Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 10, 1–27 DOI: 10.5604/01.3001.0016.0657
Studia metodologiczne / Methodological studies
Zgłoszony/submitted: 20.12.2021, poprawiony/revised: 26.09.2022, zaakceptowany/accepted: 4.10.2022

Estimation risk taking into consideration the effect of forecasting scheme: robust inference about VaR[1]

Marta Małecka[a]

Abstract. The paper addresses the issue of estimation risk in VaR testing. The occurrence of estimation risk (also called parameter uncertainty) implies that the observed VaR violation process may not fulfil the standard requirements that underpin the testing framework. As a result, VaR tests may reject correct VaR models due to estimation errors committed when predicting the VaR. The paper examines the robustness of VaR tests to estimation risk. The research is based on an observation indicating that certain elements of a forecasting scheme have a significant influence on estimation risk. Thus, the article extends the previous studies to include several more realistic forecasting schemes than those based solely on a fixed window.

The aim of the research is twofold: firstly, to find methods of mitigating the negative impact of estimation risk on VaR tests, and secondly, to provide a comprehensive comparison of VaR testing methods with reference to the issue of estimation risk. The conducted analyses demonstrate that a proper adjustment of the forecasting scheme yields better results in terms of the accuracy of the tests than correcting estimation errors by means of the subsampling technique.

Keywords: VaR tests, estimation risk, parameter uncertainty
JEL: C12, C52, C53, G17

Ryzyko estymacyjne uwzględniające schemat prognozowania – wnioskowanie o VaR za pomocą metod odpornych

Streszczenie. Artykuł dotyczy problemu ryzyka estymacyjnego przy testowaniu VaR. Występowanie ryzyka estymacyjnego (zwanego również niepewnością parametrów) oznacza, że obserwowany proces przekroczeń VaR może nie spełniać standardowych wymogów określających ramy testowe. W konsekwencji testy VaR mogą odrzucać prawidłowe modele VaR ze względu na błędy estymacji popełnione podczas wyznaczania prognoz VaR. W badaniu omawianym w artykule oceniana jest odporność testów VaR na ryzyko estymacyjne. U podstaw badania leży spostrzeżenie, że ryzyko estymacyjne w istotny sposób zależy od elementów schematu prognozowania. Z tego powodu w badaniu uwzględniono schematy prognozowania bardziej realistyczne niż schemat oparty na ustalonym oknie, co stanowi rozszerzenie w stosunku do wcześniej prowadzonych badań.

Cel badania jest dwojaki: znalezienie metod, które pozwalałyby zniwelować negatywny wpływ ryzyka estymacji na testy VaR, oraz kompleksowe porównanie metod testowania VaR

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji Multivariate Statistical Analysis MSA 2021, która odbyła się w dniach 8–10 listopada 2021 r. w Łodzi. / The article is based on a paper delivered at the Multivariate Statistical Analysis MSA 2021 conference, held on 8–10 November 2021 in Łódź.

[a] Uniwersytet Łódzki, Katedra Metod Statystycznych, Polska / University of Lodz, Department of Statistical Methods, Poland. ORCID: https://orcid.org/0000-0003-4465-9811. E-mail: marta.malecka@uni.lodz.pl.

w odniesieniu do problemu ryzyka estymacyjnego. Przeprowadzone analizy wskazują m.in. na to, że odpowiednie dostosowanie schematu prognozowania daje lepsze wyniki pod względem dokładności testów niż korygowanie błędów estymacji techniką podpróbkowania.

Słowa kluczowe: testy VaR, ryzyko estymacyjne, niepewność parametrów

1. Introduction

According to the Basel III and Basel IV accords,[2] the current global banking regulations recommend including VaR (Value-at-Risk) tests in banks' internal risk management systems. As a result, the VaR testing framework continues to be an important issue discussed in the financial and statistical literature. Recent studies on this subject have revealed a new methodological problem – the inaccuracy of testing methods due to estimation risk. For example, Escanciano and Olmo (2010) found that estimation risk changes the size of VaR tests. The size, together with the power, are two fundamental properties of a statistical test. They ensure that proper models are not overrejected (i.e. rejected more often than indicated by the chosen significance) and incorrect models are efficiently detected.

The influence of estimation risk on the size of VaR tests results from distortions of the VaR violation process, underpinning the backtesting procedures. These distortions occur because the violation process is based on the estimated, not the actual VaR. On the other hand, test statistic distributions typically rely on the violation process based on the actual VaR. As a consequence of this inconsistency, the VaR models may be rejected not only due to their incorrectness, but also as a result of estimation errors inherent in these models. Intuitively, the scale of estimation risk heavily depends on the forecasting scheme applied to produce VaR forecasts. The components of these forecasting schemes, such as the window choice or frequency of parameter correction, may potentially reduce the estimation errors and their negative impact on subsequent statistical inference. This issue, however, has not yet been studied. The existing proposals as to how to deal with the problem of size distortions caused by estimation risk (Escanciano & Olmo, 2010, 2011) do not investigate forecasting schemes. Instead, they are based on a fixed forecasting scheme and suggest solving this problem by using resampled distributions. According to these proposals, resampled distributions should replace asymptotic distributions. They should be obtained in a way that takes into account estimation risk. Our approach to this problem does not resort to resampling, but focuses on forecasting schemes instead.

In order to overcome the negative impact of estimation risk, we study this problem in the context of various realistic forecasting schemes. We design a Monte Carlo (MC) study that includes the rolling and recursive scheme along with the fixed one. This implies the need to carry out nested simulations, which substantially

[2] These accords contain supervisory recommendations issued by the Basel Committee on Banking Supervision (BCBS), which is the international committee setting prudential regulations of banks. It has 45 members including central banks and bank supervisors from 28 jurisdictions.

increase the computational demands of the study; nevertheless it is essential for drawing practical conclusions. We demonstrate that when estimation risk occurs, the careful selection of the forecasting scheme is of particular importance, since it has critical influence on the properties of the test. The results of our research lead to different conclusions than those presented in previous studies, as we provide evidence that adjusting the forecasting scheme yields better results in terms of test accuracy than correcting the estimation error by means of resampling techniques. Thus, we argue that the tests may be effectively applied with the use of asymptotic distributions instead of time-consuming and computationally-intensive simulation methods. Our proposal, however, should involve performing parameter corrections in VaR models at a suitable frequency. These conclusions are of great practical importance considering the supervisory recommendation to incorporate VaR tests in the internal procedures of banks.

Our approach is also novel because it encompasses VaR tests belonging to various classes[3] in one study. By doing so, it expands the range of the earlier comparisons of VaR testing methods provided by Berkowitz et al. (2011) and Pajhede (2017),[4] who did not address the problem of estimation risk. Our study examines the estimation risk effects for VaR tests belonging to six distinct classes. The first class encompasses unconditional coverage tests, represented by the pioneering procedure by Kupiec (1995). This test is important to the banking industry, as it establishes a procedure which underpins international regulatory rules (BCBS, 2017). The remaining five distinct classes of conditional coverage tests are: the Markovian, correlation-based, regression-based, duration-based and spectral tests. In the first group of the above-mentioned tests, the early Markov-chain Christoffersen's one was the leading reference from 1998 until 2017, when it was generalised by Pajhede (2017). Our study, therefore, is based on the generalised Pajhede version of this test. Out of the group of the correlation-based procedures, we focus on the test of the Ljung-Box-type, which draws directly upon the autocorrelation function of violations. This approach is particularly important, as, since Berkowitz et al. (2011) proposed using it for the evaluation of VaR models, it has been commonly treated as a benchmark in empirical studies. Out of the group of regression-based VaR tests, we chose the binary-choice procedure developed by Dumitrescu et al. (2012), who generalised and extended the early test by Engle and Manganelli (2004). To represent the class of the duration-based VaR tests, we selected the geometric-VaR test by Pelletier and Wei (2016). This method generalises the geometric approach proposed by Christoffersen

[3] We limit the scope of our study to univariate VaR tests. Previous research also proposed to employ the multivariate approach, i.e. to use multiple VaR levels (Colletaz et al., 2013; Gordy & McNeil, 2018; Hurlin & Tokpavi, 2006; Kratz et al., 2018; Leccadito et al., 2014; Wied et al., 2016) as a natural extension to these methods.

[4] The recent overview of VaR testing methods provided by Pajhede (2017) encompassed the Markov-chain class tests, autocorrelation tests, regression-based tests and duration-based tests; however, neither did it include the generalisation of the early regression-based test of Dumitrescu et al. (2012) nor the generalisation of the duration-based test of Pelletier and Wei (2016). It also did not include any of the spectral methods.

and Pelletier (2004) and follows a number of other VaR testing procedures related to the geometric distribution (Berkowitz et al., 2011; Candelon et al., 2011; Engle & Russel, 1998; Haas, 2005; Krämer & Wied, 2015; Ziggel et al., 2014). The spectral methods, following research by Małecka (2016), are represented in our study by the Anderson-Darling VaR test, which draws upon the spectral approach proposed by Berkowitz et al. (2011).

