha) - nL_{n-1}(d;\alpha)\right]$$

with initial values $L_{-1} = 1$ and $L_1 = 1 - \alpha d$ and L being polynomials of the Laguerre family. The test statistics for the continuous case and the orthonormal polynomials associated with the exponential distribution is then given by

$$C_{CC}^{Exp}(h) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} L_j(d_i;\alpha)\right)' \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} L_j(d_i;\alpha)\right)$$

which again following a $\chi^2(h)$ distribution under the null.

3 Simulation Study

The following simulation studies aim at detecting the problems arising from conducting backtests with univariate time series. For this purpose we simulate GARCH(1,1) processes

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \theta_0 + \theta_1 Y_{t-1}^2 + \theta_2 \sigma_{t-1}^2. \end{aligned}$$

with parameter vector $\theta' = (\theta_0, \theta_1, \theta_2) = (0.1, 0.1, 0.85)$ and different lengths of in-sample period R and out-of-sample horizon P. The in-sample period with $R = (250, 500, 750, 1000, 1500)$ is used for the estimation of the respective parameters and the out-of-sample period $P = (250, 500, 750, 1000, 1500)$ is used for the evaluation of the backrest. The VaR for the respective series with confidence level of $\alpha = 0.01$ is calculated in the next step. Following this, the hit sequence $\{I_t\}$ is computed. In order to test the accuracy of the VaR computations the test statistics of the aforementioned backtests are calculated. The procedure is replicated 5000 times. Table 1 shows the results of the Monte Carlo study. For each combination of in-sample and out-of-sample length, the respective empirical size is calculated from the computed test statistics and the nominal coverage is chosen as amounting to $\alpha = 0.05$. The first three columns summarize the results for the Kupiec test and the tests suggested by Christoffersen (independence and conditional coverage test), while the remaining columns show size results for duration-based backtests for which the sequence $\{d_t\}$ of the time span between the respective hits of sequence $\{I_t\}$ has been taken into account. While tests (4) to (6) are based on the null of a geometric distribution with $h = 1, 3, 5$, tests (7) to (9) report the results for the tests where the distribution under the null is supposed to be continuous with the same number of orthogonal polynomials as under the discrete assumption.

	P	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
R=250	250	0.0930	0.0322	0.0808	0.0486	0.0512	0.0334	0.0138	0.0134	0.0118
	500	0.2240	0.0428	0.1208	0.1758	0.1020	0.0730	0.0344	0.0390	0.0366
	750	0.2262	0.0578	0.1832	0.1840	0.1696	0.1392	0.0718	0.0746	0.0660
	1,000	0.2786	0.0684	0.2286	0.2396	0.2016	0.1660	0.0962	0.0952	0.0816
	1,500	0.3452	0.0756	0.3148	0.3454	0.2828	0.2426	0.1472	0.1458	0.1224
R=500	250	0.0664	0.0328	0.0622	0.0350	0.0388	0.0246	0.0066	0.0080	0.0072
	500	0.1682	0.0412	0.0802	0.1250	0.0682	0.0468	0.0224	0.0270	0.0250
	750	0.1612	0.0640	0.1300	0.1198	0.1128	0.0936	0.0470	0.0574	0.0524
	1,000	0.2138	0.0652	0.1712	0.1746	0.1454	0.1192	0.0666	0.0698	0.0600
	1,500	0.2472	0.0694	0.2296	0.2478	0.1834	0.1500	0.0872	0.0854	0.0744
R=750	250	0.0628	0.0368	0.0582	0.0314	0.0348	0.0236	0.0056	0.0064	0.0074
	500	0.1576	0.0414	0.0680	0.1102	0.0610	0.0456	0.0168	0.0234	0.0252
	750	0.1460	0.0605	0.1216	0.1065	0.0998	0.0849	0.0399	0.0514	0.0448
	1,000	0.1973	0.0621	0.1502	0.1581	0.1247	0.1000	0.0523	0.0589	0.0507
	1,500	0.2058	0.0748	0.2104	0.2064	0.1550	0.1260	0.0652	0.0764	0.0628
R=1,000	250	0.2058	0.0748	0.2104	0.2064	0.1550	0.1260	0.0652	0.0764	0.0628
	500	0.1430	0.0424	0.0634	0.1036	0.0556	0.0412	0.0166	0.0222	0.0230
	750	0.1300	0.0556	0.1076	0.0956	0.0918	0.0734	0.0378	0.0466	0.0394
	1,000	0.1678	0.0690	0.1440	0.1366	0.1096	0.0968	0.0568	0.0574	0.0508
	1,500	0.1877	0.0757	0.1941	0.1877	0.1522	0.1208	0.0673	0.0743	0.0625
R=1,500	250	0.1678	0.0690	0.1440	0.1366	0.1096	0.0968	0.0568	0.0574	0.0508
	500	0.1404	0.0378	0.0624	0.1000	0.0534	0.0384	0.0160	0.0224	0.0236
	750	0.1206	0.0620	0.1058	0.0890	0.0844	0.0674	0.0316	0.0402	0.0358
	1,000	0.1486	0.0604	0.1188	0.1152	0.0952	0.0822	0.0444	0.0494	0.0434
	1,500	0.1652	0.0752	0.1856	0.1656	0.1318	0.1062	0.0622	0.0678	0.0558

