Journal of Banking and Finance 72 (2016) 121–132

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Evaluating Value-at-Risk forecasts: A new set of multivariate backtests☆

Dominik Wied[a], Gregor N.F. Weiß[b,*], Daniel Ziggel[c]

[a] Universität zu Köln and Technische Universität Dortmund, Meister-Ekkehart-Str. 9, 50923 Köln, Germany
[b] Universität Leipzig and Technische Universität Dortmund, Grimmaische Str. 12, 04109 Leipzig, Germany
[c] FOM Hochschule für Oekonomie & Management, Feldstraße 88, 46535 Dinslaken, Germany

ARTICLE INFO

Article history:
Received 5 February 2016
Accepted 29 July 2016
Available online 3 August 2016

JEL Classification:
C52
C53
C58

Keywords:
Model risk
Multivariate backtesting
Value-at-Risk
Systemic risk

ABSTRACT

We propose two new tests for detecting clustering in multivariate Value-at-Risk (VaR) forecasts. First, we consider CUSUM-tests to detect non-constant expectations in the matrix of VaR-violations. Second, we propose χ^2-tests for detecting cross-sectional and serial dependence in the VaR-forecasts. Moreover, we combine our new backtests with a test of unconditional coverage to yield two new backtests of multivariate conditional coverage. Results from a simulation study underline the usefulness of our new backtests for controlling portfolio risks across a bank's business lines. In an empirical study, we show how our multivariate backtests can be employed by regulators to backtest a banking system.

1. Introduction

Over the past two decades, Value at Risk (VaR) has become the prevalent measure for assessing the risk of financial investments. Its widespread use in banking was recognized under the 1996 Market Risk Amendment to the first Basel Accord which allowed banks to employ internal forecasting models to calculate their required regulatory capital. Since then, VaR has become the industry standard for measuring and managing portfolio risk (not only for banks but also, e.g., for insurance companies due to Solvency II) even though it lacks the desirable property of a coherent risk measure (see Artzner et al., 1999) for non-Gaussian Profit & Loss (P/L) distributions. Consequently, not only regulators but also the firms that use VaR themselves have long been interested in assessing the forecasting accuracy of their VaR-models through formal backtesting. Nowadays, risk measures such as the Expected Shortfall, which explicitly take the amount of losses into account, are of increasing importance. Nevertheless, as these measures are still based on the VaR, appropriate backtesting has not lost its importance. In this paper, we address the highly important task of backtesting several VaR-forecasts of different business lines, sub-portfolios or banks across several points in time. We propose two new multivariate backtests that can be used by both risk managers in individual banks (for backtesting the risk of several business lines) and by regulators (for backtesting a whole banking system).

The backtesting of a VaR-model comprises a comparison of the model's out-of-sample VaR-forecasts and the investment's actual returns. If the investment is a single trading position or a portfolio it yields a univariate time series of VaR-forecasts and VaR-violations. In the last few years, several formal backtests have been proposed in the literature for the case of a univariate sequence of VaR-violations with tests concentrating on the correct number of violations (unconditional coverage, uc in short), the independence of the sample of violations, and both properties at the same time (tests of conditional coverage, *cc* in short) (see, e.g., Kupiec, 1995; Christoffersen, 1998; Berkowitz, 2001; Christoffersen and Pelletier, 2004; Engle and Manganelli, 2004; Haas, 2005; Candelon et al., 2011; Berkowitz et al., 2011; Pelletier and Wei, 2015). Recently, Ziggel et al. (2014) proposed a set of tests that additionally test for identically distributed violations. None of these backtests, however, can be easily extended to the multivariate case

☆ We thank Carol Alexander (the editor) and an anonymous referee for their helpful comments. Financial support by the Collaborative Research Center "Statistical Modeling of Nonlinear Dynamic Processes" (SFB 823, projects A1 and A7) of the German Research Foundation (DFG) is gratefully acknowledged. We thank Robert Löser for calculating the general formulas of Σ_s.

* Corresponding author.
E-mail addresses: dwied@uni-koeln.de (D. Wied), weiss@wifa.uni-leipzig.de (G.N.F. Weiß), daniel.ziggel@quasol.de (D. Ziggel).

http://dx.doi.org/10.1016/j.jbankfin.2016.07.014
0378-4266/

in which VaR-violations might not only be correlated across time but also across business lines.