The aim of our study is twofold: firstly, to find methods of mitigating the negative impact of the estimation risk on VaR tests, and secondly, to provide a comprehensive comparison of VaR testing methods with reference to the issue of estimation risk. The key element of the study, which is an extension of previous research on estimation risk in VaR tests, is the inclusion of various realistic forecasting schemes.

2. Literature review

The notion of estimation risk in the VaR context (also called the model risk) was defined by Escanciano and Olmo (2010). Prior to their work, studies on testing VaR centred around establishing criteria for distinguishing correct VaR models and proposals of how to verify these criteria. The main achievements in defining these economically relevant, testable criteria are attributed to Kupiec (1995), who proposed the verification of the VaR violation probability, and Christoffersen (1998), who argued that the probability of violating VaR should not only coincide with the chosen VaR coverage level (unconditional coverage criterion – VaR level) but also be constant in time (conditional coverage criterion). The criteria put forward by Kupiec and Christoffersen were accompanied by proposals of tests embedded in the likelihood ratio and first-order Markov-chain framework. These pioneering tests were followed by numerous other proposals of how to verify the unconditional coverage and independence criteria. Among them, there were proposals to use standard approaches like the correlation-based Ljung-Box test or the regression-based tests developed by Dumitrescu et al. (2012) and Engle and Manganelli (2004). The Markov-chain approach with higher-order dependencies was advocated by Pajhede (2017). A large group of the methods proposed as the follow-up to the pioneering tests mentioned before can be classified as duration-based tests. They use the transformation of VaR violations into durations and include tests proposed by Berkowitz et al. (2011), Candelon et al. (2011), Christoffersen and Pelletier (2004), Engle and Russel (1998), Haas (2005), Krämer and Wied (2015), Pelletier and Wei (2016) and Ziggel et al. (2014).

A different perspective was adopted towards spectral tests, which are based on the Fourier transformation of the autocorrelation function. Such an approach was used to test VaR by Berkowitz et al. (2011) and Gordy and McNeil (2018). An important development of these methods were multi-level and multivariate tests by Berkowitz (2001), Colletaz et al. (2013), Hurlin and Tokpavi (2006), Kratz et al. (2018), Leccadito et al. (2014) and Wied et al. (2016). These competing VaR testing procedures were assessed on the basis of two fundamental statistical properties – the size and the power, which is the standard convention.

Very few papers have taken into account the problem of size distortions caused by estimation risk reported by Escanciano and Olmo (2010). The suggestions of how to handle this problem evolved around simulation methods based on resampling, as initiated by the study of Escanciano and Olmo (2011). They suggested simulation methods are 'natural simpler alternatives' to the asymptotic theory. The authors argued that asymptotic distributions incorporating estimation risk for VaR tests encounter technical problems which could be overcome through robust resampling techniques: the block bootstrap and subsampling. According to their research, these techniques may be used to approximate the actual sampling distributions of VaR test statistics. Their study showed that resampled distributions have an advantage over theoretical ones; however, the research was limited to the assumptions of the fixed forecasting scheme only. Since then, several novel proposals of tests have been presented with regard to estimation risk. Candelon et al. (2011) investigated the influence of estimation risk on their GMM-based VaR test and the possibility of reducing it by implementing subsampling. Dumitrescu et al. (2012) included the bootstrapping technique instead of subsampling as a remedy to the problem of estimation risk occurring in the dynamic-binary-choice VaR test. Pelletier and Wei (2016), who recently proposed an extension to the geometric duration-based VaR test, indicated that the issue of estimation risk should be subject to further studies. The general conclusion from these studies is that the effects of estimation risk drastically decrease the accuracy of tests (where accuracy is understood as a test size compliant with the selected significance level). The subsampling and the block-bootstrap methods reduce this negative influence. However, even when corrected with the use of these methods, the true test size is clearly distinct from the chosen significance level.

The studies above employed subsampling or block-bootstrapping as a means to addressing the problem of estimation risk, as recommended by Escanciano and Olmo (2011). That research was based on the fixed forecasting scheme. The fixed forecasting scheme assumes that parameters are estimated only once from a fixed number of initial observations. We broaden the previous studies by waiving this assumption and including several commonly used forecasting schemes. We show that the proper adjustment of the analysed components produces better effects in dealing with estimation risk than the previously proposed methods.

3. Estimation errors in VaR tests

The criteria developed for backtesting VaR models are based on the properties of the violation process, which compares the forecasted VaR with the realised returns. To formally define these criteria, let us denote R_t as the continuous return from the portfolio at time t, Ω_t is the set of all information at time t, and $\mathcal{F}_t$ defines the σ-algebra generated by the subsets of Ω_t. The p%-VaR $q_p(R_t)$ is the conditional p-quantile of R_t, i.e. $\mathbb{P}(R_t < q_p(R_t)|\mathcal{F}_t) = p$. Let us assume that VaR forecasts are produced from the parametric model $\mathcal{M} = \{VaR_{t|t-1,p}(\theta): \theta \in \Theta \subset \mathbb{R}^n\}$. The violation process is defined by:

$$I_{t,\theta}(p) = \mathbb{I}\{R_t < VaR_{t|t-1,p}(\theta)\}, \tag{1}$$

where $\mathbb{I}_A(x) = 1$ if $x \in A$, and $\mathbb{I}_A(x) = 0$ if $x \notin A$.

Equation (1) clearly demonstrates that the violation process and its properties which give rise to the VaR testing criteria are heavily dependent on parameter vector θ. The nuisance parameters in θ introduce estimation risk to the backtesting framework, which tends to be neglected in standard backtesting procedures and thus requires verification. As noted by Escanciano and Olmo (2010), these procedures are typically based on the simplifying assumption that such $\theta^\star \in \Theta$ exists that $VaR_{t|t-1,p}(\theta^\star) = q_p(R_t)$. Moreover, it is usually assumed that $\theta^\star$ is known. These assumptions may be jointly written as follows:

$$VaR_{t|t-1,p}(\theta) = VaR_{t|t-1,p}(\theta^\star) = q_p(R_t). \tag{2}$$

According to (2) and provided that the VaR model is correct, the underlying violation process shown in (1) reduces to:

$$I_t(p) = \mathbb{I}\{R_t < q_p(R_t)\}. \tag{3}$$

Based on the process described in (3), the VaR testing criteria are formulated as

$$\mathbb{P}(I_{t,\theta}(p) = 1) = p \tag{4}$$

and

$$\mathbb{P}(I_{t,\theta}(p) = 1|\mathcal{F}_{t-1}) = \mathbb{P}(I_{t,\theta}(p) = 1), \tag{5}$$

which are called the unconditional coverage and independence criterion, respectively. Jointly, these criteria require the conditional probability of violation to be constant and equal to p, i.e.

$$\mathbb{P}(I_{t,\theta}(p) = 1|\mathcal{F}_{t-1}) = p, \tag{6}$$

which is called the conditional coverage criterion.

Conditions (4)–(6) hold true only when (3) is the true violation process. In a realistic setting, however, process (3) is unobservable, while the observed VaR violations are the realisations of (1). These realisations correspond to the out-of-sample VaR forecasts for the observed return series. Assuming series of a length of T, i.e. $\{R_1, \dots, R_T\}$, the beginning R observations are used only to estimate model $\mathcal{M}$ and produce forecasts for the remaining period. The forecasted VaR series $\{VaR_{R+1|R,p}(\theta), \dots, VaR_{T|T-1,p}(\theta)\}$ of the $P = T - R$ length, compared with the P out-of-sample returns $\{R_{R+1}, \dots, R_T\}$, forms observed violation series $\{I_{R+1,\theta}(p), \dots, I_{T,\theta}(p)\}$, where criteria (4)–(6) do not hold true. However, following the standard convention, the asymptotic distributions of the VaR test statistics are based on the assumption that this violation sample is the series of realisations of (3) instead of (1). This may cause the results of the backtesting to be misleading. Also, the typical studies that evaluate the fundamental statistical properties of test statistics – the size and the power – are based on simulations that erroneously assume that, under the correct model, (3) is the violation process.

Two ways of dealing with the problem of estimation risk were proposed by Escanciano and Olmo (2011). The first method involved the development of an asymptotic theory that incorporated the influence of the estimation risk on the evaluation of risk models. Based on this theory, the correct asymptotic distributions of two VaR tests were derived: one dedicated to the unconditional coverage and the other to the conditional coverage hypothesis. However, the authors noted that such distributions may be difficult to derive analytically for other available testing procedures. Moreover, the quick inflow of new, generalised and extended tests makes such an analytical approach infeasible. In order to address these issues, Escanciano and Olmo suggested that the problem of estimation risk be resolved by the employment of simulation methods. They proposed two resampling techniques: subsampling and the block bootstrap to approximate the true sampling distribution of the test statistics.