Table 1: Results - Size, $\alpha = 0.01$

The first observation to be made is that the majority of the backtests are oversized and hence reject the null too often. Thus, even if the null is true, the backtests classify the VaR to be inaccurate. However, some of the duration-based backtests tend to be undersized especially if P and R are both small. Secondly, the smaller the ratio $\pi = P/R$ of out-of-sample length to in-sample length, the lower is the distortion, that is the difference between the empirical and nominal size. For example, for $R = 250$ the Kupiec test is distorted by 29.52% for $P = 1,500$ and the lower the in-sample period the smaller is the distortion. When the out-of-sample length is reduced to $P = 250$ the size is distorted by 4.3%. This is due to the reason that the smaller the amount of data available for estimation of parameters in comparison to P the higher is the estimation risk involved which leads to less accurate projections of VaR. Duration-based backtests tend to have lower size distortions in general.

Acknowledging model risk, Escanciano and Olmo [2010] provided tests corrected for estimation risk. When correcting the variance of the backtest by Kupiec and taking into account the demeaned hit sequence $\{I_t\}$ the test should not be rejected as often as is the case with the uncorrected test. Therefore, it should be expected that the size distortions decrease by applying the estimation risk corrected backtest by Escanciano and Olmo [2010]. We again conduct a Monte Carlo experiment as outlined above with 500 replications and $R, P = (250, 500, 750, 1000)$ and computed S_P and $\tilde{S}_P$. Size results are reported in Table 2.

	R = 250				R = 500			
P	250	500	750	1,000	250	500	750	1,000
S_P	0.138	0.182	0.250	0.268	0.108	0.154	0.228	0.194
$\tilde{S}_P$	0.088	0.096	0.082	0.118	0.074	0.078	0.092	0.074

	R = 750				R = 1,000			
P	250	500	750	1,000	250	500	750	1,000
S_P	0.128	0.142	0.228	0.184	0.100	0.090	0.180	0.156
$\tilde{S}_P$	0.090	0.098	0.084	0.064	0.084	0.062	0.078	0.084

Table 2: Results

For each combination of R and P the effect of the variance correction results in a much lower empirical coverage for $\tilde{S}_P$ and for low π empirical and nominal coverage do hardly deviate from each other.

In Figure 1, the density of the true asymptotic distribution of S_P and $\tilde{S}_P$, ie the normal distribution, as well as the kernel density estimation of the test statistic S_P as well as $\tilde{S}_P$ of the corrected test for $R = 250$ and $P = 500$ and $\alpha = 0.05$ are plotted. Whereas the density of S_P deviates considerably from its asymptotic distribution, the kernel density of the corrected backtest comes much closer to it.