The idea of the subsampling approach consists in using a subsampling approximation of the actual statistic distribution. This approximation is formed in a way that mimics the actual processes of estimating, forecasting VaR and testing. The approximate distribution is based on values of the test statistic computed on smaller subsets of data – subsamples. The s-th subsample from return data $\{R_1, R_2, \dots, R_T\}$ is defined as $\{R_s, R_{s+1}, \dots, R_{s+u-1}\}$, where $s \in \{1, \dots, T-u+1\}$ and u is the subsample size. Then these subsamples are divided into two parts that correspond to the division of the initial sample into the in-sample part of length R and the out-of-sample part of length P. Consequently, each subsample has the in-sample part of size R_u and the out-of-sample part of size P_u, where $\frac{P_u}{R_u} = \frac{P}{R}$. The in-sample parts are used for estimating model $\mathcal{M}$ and the out-of-sample parts to compute the realisations of the test statistic. This technique was adopted by Candelon et al. (2011), who studied the effects of the estimation risk on their GMM VaR tests.

As regards the bootstrap approach, the sampling distribution of the test statistic derives from estimating model $\mathcal{M}$ on the basis of resamples of size R drawn with the replacement from the standardised residual series. These residuals are obtained with the use of the parameter estimates from the in-sample parts $\{R_1, R_2, \dots, R_R\}$. Following the resampling phase, the residuals are transformed into bootstrap samples $\{R_1^B, R_2^B, \dots, R_R^B\}$, which produce bootstrap parameter estimates and out-of-sample bootstrap VaR forecasts. In effect, a new violation series is formed and used for testing. A sufficient number of repetitions yields series of realisations of the test statistic, which approximates its distribution. Such a technique was applied in the study by Dumitrescu et al. (2012), where the dynamic binary choice VaR test was developed.

The studies by Candelon et al. (2011) and Dumitrescu et at. (2012) led to two main conclusions: test accuracy drastically deteriorates when estimation errors occur, and resampling methods reduce the effects of the estimation errors. However, even corrected with the help of these methods, the true test size was clearly distinct from the chosen significance level. An important limitation to these studies from the practical point of view was the simplifying assumption that VaR forecasts were obtained from the fixed forecasting scheme. This scheme takes for granted that parameters are estimated only once from a fixed number of beginning observations. We ignore this assumption and investigate the influence of estimation errors under several commonly used forecasting schemes.

The forecasting schemes we analyse include three main types: fixed, rolling and recursive. While the fixed scheme relies on the same parameter estimates throughout the whole sample, the other schemes involve parameter corrections, which are made at a given frequency. To reflect these components in the violation process, we adopt the following notation:

- $\{I_{t,\theta^*}(p)\}$ – the violation process based on the true VaR;
- $\{I_{t,\theta_f}(p)\}$ – the violation process based on VaR predictions made by a fixed forecasting scheme;
- $\{I_{t,\theta_{rol}^d}(p)\}$ – the violation process based on VaR predictions made by a rolling forecasting scheme with a parameter correction made every d days;
- $\{I_{t,\theta_{rec}^d}(p)\}$ – the violation process based on VaR predictions made by a recursive forecasting scheme with a parameter correction made every d days.

The analysis of the above forecasting schemes is directly related to the first aim of our study – to find ways of coping with the problem of estimation risk. As opposed to the previous studies, our research examines the possibilities of reducing the effects of estimation risk by adjusting the components of the forecasting scheme. In order to indicate such possibilities, we study the influence of the forecasting scheme components on the accuracy of the VaR test. We compare our outcomes with the results of the estimation errors correction through the previously proposed resampling methods.

Our research extends the previous studies on the undertaken subject also in the sense that it includes VaR tests belonging to six distinct classes, which allows for a broad comparison of these methods. In addition to the unconditional coverage test that underpins the international regulatory framework, we examine joint conditional coverage tests. Within this group, we first select the representatives of the testing procedures that directly refer to the correlation structure of the violation process. We consider methods developed in the standard Markov-chain convention or with the use of the autocorrelation function estimates. We then include indirect methods using the regression-based or the duration-based approach. Finally, we examine methods that were developed with the use of the spectral theory. Whenever available, we use the generalised versions of the test statistics.

4. Backtesting framework

Depending on the class, VaR backtests either refer to conditions (4)–(6) directly or operate on some implications of these conditions. They are most often designed in a way that allows tests of joint hypothesis (6). This hypothesis is typically reduced to a simplified criterion:

$$\mathbb{P}\big(I_{t,\theta}(p) = 1 \big| I_{t-1,\theta}(p), I_{t-2,\theta}(p), \dots\big) = p, \tag{7}$$

which is a special case of (6). This criterion is equivalent to the requirement that the $\{I_{t,\theta}(p)\}$ process is i.i.d.[5] Bernoulli with parameter p. While recent VaR tests generally correspond with the conditional coverage criterion in either of the two forms: (6) or (7), the pioneer VaR test by Kupiec (1995) was dedicated to checking unconditional coverage (4) exclusively. This test is still important for two reasons: firstly, it continues to be used by the international regulator and secondly, it serves to complement more complex procedures that verify VaR violation independence but have low power against incorrect unconditional coverage. This test checks (4) through the likelihood ratio statistic:

$$LR_{uc}^{K} = -2(P_1\log(p) + (P - P_1)\log(1-p) - P_1\log(\hat{p}) - (P - P_1)\log(1-\hat{p})), \tag{8}$$

[5] Independent and identically distributed.

where P_1 is the number of VaR violations in the out-of-sample P returns $\{R_{R+1}, R_2, \dots, R_T\}$ and $\hat{p}$ is the ML estimate of the violation probability given by $\hat{p} = \frac{P_1}{P}$.

The Markov-chain framework proposed by Christoffersen (1998), which extends the LR_{uc}^{K} approach to a joint test of the conditional coverage hypothesis, operates on first-order dependencies. To improve its power, it was generalised by Pajhede (2017) into a generalised Markov-chain VaR test. This test is based on two types of generalised probabilities: the excited p_E and the steady p_S. They describe the probability of VaR violation in $I_{t,\theta}(p)$ on condition that the previous violation has occurred (p_E) or has not occurred (p_S) in $\{I_{t-1,\theta}(p), I_{t-2,\theta}(p), \dots, I_{t-h,\theta}(p)\}$. The ML estimates of these probabilities are $\hat{p}_E = T_{11}/(T_{10} + T_{11})$ and $\hat{p}_S = T_{01}/(T_{00} + T_{01})$, where $P_{11} = \sum_{t=R+1}^{T} I_t J_{t-1}$, $P_{01} = \sum_{t=R+1}^{T} I_t (1 - J_{t-1})$, $P_{10} = \sum_{t=R+1}^{T} (1 - I_t) J_{t-1}$, $P_{00} = \sum_{t=R+1}^{T} (1 - I_t)(1 - J_{t-1})$, and $J_{t-1} =$ $= 1\{\sum_{i=1}^{h} I_{t-i} > 0\}$.[6] The estimates of the excited and the steady probabilities are used to build the generalised Markov-chain likelihood ratio test statistic:

$$\begin{aligned} LR_{cc}^{M,gen} = -2\big((T_{01} + T_{11})\log(p) + (T_{00} + T_{10})\log(1 - p) + \\ -T_{11}\log(\hat{p}_E) - T_{10}\log(1 - \hat{p}_E) - T_{01}\log(\hat{p}_S) - T_{00}\log(1 - \hat{p}_S)\big). \end{aligned} \tag{9}$$

Another direct way to conduct VaR backtesting on a custom number of dependencies between violations is through sample autocorrelations:

$$\rho(k) = \frac{1}{P - k} \sum_{t=R+k+1}^{T} \frac{(I_{t,\theta}(p) - p)(I_{t-k,\theta}(p) - p)}{p(1 - p)}. \tag{10}$$

This idea was utilised for example by Berkowitz et al. (2011), who proposed a VaR correlation test in the spirit of Ljung and Box (1978). This test uses the statistic in the following form:

$$LB_{cc} = P(P + 2) \sum_{k=1}^{m} \frac{(\rho(k))^2}{P - k}, \tag{11}$$

where m is a chosen lag order.

[6] This corrects formula (18) from Pajhede (2017), where $J_{t-1} = 1\{\sum_{i=1}^{k} I_{t-1} > 0\}$, which in our notation takes the form of $J_{t-1} = 1\{\sum_{i=1}^{h} I_{t-i} > 0\}$.

The regression-based approach, which constitutes another class of tests, is built on the observation that under (6) the conditional expectation of demeaned process $I_{t,\theta}(p) - p$ is equal to zero, i.e.

$$\mathbb{E}(I_{t,\theta}(p) - p|\mathcal{F}_{t-1}) = 0. \tag{12}$$

This expectation can be modelled by linear regression. In that case restriction (12) can be verified by means of testing the significance of the regression parameters. However, the binary character of the dependent variable implies that residuals from such a regression are heteroscedastic. Thus, the early proposition based on simple linear regression is replaced with tests that use the binary choice models. In the generalised test, proposed by Dumitrescu et al. (2012), the dynamic binary response model

$$\mathbb{P}(I_{t,\theta}(p) = 1|\mathcal{F}_t) = E(I_{t,\theta}(p)|\mathcal{F}_t) = F(\pi_t) \tag{13}$$

uses c.d.f.[7] F (usually Gaussian or exponential) and index π_t, which is given by the autoregressive equation. The general form of its representation is as follows:

$$\begin{aligned}\pi_t = \eta + \sum_{i=1}^{q_1} \lambda_i \pi_{t-i} + \sum_{i=1}^{q_2} \phi_i I_{t-i,\theta}(p) + \sum_{i=1}^{q_3} \psi_i l(x_{t-i}) + \\ + \sum_{i=1}^{q_4} \gamma_i l(x_{t-i}) I_{t-i,\theta}(p),\end{aligned} \tag{14}$$

where l links the variables from Ω_t with π_t and q_1, q_2, q_3 and q_4 are chosen lag orders. The restrictions on the parameters, corresponding to (12) are tested through the likelihood ratio of the form:

$$\begin{aligned}LR_{cc}^{db} = -2(\log L\big(\{I_{R+1,\theta}(p), \dots, I_{T,\theta}(p)\}\big|\hat{\eta}, \hat{\lambda}, \hat{\phi}, \hat{\psi}, \hat{\gamma}\big) + \\ - \log L(\{I_{R+1,\theta}(p), \dots, I_{T,\theta}(p)\}|\eta = F^{-1}(p), \lambda = \phi = \psi = \gamma = 0)),\end{aligned} \tag{15}$$

where $\log L(\{I_{R+1,\theta}(p), \dots, I_{T,\theta}(p)\}|\hat{\eta}, \hat{\lambda}, \hat{\delta}, \hat{\psi}, \hat{\gamma}) = \sum_{t=R+1}^{T} (I_{t,\theta}(p) \log F(\pi_t) + - (1 - I_{t,\theta}(p)) \log (1 - F(\pi_t)))$, $\lambda = (\lambda_1, \dots, \lambda_{q_1})$, $\phi = (\phi_1, \dots, \phi_{q_2})$, $\psi = (\psi_1, \dots, \psi_{q_3})$, $\gamma = (\gamma_1, \dots, \gamma_{q_4})$.

Another way of verifying criterion (6) is through the duration-based framework. It transforms violation process $\{I_{t,\theta}(p)\}$ into a $\{D_{j,\theta}(p)\}$ process of durations and requires that the latter process follows the geometric distribution with parameter p. The transformation into durations is defined as:

[7] Cumulative distribution function.

$$D_{j,\theta}(p) = t_j - t_{j-1}, \tag{16}$$

where t_j denotes the time of the j-th VaR violation in process $\{I_{t,\theta}(p)\}$. The generalised geometric-VaR test by Pelletier and Wei (2016) confronts the hazard function of the geometric distribution with the following general form of a hazard function:

$$\lambda_d^j = ad^{b-1}e^{-cVaR_{t_{j-1}+d}}, \tag{17}$$

where $0 \leq a < 1$, $0 \leq b \leq 1$, $c \geq 0$. The relevant restrictions on its parameters are verified through the likelihood ratio:

$$\begin{gathered} LR_{cc}^{geom,VaR} = -2(\log L\big(\{D_{1,\theta}(p), \dots, D_{N,\theta}(p)\}\big|\hat{a}, \hat{b}, \hat{c}\big) + \\ - \log L(\{D_{1,\theta}(p), \dots, D_{N,\theta}(p)\}|a = p, b = 1, c = 0)), \end{gathered} \tag{18}$$

where
$\log L\big(\{D_{1,\theta}(p), \dots, D_{N,\theta}(p)\}|a, b\big) = C_1 \log S\left(D_{1,\theta}(p)\right) + (1 - C_1) \log f\left(D_{1,\theta}(p)\right) +$
$+ \sum_{j=2}^{N-1} \log f(D_{j,\theta}(p)) + C_N \log S(D_{N,\theta}(p)) + (1 - C_N) \log f(D_{N,\theta}(p))$, $f(d) =$
$= (1 - \lambda_1^j)(1 - \lambda_2^j) \dots (1 - \lambda_{d-1}^j)\lambda_d^j$, $S(d) = 1 - (1 - \lambda_1^j)(1 - \lambda_2^j) \dots (1 - \lambda_{d-1}^j)$,
$\{D_{1,\theta}(p), \dots, D_{N,\theta}(p)\}$ is the sample of durations and $\{C_1, \dots, C_N\}$ indicates censoring.

A different transformation of the $\{I_{t,\theta}(p)\}$ process is used within the spectral VaR testing framework. These methods rely on spectral density – the spectral transform of the autocovariance function of the violation process. Since under the i.i.d. Bernoulli violations, this spectral transform is a flat function, the tests consist in comparing the observed spectral density with the theoretical flat shape. The Anderson-Darling test statistic proposed by Małecka (2016) is given by:

$$SD_{cc}^{AD} = \int_0^1 \frac{U(t)^2}{t(1-t)} dt, \tag{19}$$

where $U(t)$ is the function that measures the distance between the estimated and theoretical spectral density. It is given by:

$$\begin{gathered} U_P(t) = \sqrt{2P} \int_0^{\pi t} \left(\frac{\mathcal{P}_P(\omega)}{\sigma(0)} - \frac{1}{2\pi}\right) d\omega = \frac{\sqrt{2}}{\pi} \sum_{k=1}^{P-1} \sqrt{P} \rho(k) \frac{\sin k\pi t}{k}, \\ t \in [0, 1], \end{gathered} \tag{20}$$

where $\mathcal{P}_P(\omega) = \frac{1}{2\pi}\sum_{k=-(P-1)}^{P-1} \sigma(k)e^{-ik\omega}$ is the periodogram estimate of the spectral density, $\sigma(k)$ and $\rho(k)$, respectively, are the estimates of the k-th order autocovariances and autocorrelations of the violation process.

Under standard conditions, the likelihood ratio statistics are asymptotically χ^2 distributed with the number of degrees of freedom corresponding to the number of tested restrictions. Therefore, the Kupiec LR_{uc}^{K}, the Markov-chain $LR_{cc}^{M,gen}$, the Ljung-Box LB_{cc} and the dynamic binary LR_{cc}^{db} tests use χ_1^2, χ_2^2, χ_k^2 and $\chi^2_{q_1+q_2+q_3+q_4+1}$ distributions, respectively. The theoretical distribution of the duration-based statistic diverges from the standard asymptotic likelihood ratio distribution due to the specific boundary case. It has been shown to be a mixture distribution of form $0.25\chi_1^2 + 0.5\chi_2^2 + 0.25\chi_3^2$ (Małecka, 2016). The SD_{cc}^{AD} test is based on the Anderson-Darling distribution (Anderson & Darling, 1952).

5. Robust inference

5.1. Size of VaR tests

The standard VaR backtesting procedures are based on the theoretical distributions given in Section 4. Whether such tests guarantee a reliable classification of VaR models depends on several aspects. Firstly, tests based on these distributions need to be well-sized. Here, by ‘well-sized’ we understand well-sized under the absence of estimation risk. This means that, under the absence of estimation risk, the frequency of rejecting correct models should coincide with the chosen significance level. Since all the considered distributions are based on limit theorems, this property may be violated as a result of applying asymptotic properties to finite samples. Secondly, the tests need to be robust to estimation risk. As argued in Section 3, the convergence of the test statistics to their theoretical distributions may be affected by this additional factor. The estimation risk results from the noise in the violation process that follows from the procedure of estimating and forecasting VaR with the effect that proper risk models may be overrejected due to estimation errors. Thirdly, the tests should be effective at detecting incorrect models. In particular, they should be powerful both against incorrect unconditional and conditional coverage. Out of these three aspects, we focus on the influence of estimation risk on test accuracy. However, to properly combine all the elements of the testing procedure, the evaluation of the effects of the estimation risk is preceded by a size study that assumes the absence of such a risk. This study serves two purposes: it indicates the tests which have the potential to be well-sized in finite samples when based on asymptotic distributions, and it provides benchmark size estimates for a further robustness check. The tests that do not meet the requirements of this preliminary size assessment are excluded from the further phases of the procedure. What then follows is the core part of the process which introduces estimation risk and finally, both parts are complemented by the power evaluation.

The finite sample test sizes are evaluated by means of the MC method based on rejection frequencies observed in simulations which assume correct conditional coverage. To ensure the comparability with previous studies (Candelon et al., 2011; Dumitrescu et al., 2012; Escanciano & Olmo, 2010, 2011), our simulations are based on the t-GARCH process of the $R_t = \sqrt{h_t}\epsilon_t$ form, where ϵ_t follows Student t distribution t_v and the variance is represented by:

$$h_t = \omega_1 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}^2, \tag{21}$$

with the following parameter values: $\omega_1^* = 0.0001$, $\alpha_1^* = 0.1$, $\beta_1^* = 0.85$ and $v^* = 10$. Therefore, model $\mathcal{M}$ in our study is $\mathcal{M} = \{VaR_{t|t-1,p}(\omega_1, \alpha_1, \beta_1, v)$: $\omega_1, \alpha_1, \beta_1 > 0, \alpha_1 + \beta_1 < 1\}$. The returns generated by the process above are compared with the VaR estimates set to $VaR_{t|t-1,p}(\omega_1^*, \alpha_1^*, \beta_1^*) = \sqrt{h_t}F_{t_{v^*}}^{-1}(p)$, where F_{t_v} denotes the c.d.f. of Student t distribution t_v. The resulting violation process – $\{I_{1,\theta^*}(p), \dots, I_{P,\theta^*}(p)\}$ – is then used to conduct testing procedures. The sample sizes are as follows: $P = 250, 500, 750, 1{,}000, 1{,}250, 1{,}500$ and the standard significance levels: $\alpha = 0.01, 0.05, 0.1$. The VaR coverage level is first set to the typical $p = 5\%$ and subsequently to $p = 1\%$, which corresponds to the current trends in regulatory standards. The rejection frequencies are computed over 10,000 MC repetitions.