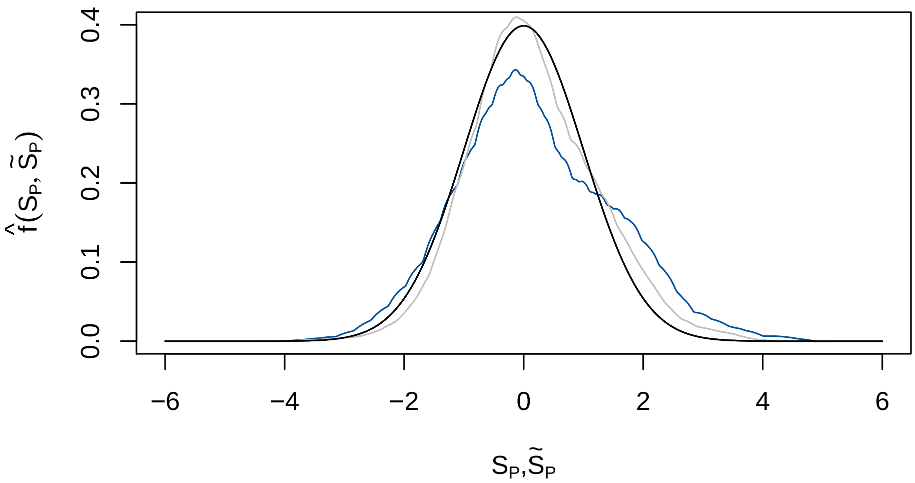

Figure 1: Density of normal distribution ($\mu = 0$, $\sigma = 1$) (black), Kernel density estimate of S_P (blue), Kernel density estimate of $\tilde{S}_P$ (gray) for $R = 250$, $P = 500$ and $\alpha = 0.05$

However, for the Basel II relevant period length of $R = 250$ and the VaR level of $\alpha = 0.01$ size distortions remain at a considerable level of about 3%. The problem therefore remains that the test rejects too often. Looking at the size distortions of the tests proposed by Escanciano and Olmo [2010] we can see that even when accounting for estimation risk the problem prevails. In their follow-up paper for including misspecification risk in their backtest, Escanciano and Olmo [2011] acknowledge that their modified test still suffers from problems of high size distortions also in case of very small in-sample lengths. To put it in a nutshell, all classes of univariate backtests proposed (although duration-based backtests to a lesser extent) have problems when it comes to short in-sample horizons.

Although the corrected backtests result in a reduction of the size distortion, the tests tend to reject too often. Even though the correction for estimation risk has been conducted the problem especially prevails in the Basel II scenario for $R = 250$ and VaR confidence level of $\alpha = 0.01$. In this set-up duration-based backtests with orthonormal approximation of the distribution under the null seem to be the most promising alternative.

4 Conclusion

In our paper we analyze the problems of backtests that have been suggested so far. Backtests based on hit and duration sequences in a univariate framework show heavy size distortions. A solution for this is to account for model risk and correct the asymptotic variance of the backtest and thereby reduce the distortion. The problems of univariate backtesting resulting in considerable size distortions for the relevant Basel II set-up, however, cannot be alleviated by modifying backtests in a way that accounts for estimation risk or misspecification risk. When financial institutions conduct backtesting, they face restrictions from the regulation side where the in-sample length is set to $R = 250$. A reduction of the out-of-sample length does not suffice to reduce the empirical size. Using inaccurate backtests has severe implications and higher risk-based capital results as the factor for its calculation of directly linked to the number of hits.

A solution suggested by Danciulescu [2010] as well as Berkowitz et al. [2009] is to conduct multivariate backtesting as a mean to overcome these problems. They argue that the sample size is thereby increased and information is more efficiently used for this purpose. In our Monte Carlo study backtests based on orthonormal polynomials performed best. Extending these backtest in a multivariate surrounding would therefore be an alternative to the common approaches. Backtesting with multivariate orthonormal polynomials includes the assumption that under the null the duration sequences follow a respective discrete or continuous multivariate distribution and that this distribution is approximated by Laguerre polynomials in the continuous case. The idea of multivariate backtesting with Laguerre polynomials is a topic to be pursued in further research.

References

BCBS. Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. Technical report, Basel Committee on Banking Supervision, 1996.

J. Berkowitz, P.F. Christoffersen, and D. Pelletier. Evaluating value-at-risk models with desk-level data. *Management Science, Articles in Advance*, 31:1–15, 2009.

S.D. Campbell. A review of backtesting and backtesting procedures. *Finance and Economics Discussion Series, Federal Reserve Board*, 21, 2005.

B. Candelon, G. Colletaz, C. Hurlin, and S. Tokpavi. Backtesting value-at-risk: A gmm duration-based test. *Journal of Financial Econometrics*, 9:314–343, 2011.

P. F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39:841–862, 1998.

P. F. Christoffersen and D. Pelletier. Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2:84–108, 2004.

C. Danciulescu. Backtesting value-at-risk models: A multivariate approach. Technical report, Indiana University, Center for Applied Economics & Policy Research Working Paper No. 004-2010, 2010.