The study combines VaR backtesting procedures from various classes. The first procedure is the LR_{uc}^K unconditional coverage test by Kupiec (1995), which continues to be the underlying procedure of the Basel standards (BCBS, 2017). This test is followed by conditional coverage tests belonging to five classes. Wherever available, we use the generalised versions of the test statistics. Thus, from the Markov-chain framework, we apply the generalised $LR_{cc}^{M,gen}$ test by Pajhede (2017). Relying on the primary study of its properties, we select parameters p_E and p_S to be the violation probabilities on condition that the previous violation occurred in the last five observations. The same lag order is chosen in the LB_{cc} Ljung-Box-type test, which is based on Berkowitz et al. (2011). This procedure represents the correlation class of VaR tests. In the LR_{cc}^{db} dynamic binary test, representing the class of regression-based methods, we adopt the following representation of the π_t index: $\pi_t = \eta + \lambda_1 \pi_{t-1}$. This choice follows from the accuracy results presented by Dumitrescu et al. (2012), who developed this test. The broad class of the duration-based methods is represented by the $LR_{cc}^{geom,VaR}$ generalised statistic proposed by Pelletier and Wei (2016), which uses time dependencies in the violation process as well as lagged VaR forecasts as explanatory variables. To represent the spectral methods, we apply the SD_{cc}^{AD} test based on the Anderson-Darling statistic, as advocated by Małecka (2016).

The size results under the absence of estimation risk, presented in Table 1, show substantial differences in test accuracy both among the specific procedures and between the VaR coverage levels. There is clearly less accuracy when operating on the lower, 1% VaR level. For 1% VaR, the only tests that may be treated as satisfactory in terms of size are the LR_{uc}^{K} Kupiec test and the SD_{cc}^{AD} spectral test. In the latter case, this is additionally limited to the highest significance level of 0.1. One more restriction that holds for such a low VaR level is that the sample size should consist of at least 750 observations for both tests. As regards the 5% VaR level, the accuracy of the test improves greatly. In this case, most considered procedures seem to be well-sized at all significance levels. These are the LR_{uc}^{K}, $LR_{cc}^{M,gen}$, LB_{cc}, $LR_{cc}^{geom,VaR}$, and SD_{cc}^{AD} tests. The recommendable sample sizes depend on the type of procedure, but for most tests, they start with 250 or 500 observations. Only the LB_{cc} Ljung-Box and SD_{cc}^{AD} spectral tests seem to require longer samples of e.g. 1,000 or more observations. One clear exception emerges in this general description of the convergence of the true test sizes to nominal levels – the LR_{cc}^{db} dynamic binary test. This procedure appears to be systematically undersized. Its true size most often proves to be below half of the nominal significance level. Due to such a large level of inaccuracy, this test is considered not fit for implementation based on the theoretical distribution. Thus, we exclude it from the further parts of our study. Systematic discrepancies are also observed for the Ljung-Box test. This test, in contrast to the LR_{cc}^{db}, seems to be oversized. However, in this case, the inaccuracies are on a much smaller scale, so we decide to include this test in our research.

As the LR_{uc}^{K}, $LR_{cc}^{M,gen}$, LB_{cc}, $LR_{cc}^{geom,VaR}$, and SD_{cc}^{AD} procedures appear to satisfy the correct-size property under the absence of estimation risk, we treat them as having the potential to be well-sized also under the presence of such a risk. In these circumstances, they may be recommended for use with asymptotic distributions. However, their robustness to estimation errors needs to be checked in relevant simulation experiments, which we carry out in the next part of our study.

Table 1. Size estimates for the VaR tests under the absence of estimation risk by significance levels

Series length	5% VaR					
	LR_{uc}^{K}	$LR_{cc}^{M,gen}$	LB_{cc}	LR_{cc}^{db}	$LR_{cc}^{geom,VaR}$	SD_{cc}^{AD}
0.01						
250	0.0084	0.0101	0.0482	0.0049	0.0097	0.0162
500	0.0122	0.0165	0.0342	0.0050	0.0096	0.0124
750	0.0133	0.0097	0.0207	0.0050	0.0114	0.0095
1,000	0.0098	0.0100	0.0190	0.0046	0.0113	0.0112
1,250	0.0106	0.0101	0.0170	0.0040	0.0122	0.0124
1,500	0.0099	0.0114	0.0168	0.0060	0.0099	0.0106
0.05						
250	0.0582	0.0529	0.1160	0.0309	0.0447	0.0416
500	0.0536	0.0518	0.0810	0.0274	0.0504	0.0403
750	0.0559	0.0538	0.0666	0.0230	0.0574	0.0419
1,000	0.0525	0.0484	0.0664	0.0230	0.0545	0.0538
1,250	0.0433	0.0518	0.0635	0.0230	0.0549	0.0497
1,500	0.0496	0.0534	0.0619	0.0280	0.0471	0.0489
0.1						
250	0.1121	0.1483	0.1446	0.0459	0.0928	0.0692
500	0.1018	0.1180	0.1237	0.0488	0.1067	0.0729
750	0.1146	0.1019	0.1110	0.0570	0.1096	0.0763
1,000	0.1092	0.1028	0.1132	0.0400	0.1033	0.0896
1,250	0.1048	0.0978	0.1107	0.0510	0.1086	0.0947
1,500	0.0924	0.1050	0.1063	0.0490	0.0967	0.1027

(cont.)

Series length	1% VaR					
	LR_{uc}^{K}	$LR_{cc}^{M,gen}$	LB_{cc}	LR_{cc}^{db}	$LR_{cc}^{geom,VaR}$	SD_{cc}^{AD}
0.01						
250	0.0039	0.0037	0.1173	0.0030	0.0010	0.0355
500	0.0042	0.0031	0.1978	0.0089	0.0034	0.0416
750	0.0076	0.0042	0.0599	0.0030	0.0060	0.0449
1,000	0.0120	0.0074	0.0691	0.0070	0.0099	0.0447
1,250	0.0102	0.0068	0.0513	0.0020	0.0091	0.0394
1,500	0.0120	0.0068	0.0593	0.0010	0.0070	0.0368
0.05						
250	0.0142	0.0174	0.1173	0.0088	0.0111	0.0438
500	0.0627	0.0244	0.1978	0.0176	0.0218	0.0589
750	0.0362	0.0259	0.2984	0.0124	0.0383	0.0768
1,000	0.0556	0.0298	0.0883	0.0270	0.0423	0.0800
1,250	0.0663	0.0436	0.1087	0.0150	0.0364	0.0785
1,500	0.0563	0.0374	0.0885	0.0180	0.0372	0.0810
0.1						
250	0.0422	0.0432	0.1173	0.1071	0.0281	0.0611
500	0.0627	0.0761	0.1978	0.0590	0.1025	0.0718
750	0.0974	0.0644	0.2998	0.0382	0.0891	0.0987
1,000	0.1155	0.0864	0.1462	0.0510	0.0715	0.1033
1,250	0.1193	0.0716	0.1259	0.0430	0.0714	0.1091
1,500	0.1248	0.0720	0.1401	0.0410	0.0743	0.1081

Source: author's calculations.

5.2. Estimation risk

Our study of the robustness to estimation risk focuses on the components of the forecasting scheme, i.e. the type of the scheme and the frequency of the parameter correction. To the best of our knowledge, the influence of the forecasting scheme on the VaR testing framework has not yet been examined. The available studies are limited to fixed forecasting schemes only. These papers show the dramatic effects that estimation errors have on test accuracy and advocate the use of the subsampling (Candelon et al., 2011) or bootstrap technique (Dumitrescu et al., 2012) to correct these errors. In interpreting their results, however, two factors are of practical importance. Firstly, such results cannot be generalised, since, by definition, the fixed forecasting scheme creates the largest estimation risk. Secondly, the industry standard requires the adoption of other schemes and the correction of parameters with a predefined frequency. Therefore, our study includes several realistic forecasting schemes.

The MC study of the influence of estimation errors on test accuracy requires generating a T-element sample of the returns from model $\mathcal{M}$, which serves to provide both the in-sample data used for estimating the parameters of $\mathcal{M}$, and the out-of-sample data used to forecast the violation process and conduct the testing. The rejection frequency observed over a sufficiently large number of repetitions during the procedure of estimation and testing gives the approximate test size. Due to the inclusion of the estimation process, such a proxy for the test size involves estimation risk.