R.F. Engle and J. Russell. Econometric analysis of discrete-valued irregularly-spaced financial transactions data using a new autoregressive conditional multinomial model. Technical report, University of California San Diego Dept. of Economics, Discussion paper no. 98-10, 1998.

J.C. Escanciano and J. Olmo. Estimation risk effects on backtesting for parametric value-at-risk models. Technical report, Indiana University, Center for Applied Economics and Policy Research, 2007.

J.C. Escanciano and J. Olmo. Backtesting parametric value-at-risk with estimation risk. *Journal of Business and Economis Statistics*, 28:36–51, 2010.

J.C. Escanciano and J. Olmo. Robust backtesting tests for value-at-risk models. *Journal of Financial Econometrics*, 9:132–161, 2011.

C. Francq and J.-M. Zakoïan. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10:605–637, 2004.

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

J.A. Lopez. Methods for evaluating value-at-risk estimates. Technical report, Federal Reserve Bank of San Francisco, Economic Review, 1999.

Appendix

Quasi-Maximum-Likelihood estimation of GARCH(1,1)

As in Francq and Zakoïan [2004] and Escanciano and Olmo [2007].

Model is a pure GARCH(1,1) $Y_t = \mu + \sigma_t \varepsilon_t$ with $\sigma_t^2 = \theta_0 + \theta_1 Y_{t-1}^2 + \theta_2 \sigma_{t-1}^2$ with $\mu = 0$, innovation $\varepsilon_t = Y_t / \sigma_t \overset{iid}{\sim} t(\nu)$ and parameter vector $\theta = (\theta_0, \theta_1, \theta_2)$.

Asymptotic normality of QMLE:

$$\sqrt{T}(\hat{\theta} - \theta)' \xrightarrow{d} N(0, V)$$
$$V = J^{-1} I J^{-1}$$

Conditional Gaussian quasi-log-likelihood:

$$L = \sum \frac{1}{\sqrt{2\pi\sigma_t^2}} exp\left(-\frac{Y_t^2 - \mu}{2\sigma_t^2}\right)$$
$$\tilde{l}_t = -\frac{1}{2} log(2\pi) - \frac{1}{2} log(\sigma_t^2) - \frac{1}{2}\frac{Y_t^2}{\sigma_t^2} = -\frac{1}{2}\left\{ log(2\pi) + log(\sigma_t^2) + \frac{Y_t^2}{\sigma_t^2} \right\}$$

Score:

$$\frac{\partial \tilde{l}_t}{\partial \theta} = -\frac{1}{2}\left\{ \frac{\partial (log(\sigma_t^2))}{\partial \theta} + \frac{\partial (\frac{Y_t^2}{\sigma_t^2})}{\partial \theta} \right\} = -\frac{1}{2}\left\{ \frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} - \frac{Y_t^2}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta} \right\}$$
$$= -\frac{1}{2}\left\{ 1 - \frac{Y_t^2}{\sigma_t^2} \right\}\left\{ \frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} \right\} = -\frac{1}{2}\{1 - \varepsilon_t^2\}\left\{ \frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} \right\}$$

Hessian:

$$\frac{\partial^2 \tilde{l}_t}{\partial \theta \partial \theta'} = -\frac{1}{2}\left\{ -Y_t^2 \frac{\partial \sigma_t^{-2}}{\partial \theta}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{\partial \sigma_t^{-2}}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta} + \frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right)\right\}$$
$$= -\frac{1}{2}\left\{ -Y_t^2 \frac{\partial \sigma_t^{-2}}{\partial \theta}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{\partial \sigma_t^{-2}}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta}\right) + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right)\right\}$$
$$= -\frac{1}{2}\left\{ \frac{\partial \sigma_t^{-2}}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta}\left[-\frac{Y_t^2}{\sigma_t^2} + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\right] + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right)\right\}$$
$$= -\frac{1}{2}\left\{ -\frac{1}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta}\left(1 - 2\frac{Y_t^2}{\sigma_t^2}\right) + \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right)\right\}$$
$$= -\frac{1}{2}\left\{ \left(1 - \frac{Y_t^2}{\sigma_t^2}\right)\left(\frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right) + \left(2\frac{Y_t^2}{\sigma_t^2} - 1\right)\frac{1}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta}\right\}$$
$$= -\frac{1}{2}\left\{ (1 - \varepsilon_t^2)\left(\frac{1}{\sigma_t^2}\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'}\right) + (2\varepsilon_t^2 - 1)\frac{1}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta}\frac{\partial \sigma_t^2}{\partial \theta}\right\}$$