In order to include the forecasting scheme effects, we start with a fixed forecasting scheme which assumes that parameters are estimated once, from the R beginning observations. These estimates serve to produce VaR forecasts throughout the remaining P-observation period, where $P = T - R$. Based on the obtained forecasts, we generate violation process $\{I_{R+1,\theta_f}(p), \ldots, I_{T,\theta_f}(p)\}$. Then, we proceed to the rolling scheme, in which the estimation window of length P is moved over the sample and used to produce one-day-ahead forecasts. We apply this scheme at different frequencies of parameter correction: every day, every five days and every 10 days. We obtain the following violation processes: $\{I_{R+1,\theta_{rol}^1}(p), \ldots, I_{T,\theta_{rol}^1}(p)\}$, $\{I_{R+1,\theta_{rol}^5}(p), \ldots, I_{T,\theta_{rol}^5}(p)\}$ and $\{I_{R+1,\theta_{rol}^{10}}(p), \ldots, I_{T,\theta_{rol}^{10}}(p)\}$, respectively, from these procedures. For the recursive scheme, the procedures are analogous, except that the initial estimation window of length P is not moved but extended to include more observations. Everyday, five-day and 10-day parameter corrections serve to generate the following violation processes: $\{I_{R+1,\theta_{rec}^1}(p), \ldots, I_{T,\theta_{rec}^1}(p)\}$, $\{I_{R+1,\theta_{rec}^5}(p), \ldots, I_{T,\theta_{rec}^5}(p)\}$ and $\{I_{R+1,\theta_{rec}^{10}}(p), \ldots, I_{T,\theta_{rec}^{10}}(p)\}$, respectively. Each of these violation processes involves a different level of estimation risk. We assess the impact of this risk on the test size by applying all of the considered VaR tests to all of the generated violation processes during each MC repetition. The size estimates produced from these procedures are called uncorrected size estimates. Despite the fact that the procedures involve parameter corrections, the word 'uncorrected' signifies consistency with the previous studies and refers to the use of standard asymptotic distributions instead of the resampled ones. The uncorrected sizes corresponding to the various forecasting schemes are computed over 5,000 MC repetitions.

A natural extension to our robustness check is the incorporation of the forecasting scheme components which we study as a means of controlling the estimation risk. For this purpose, we compare the scale of the reduction in the estimation risk achieved with the help of the most suitable forecasting scheme with the performance of the subsampling technique,[8] proposed previously to address the issue of estimation risk. The sizes corrected by the subsampling method are obtained by using subsampled approximates of the test statistic distributions. The subsampled distrubutions are generated at each MC repetition. Generating them requires that the subsets of length u, of form $\{R_s, R_{s+1}, \dots, R_{s+u-1}\}$, $s = 1, \dots, T-u+1$ are divided into the in-sample parts of length R_u and the out-of-sample parts of length P_u, so that $\frac{P_u}{R_u} = \frac{P}{R}$ and $R_u + P_u = u$. These subsets are moved across the data, which gives $N_u = T - u + 1$ repetitions. For each $s = 1, \dots, T-u+1$, the following substeps are conducted:[9]

1. estimation of the parameters of model $\mathcal{M}$ from the in-sample part: $\{R_s, R_{s+1}, \dots, R_{s+R_u-1}\}$;
2. producing a VaR violation process: $\{I_{s+R_u}(p), \dots, I_{s+u-1}(p)\}$;
3. computing the test statistic from: $\{I_{s+R_u}(p), \dots, I_{s+u-1}(p)\}$.

The steps above are repeated 1,000 times to receive the subsampling-corrected distribution. With this distribution, we compute the p-value for the test statistics and make test decisions. From a large number of repetitions which involve a whole simulation procedure and together with the nested simulations from substeps 1–3, we can approximate the test sizes into what we call subsampling-corrected sizes. Since such nested simulations are very time-consuming, we set the number of MC repetitions to 1,000. Together with 1,000 repetitions of substeps 1–3, it gives $1{,}000 \cdot 1{,}000 = 1{,}000{,}000$ repetitions.

The study of the effects of estimation risk includes all the previously considered tests except for the LR_{cc}^{db} procedure. This test is rejected as it generates the largest

[8] The choice of the subsampling technique as a representative of the resampling methods was determined by our preliminary study, where this method proved more accurate than the bootstrap one.

[9] A detailed description of the subsampling technique for correcting estimation risk can be found e.g. in Candelon et al. (2011).

size distortions, which show the irrelevance of using its asymptotic distribution in finite samples. Therefore, it does not fulfil our aim of indicating asymptotic tests that are both accurate and robust to estimation risk. We provide the results for a 5% VaR and a 0.05 significance level.

The estimates of the uncorrected and corrected sizes are presented in Table 2. The results for the fixed forecasting scheme confirm the conclusions from Candelon et al. (2011) about the large influence of estimation risk on the accuracy of tests, i.e. estimation errors cause the test accuracy to decrease dramatically. The uncorrected test sizes reach the level of over 0.1 instead of the nominal 0.01 level, 0.3 instead of 0.05 and 0.4 instead of 0.1. These results almost exactly match the scale of influence reported by Candelon et al. (2011).

Table 2. Size estimates of VaR tests under the presence of estimation risk by significance levels

Assumption	LR_{uc}^{K}	$LR_{cc}^{M,gen}$	LB_{cc}	$LR_{cc}^{geom,VaR}$	SD_{cc}^{AD}
0.01					
Absence of estimation risk	0.0098	0.0100	0.0190	0.0113	0.0112
Fixed scheme: uncorrected asymptotic distribution	0.1594	0.1476	0.1052	0.1833	0.0469
subsampling corrected distribution	0.0266	0.0280	0.0300	0.0280	0.0250
Rolling scheme: 10-day parameter correction	0.0037	0.0036	0.0334	0.0244	0.0148
5-day parameter correction	0.0031	0.0033	0.0294	0.0234	0.0133
everyday parameter correction	0.0035	0.0039	0.0240	0.0211	0.0091
Recursive scheme: 10-day parameter correction	0.0150	0.0125	0.0271	0.0173	0.0142
5-day parameter correction	0.0141	0.0114	0.0261	0.0179	0.0131
everyday parameter correction	0.0148	0.0107	0.0240	0.0169	0.0114
0.05					
Absence of estimation risk	0.0525	0.0484	0.0664	0.0545	0.0538
Fixed scheme: uncorrected asymptotic distribution	0.2667	0.2664	0.2036	0.3498	0.1261
subsampling corrected distribution	0.0680	0.0650	0.0540	0.1160	0.0620
Rolling scheme: 10-day parameter correction	0.0213	0.0319	0.1206	0.1046	0.0679
5-day parameter correction	0.0201	0.0293	0.1126	0.1035	0.0626
everyday parameter correction	0.0199	0.0284	0.1007	0.0997	0.0544
Recursive scheme: 10-day parameter correction	0.0607	0.0585	0.0951	0.0847	0.0597
5-day parameter correction	0.0594	0.0561	0.0922	0.0846	0.0566
everyday parameter correction	0.0584	0.0561	0.0880	0.0836	0.0547
0.1					
Absence of estimation risk	0.1092	0.1028	0.1132	0.1033	0.0896
Fixed scheme: uncorrected asymptotic distribution	0.3393	0.3564	0.2683	0.4610	0.1932
subsampling corrected distribution	0.0833	0.1100	0.0800	0.2161	0.0762
Rolling scheme: 10-day parameter correction	0.0591	0.0716	0.1957	0.1961	0.1251
5-day parameter correction	0.0578	0.0681	0.1841	0.1959	0.1165
everyday parameter correction	0.0585	0.0645	0.1719	0.1925	0.1049
Recursive scheme: 10-day parameter correction	0.1106	0.1111	0.1580	0.1649	0.1127
5-day parameter correction	0.1113	0.1104	0.1550	0.1640	0.1104
everyday parameter correction	0.1115	0.1077	0.1475	0.1659	0.1059

Note. All size results in the table assume testing a 5% VaR. The results indicating the absence of estimation risk have been provided for the purpose of comparison.

Source: author's calculations.

The results also confirm that the subsampling technique reduces the effects of the estimation errors. The match between the subsampling-corrected sizes and the nominal levels depends on the adopted procedure. However, looking at all the procedures, the corrected sizes fit in a 0.02–0.03 interval for the nominal level of 0.01, in 0.05–0.11 for 0.05, and in 0.08–0.21 for 0.1. These size results are closer to the desired values than the uncorrected ones, but still, the quality of the match may be questioned during practical applications.

From the point of view of our aims, the key conclusions refer to the effects that various forecasting schemes have on estimation risk. The inclusion of the rolling and recursive schemes into the studied processes significantly changes the influence of estimation risk on the test sizes. This effect is large enough to alter the conclusions presented in the previous studies, advocating for the application of resampling methods. Our research demonstrates that in most cases, the employment of a properly selected forecasting scheme results in better-sized tests than the use of a subsampling technique. In two cases – for $LR_{cc}^{geom,VaR}$ and SD_{cc}^{AD} – any forecasting scheme with a parameter correction, regardless of whether it is rolling or recursive, yields better results in terms of the test size than the subsampling method. For LR_{uc}^{K} and $LR_{cc}^{M,gen}$, the recursive scheme is recommended. This scheme takes full advantage of the available sample size for estimating the parameters, so the sample length is crucial in these cases. For these two tests, the results obtained with the help of the recursive scheme at any frequency of parameter correction are more accurate than the ones coming from the subsampled distributions. As regards the improvements obtained in various tests resulting from the changes of the forecasting scheme, the largest gains in test accuracy occur along a shift from a fixed forecasting scheme to a rolling or recursive scheme. The differences then between daily, five-day or 10-day corrections are systematic although rather minor.

The only test where subsampling outperforms testing based on asymptotic distribution and where the parameter corrections in any of the schemes do not reduce this effect is the Ljung-Box LB_{cc} test. This confirms our results from Section 5.1 of our study of the test size, where it was examined assuming the absence of estimation risk and assuming the test being systematically oversized. After changing the forecasting scheme, this effect is still present. Therefore, we conclude that the use of asymptotic distribution is not recommended for this test.

Our conclusions about the influence of the parameter corrections on reducing the negative effects of estimation risk are important in terms of the practical applications of VaR tests. We demonstrate that most of the tests may be effectively implemented with asymptotic distributions without the need to apply simulation methods. This translates into the time needed for implementing the testing procedures. The time needed by subsampling for computational purposes is incomparably longer than the time needed for parameter corrections. What is more, the complexity associated with introducing parameter corrections into the estimation procedure is negligible compared to the complexity connected with the implementation of the subsampling technique.

5.3. Power of VaR tests

The power evaluation complements our study of test properties and serves to choose the procedures that are most effective in detecting incorrect VaR models. This part of our study focuses on the tests classified as well-sized and robust to estimation risk under a suitable forecasting scheme. These are: the Kupiec LR_{uc}^{K}, the generalised Markov $LR_{cc}^{M,gen}$, the geometric-VaR $LR_{cc}^{geom,VaR}$, and the spectral SD_{cc}^{AD} tests.

The power study we conduct involves various forms of distortions from the null which correspond to the construction of the considered tests. Since we examine both unconditional and conditional coverage tests, we need two types of alternatives. Moreover, in the class of the conditional coverage tests, some procedures (e.g. the spectral test) are designed mainly with the aim of testing the independence rather than the joint conditional coverage hypothesis. Their classification as conditional coverage tests follows from the inclusion of parameter p in their statistics; however, intuitively, they exhibit low power against the incorrect unconditional coverage. To take into account all these cases, we evaluate the test power in a two-stage simulation study. The first set of experiments involves various violations of the unconditional coverage postulate, while the experiments carried out during the second stage are designed to violate the independence property exclusively (the unconditional coverage and independence properties jointly form the conditional coverage property).

Each stage of the power study consists of several variants characterised by different distances from the null. For experiments corresponding to the unconditional coverage hypothesis, these variants are implemented in a straightforward way through the manipulation of parameter p. This parameter is expressed as $p = \delta p^*$, where p^* denotes the VaR coverage assumed under the null. The considered levels of δ are: $\delta = 0.5,\ 0.75,\ 1.25,\ 1.5$. We report the results for $p^* = 5\%$.

Measuring the distance from the null in experiments that violate the independence property requires us to change the representation of the return process from the t-GARCH to the $Normal$-GARCH. The latter model allows for the explicit measurement of the scale of the violation of the independence property by the scale of the volatility clustering. The volatility clustering, in turn, is quantified by the autocorrelation of the squared returns. Under the normality assumption and the additional assumption of the existence of the fourth moment, the first-order autocorrelation of the squared returns in the GARCH(1,1) process is expressed analytically by $\rho_1 = \alpha_1 + \alpha_1^2\beta_1/(1 - 2\alpha_1\beta_1 - \beta_1^2)$. The existence of the fourth moment can be verified by criterion $(\alpha_1 + \beta_1)^2 + 2\alpha_1^2 < 1$. If this criterion does not hold, the first-order autocorrelation behaves similarly to $\rho_1 = \alpha_1 + \beta_1/3$ (Ding & Granger, 1996), as has already been shown. Using the exact (when available) or approximate expressions for ρ_1, we design the experiments to target the ρ_1 levels: 0.1, 0.2, 0.3 and 0.4. These levels are obtained by adjusting parameter α_1 with ω_1 and β_1 fixed at the same levels as in the size study.[10]

The results of the first stage of the power study which examines the test effectiveness in discovering the incorrect unconditional VaR level are presented in Table 3. They clearly show that all of the tests were effective in completing the abovementioned task apart from one which proves unsuitable for this purpose, i.e. the spectral SD_{cc}^{AD} test. Its poor performance can be explained by the way it is constructed. It is specifically designed for the purpose of testing the independence property, and based on the autocorrelation function. Only the inclusion of parameter p allows this test to be classified as a conditional coverage test. However, it exhibits a feature common to all correlation-based tests, i.e. insensitivity to violations of unconditional moments.[11]

Out of the three tests performing well, one is the unconditional coverage test and the two others the conditional coverage tests. Thus, the two latter tests can potentially handle well both the violations of the unconditional VaR level and the violations of the independence property. According to our results, all three tests exhibit similar efficiency in detecting incorrect VaR levels. All of them show that the true VaR level amounting to 0.75 or 1.25 of the tested level cannot be effectively detected, while all tests become efficient if the true VaR reaches 0.5 or 1.5 of the tested value. Usually samples of 500 observations are sufficiently long, but the sample length of 1,000 gives the power of at least 0.8 in all the cases.

[10] Such experiments enable us to study the power as a function of the distribution parameter. In this way we improve the existing studies, which are based on arbitrarily chosen GARCH-model-based alternatives (single or multiple), without any information about the distance from the null. Although various alternatives that have more complex representations might perform better at describing real financial processes, we believe they are less suitable for the purpose of the power study.

[11] This feature is also exhibited by the Ljung-Box LB_{cc} test; however, we do not discuss it here – we excluded it from this part of the study for other reasons.

Table 3. Estimates of the power of VaR tests against incorrect unconditional coverage

Series length	LR_{uc}^{K}				$LR_{cc}^{M,gen}$			
	distance from the null measured by δ							
	0.5	0.75	1.25	1.5	0.5	0.75	1.25	1.5
250	0.566	0.173	0.163	0.404	0.444	0.119	0.110	0.297
500	0.875	0.307	0.210	0.624	0.736	0.195	0.183	0.547
750	0.959	0.391	0.336	0.833	0.913	0.299	0.253	0.734
1,000	0.992	0.514	0.392	0.896	0.971	0.363	0.320	0.839
1,250	0.998	0.608	0.476	0.951	0.992	0.447	0.382	0.918
1,500	1.000	0.670	0.545	0.978	0.998	0.541	0.460	0.957

(cont.)

Series length	$LR_{cc}^{geom,VaR}$				SD_{cc}^{AD}			
	distance from the null measured by δ							
	0.5	0.75	1.25	1.5	0.5	0.75	1.25	1.5
250	0.508	0.155	0.087	0.254	0.073	0.059	0.048	0.050
500	0.784	0.233	0.148	0.483	0.067	0.056	0.055	0.059
750	0.931	0.345	0.211	0.685	0.061	0.052	0.055	0.059
1,000	0.978	0.413	0.291	0.808	0.055	0.051	0.057	0.055
1,250	0.994	0.504	0.357	0.904	0.052	0.044	0.050	0.052
1,500	0.998	0.572	0.414	0.943	0.047	0.044	0.055	0.057

Note. All power results in the table assume testing a 5% VaR.
Source: author's calculations.

The results of examining the power against violations of the independence property, presented in Table 4, show that all tests admitted to this part of the study seem to cope well with this issue. For example, if the sample size is at least 500, all the tests can detect, with the power of over 0.7, the violations of the independence corresponding to the volatility clustering at the level of 0.2. This level corresponds to the level of 0.2 of the autocorrelation coefficient of the squared returns in the violation process. The Markov $LR_{cc}^{M,gen}$ and the geometric-VaR $LR_{cc}^{geom,VaR}$ tests exhibit outstanding performance with short samples. In their cases, the power of over 0.7 is attainable with the shortest examined samples (250 observations). The spectral SD_{cc}^{AD} test seems to require slightly longer samples, but in this case, the power of 0.7 is attainable starting from 500 observations.