Expected value of Hessian, J:

$$J = E\left[-\frac{1}{2}\left\{(1-\varepsilon_t^2)\left(\frac{1}{\sigma_t^2}\frac{\partial^2\sigma_t^2}{\partial\theta\partial\theta'}\right) + (2\varepsilon_t^2 - 1)\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta}\right\}\right] = \frac{1}{2}\left\{E\left[\frac{1}{2}(2\varepsilon_t^2 - 1)\right]E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta}\right]\right\}$$

$$= \frac{1}{2}(2E(\varepsilon_t^2) - 1)E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta}\right] = \frac{1}{2}E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]$$

Expected value of squared score, I:

$$I = E\left[-\frac{1}{2}(1-\varepsilon_t^2)\frac{\partial\sigma_t^2}{\partial\theta}\frac{1}{\sigma_t^2}\left(-\frac{1}{2}(1-\varepsilon_t^2)\frac{\partial\sigma_t^2}{\partial\theta}\frac{1}{\sigma_t^2}\right)'\right] = E\left[\frac{1}{4}(1-\varepsilon_t^2)^2\right]E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]$$

$$= \frac{1}{4}(E(\varepsilon_t^4) + 1 - E(\varepsilon_t^2))E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right] = \frac{1}{4}(E(\varepsilon_t^4) - 1)E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]$$

$$= \frac{1}{2}(E(\varepsilon_t^4) - 1)\frac{1}{2}E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right] = \frac{1}{2}(E(\varepsilon_t^4) - 1)J$$

Hence, asymptotic covariance matrix of QMLE, V:

$$V = J^{-1}\frac{1}{2}(E(\varepsilon_t^4) - 1)JJ^{-1} = J^{-1}\frac{1}{2}(E(\varepsilon_t^4) - 1)$$

$$= \frac{1}{2}(E(\varepsilon_t^4) - 1)2\left[E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]\right]^{-1} = (E(\varepsilon_t^4) - 1)E\left[\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]^{-1}$$

Consistent estimate of V:

$$\hat{V} = (\kappa - 1)\left[P^{-1}\sum_{t=R+1}^{n}\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}\right]^{-1}$$

where

$$\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'} = \begin{pmatrix} \psi^2 & \psi\sum_{j=1}^{\infty}\theta_2^{j-1}Y_{t-j}^2 & \psi\sum_{j=1}^{\infty}\theta_2^{j-1}\sigma_{t-j}^2 \\ \psi\sum_{j=1}^{\infty}\theta_2^{j-1}Y_{t-j}^2 & \left(\sum_{j=1}^{\infty}\theta_2^{j-1}Y_{t-j}^2\right)^2 & \sum_{j=1}^{\infty}\theta_2^{j-1}Y_{t-j}^2\sum_{j=1}^{\infty}\theta_2^{j-1}\sigma_{t-j}^2 \\ \psi\sum_{j=1}^{\infty}\theta_2^{j-1}\sigma_{t-j}^2 & \sum_{j=1}^{\infty}\theta_2^{j-1}Y_{t-j}^2\sum_{j=1}^{\infty}\theta_2^{j-1}\sigma_{t-j}^2 & \left(\sum_{j=1}^{\infty}\theta_2^{j-1}\sigma_{t-j}^2\right)^2 \end{pmatrix}$$

with $\psi \equiv (1-\theta_2)^{-1}$ and where κ is the unstandardized kurtosis.

Consistent estimate of A:

$$\hat{A} = f(F_\varepsilon^{-1})F_\varepsilon^{-1}\frac{1}{P}\sum(\frac{1}{\sigma_t}\frac{\partial\sigma_t}{\partial\theta} = f(F_\varepsilon^{-1})F_\varepsilon^{-1}\frac{1}{P}\sum\begin{bmatrix} \frac{1}{2\sigma_t^2(1-\theta} \\ \frac{1}{\sigma_t^2}\sum_{j=1}^{\infty}\theta^{j-1}y_{t-j}^2 \\ \frac{1}{\sigma_t^2}\sum_{j=1}^{\infty}\theta j - 1\sigma_{t-j}^2 \end{bmatrix}$$