Table 4. Estimates of the power of VaR tests against violations of independence property

Series length	LR_{uc}^{K}				$LR_{cc}^{geom,VaR}$				SD_{cc}^{AD}			
	distance from the null measured by ρ_1											
	0.1	0.2	0.3	0.4	0.5	0.75	1.25	1.5	0.5	0.75	1.25	1.5
250	0.285	0.551	0.715	0.799	0.287	0.594	0.764	0.849	0.282	0.476	0.555	0.579
500	0.410	0.766	0.907	0.959	0.433	0.826	0.948	0.985	0.447	0.726	0.810	0.823
750	0.548	0.902	0.978	0.994	0.584	0.943	0.993	0.999	0.579	0.868	0.922	0.927
1,000	0.624	0.951	0.994	0.999	0.679	0.980	0.999	1.000	0.655	0.928	0.966	0.967
1,250	0.718	0.981	0.999	1.000	0.780	0.994	1.000	1.000	0.741	0.965	0.985	0.987
1,500	0.791	0.992	1.000	1.000	0.836	0.999	1.000	1.000	0.793	0.981	0.993	0.994

Note. All power results in the table assume testing a 5% VaR.
Source: author's calculations.

6. Conclusions

VaR testing procedures are a part of the global banking supervisory system. Originally, these procedures were developed with the use of theoretical asymptotic distributions. However, the approach based on theoretical distributions has been questioned since 2010, when estimation risk was first considered in the context of the VaR testing framework. It was argued that estimation risk disturbed the convergence of the distributions of test statistics to their standard asymptotic distributions. Thus, the replacement of the asymptotic distributions with those obtained by means of resampling methods was proposed as a remedy to the aforementioned issue.

While agreeing with the argument that estimation risk disturbs the convergence to the asymptotic distributions, we raised the question of whether the proposed solution was the best possible. We recalled the results from previous studies which have also been confirmed by our research, demonstrating that the accuracy of the tests corrected with resampling methods was still not satisfactory. We also indicated that all previous studies were based on a simplified assumption that the VaR estimates were obtained by means of the fixed forecasting scheme. This assumption had a large impact on the results, as, by definition, this type of forecasting schemes generate the largest estimation risk. Therefore, we argued that these results could not be generalised.

To examine the ways of dealing with estimation risk, we waived the assumption of the fixed forecasting scheme and studied more realistic ones – the rolling and the recursive schemes. Our results altered the conclusions from the previous studies. We showed that under the more realistic forecasting schemes, the loss of accuracy due to the use of resampled distributions exceeded the loss of accuracy resulting from the occurrence of estimation risk. Therefore, for the sake of accuracy in VaR testing, it is better to rely on asymptotic distributions and minimise the influence of estimation errors by a suitable forecasting scheme than to use resampling methods.

Our approach has two more advantages, which seem important in the practical applications of VaR testing. The methods we proposed are much more time-efficient and do not require designing simulations to conduct a VaR test. They are, therefore, optimal from the point of view of real business operations.

Apart from the above general conclusions about dealing with the issue of estimation risk, our contribution to the undertaken subject also provides a broad comparison of the VaR testing methods, belonging to six distinct classes. To the best of our knowledge, this paper is the first to present such a comparison involving estimation risk. We selected the VaR tests that fulfil three optimality postulates: the tests are well-sized in finite samples when based on asymptotic distributions, they are robust to estimation risk, and the most efficient in detecting incorrect VaR models. Two conditional coverage tests turned out to be most satisfactory in the light of the postulates listed above – the generalised Markov $LR_{cc}^{M,gen}$ and the geometric-VaR $LR_{cc}^{geom,VaR}$ tests. They are both accurately sized and exhibit similar, outstanding results in terms of their power. The $LR_{cc}^{M,gen}$ test prevailed in robustness to estimation risk, thus this procedure proved best.

Both the above-mentioned tests, being of the conditional coverage type, allow for the joint testing of the unconditional coverage and independence postulates. Another procedure appearing to perform exceptionally well in all the aspects was the spectral SD_{cc}^{AD} test. However, it is dedicated to testing the independence property solely and has no power against incorrect unconditional coverage. Therefore, it may be used only together with an unconditional coverage test. An unconditional coverage test that fulfils all our postulates is the LR_{uc}^{K} Kupiec test. Thus, the approach based on the Kupiec and spectral tests seems equivalent to the one based on the generalised Markov procedure. All these procedures may be effectively implemented with the use of asymptotic distributions, provided that VaR forecasting is conducted with a suitable forecasting scheme.

References

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, *23*(2), 193–212. https://doi.org/10.1214/aoms/1177729437.

Basel Committee on Banking Supervision. (2017). *High-level summary of Basel III reforms*. Bank for International Settlements. https://www.bis.org/bcbs/publ/d424_hlsummary.pdf.

Berkowitz, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business & Economic Statistics, 19*(4), 465–474. https://dx.doi.org/10.1198/07350010152596718.

Berkowitz, J., Christoffersen, P., & Pelletier, D. (2011). Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science*, *57*(12), 2213–2227. https://doi.org/10.1287/mnsc.1080.0964.

Candelon, B., Colletaz, G., Hurlin, C., & Tokpavi, S. (2011). Backtesting Value-at-Risk: a GMM duration-based test. *Journal of Financial Econometrics*, *9*(2), 314–343. https://doi.org/10.1093/jjfinec/nbq025.

Christoffersen, P. (1998). Evaluating Interval Forecasts. *International Economic Review*, *39*(4), 841–862. https://doi.org/10.2307/2527341.

Christoffersen, P., & Pelletier, D. (2004). Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics*, *2*(1), 84–108. https://doi.org/10.1093/jjfinec/nbh004.

Colletaz, G., Hurlin, C., & Pérignon, C. (2013). The Risk Map: A new tool for validating risk models. *Journal of Banking & Finance*, *37*(10), 3843–3854. https://doi.org/10.1016/j.jbankfin.2013.06.006.

Ding, Z., & Granger, C. W. J. (1996). Modeling Volatility Persistence of Speculative Returns: A New Approach. *Journal of Econometrics*, *73*(1), 185–215. http://dx.doi.org/10.1016/0304-4076(95)01737-2.

Dumitrescu, E.-I., Hurlin, Ch., & Pham, V. (2012). Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Tests (HAL Working Papers No. halshs-00671658). https://halshs.archives-ouvertes.fr/halshs-00671658/document.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Jornual of Business & Economic Statisitics*, *22*(4), 367–381. https://doi.org/10.1198/073500104000000370.

Engle, R. F., & Russel, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, *66*(5), 1127–1162. https://doi.org/10.2307/2999632.

Escanciano, J. C., & Olmo, J. (2010). Backtesting Parametric Value-at-Risk With Estimation Risk. *Journal of Business and Economic Statistics*, *28*(1), 36–51. https://doi.org/10.1198/jbes.2009.07063.

Escanciano, J. C., & Olmo, J. (2011). Robust Backtesting Tests for Value-at-risk Models. *Journal of Financial Econometrics*, *9*(1), 132–161. https://doi.org/10.1093/jjfinec/nbq021.

Gordy, M. B., & McNeil, A. J. (2018). *Spectral backtests of forecast distributions with application to risk management* (Finance and Economics Discussion Series No. 2018-021). https://doi.org/10.17016/FEDS.2018.021.

Haas, M. (2005). Improved duration-based backtesting of value-at-risk. *Journal of Risk*, *8*(2), 17–38. https://doi.org/10.21314/JOR.2006.128.

Hurlin, Ch., & Tokpavi, S. (2006). Backtesting value-at-risk accuracy: a simple new test. *Journal of Risk*, *9*(2), 19–37. https://doi.org/10.21314/JOR.2007.148.

Kratz, M., Lok, Y. H., & McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, *88*, 393–407. https://doi.org/10.1016/j.jbankfin.2018.01.002.

Krämer, W., & Wied, D. (2015). A simple and focused backtest of value at risk. *Economics Letters*, *137*, 29–31, https://doi.org/10.1016/j.econlet.2015.10.028.

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives, 3*(2), 73–84. https://doi.org/10.3905/jod.1995.407942.

Leccadito, A., Boffelli, S., & Urga, G. (2014). Evaluating the Accuracy of Value-at-Risk Forecasts: New Multilevel Tests. *International Journal of Forecasting*, *30*(2), 206–216. https://doi.org/10.1016/j.ijforecast.2013.07.014.

Ljung, G. M., & Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika, 65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297.

Małecka, M. (2016). Spectral VaR test statistical properties. In M. Papież & S. Śmiech (Eds.), *The 10th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings* (pp. 102–109). Foundation of the Cracow University of Economics.

Pajhede, T. (2017). Backtesting Value-at-Risk: A Generalized Markov Test. *Journal of Forecasting*, *36*(5), 597–613. https://doi.org/10.1002/for.2456.

Pelletier, D., & Wei, W. (2016). The geometric-VaR backtesting method. *Journal of Financial Econometrics*, *14*(4), 725–745. https://doi.org/10.1093/jjfinec/nbv015.

Wied, D., Weiß, G. N. F., & Ziggel, D. (2016). Evaluating Value-at-Risk forecasts: a new set of multivariate backtests. *Journal of Banking & Finance*, *72*, 121–132. https://doi.org/10.1016/j.jbankfin.2016.07.014.

Ziggel, D., Berens, T., Weiß, G. N. F., & Wied, D. (2014). A new set of improved value-at-risk backtests. *Journal of Banking & Finance*, *48*, 29–41. https://doi.org/10.1016/j.jbankfin.2014.07.005.