One motivation for considering multivariate VaR backtesting is that financial institutions are usually interested in forecasts of their trading desk's aggregate P/L distribution in contrast to VaR-forecasts of isolated investments. However, aggregating individual VaR-forecasts often yields biased results as diversification effects between (sub-)portfolios are not adequately modeled. To tackle this problem, multivariate backtests need to account for cross-sectional dependence within the portfolio.[1] While it may also be possible to directly consider VaR for aggregate portfolios (i.e., for univariate additive combinations of different investments), the results of a (univariate) backtest for these always depend on the type of aggregation. Moreover, and more importantly, a multivariate backtest avoids the problem of multiple testing, which arises if each business line is tested separately as prescribed by the regulators. Apart from applications within a single bank, our newly proposed tests should also be of great interest to bank regulators as they allow them to backtest risk forecasts for a set of banks. Our multivariate backtests could thus be used to identify times and sources of systemic risk in a banking sector. Finally, in addition to the practical relevance of our backtests for bankers and regulators, our new multivariate backtests process much more data at once thereby allowing further theoretical applications and improving the tests' power properties.

Despite the importance of multivariate backtesting, only a few papers in the literature deal with this topic with most papers leaving the development of such tests for future research (see, e.g., Berkowitz et al., 2011; Ziggel et al., 2014). To the best of our knowledge, the only exception is Danciulescu (2010) who proposes a multivariate uc and independence test. The test is based on a multivariate Portmanteau statistic of Ljung-Box type that jointly tests for the absence of autocorrelations and cross-correlations in the vector of hits sequences for different business lines. However, to the best of the authors' knowledge, there currently exists no multivariate backtest that explicitly tests for the i.i.d. property (in contrast to the mere independence) of VaR-violations.[2]

In this paper, we suggest new multivariate backtests for clusters in VaR-violations which are easy to implement and have appealing properties under the null and the alternative. Moreover, these tests can easily be extended to cc-versions. We essentially propose two different kinds of tests. First, we consider a CUSUM-test for detecting clusters that are caused by instationarities in the mean of the VaR-violations. To take the multidimensionality of the VaR-violations into account, we use the sums of the violations for different business lines and sub-portfolios for a single day. Second, we consider a χ^2-test for detecting clusters that are caused by cross-sectional and/or serial dependencies within the VaR-violations. Finally, we combine our new backtests with a test of unconditional coverage to yield two new backtests of multivariate conditional coverage. All tests are easy to implement and perform well in simulations. Additionally, all tests work without Monte Carlo simulations or bootstrap approximations. However, there are bootstrap approximations available: The one for the CUSUM-tests serves for making it more robust (which does not seem to be necessary, at least in our simulations), while the one for the χ^2-tests is potentially interesting with respect to the test's software implementation.[3]

The rest of this paper is organized as follows. In Section 2, we introduce the notation and the new multivariate backtests. The performance of the new backtests in finite samples is analyzed in simulations in Section 3. In Section 4, the outline and results of our empirical study are presented. Section 5 concludes.

2. Methodology

In this section, we introduce the notation used throughout the paper. Moreover, we define the desirable properties of VaR-violations and present our new multivariate backtests.

2.1. Notation and VaR-violation properties

First, we shortly discuss the univariate case in order to extend it in the following to a multivariate setting. Let $\{y_t\}_{t=1}^n$ be the observable part of a time series $\{y_t\}_{t\in\mathbb{Z}}$ corresponding to daily observations of the returns on an asset or a portfolio. We are interested in the accuracy of VaR-forecasts. Following Dumitrescu et al. (2012), the ex-ante VaR $VaR_{t|t-1}(p)$ (conditionally on an information set $\mathbb{F}_{t-1}$) is implicitly defined by $Pr(y_t < -VaR_{t|t-1}(p)) = p$, where p is the VaR coverage probability. Note that we follow the actuarial convention of a positive sign for a loss. In practice, the coverage probability p is typically chosen to be either 1% or 5% (see Christoffersen, 1998). This notation implies that information up to time $t-1$ is used to obtain a forecast for time t. Moreover, we define the ex-post indicator variable $I_t(p)$ for a given VaR-forecast $VaR_{t|t-1}(p)$ as

$$I_t(p) = \begin{cases} 0, & \text{if } y_t \geq -VaR_{t|t-1}(p); \\ 1, & \text{if } y_t < -VaR_{t|t-1}(p). \end{cases} \tag{1}$$

If this indicator variable is equal to 1, we will call it a VaR-violation. The indicator variables may depend on additional parameters which are assumed to be known such that there is no estimation error. In practice, this is a reasonable assumption given results by Escanciano and Olmo (2010). These authors show for some particular VaR backtests that, asymptotically, there is no estimation error if one uses a *fixed forecasting scheme* for estimating model parameters.

To backtest a given sequence of VaR-violations, Ziggel et al. (2014) state three desirable properties that the VaR-violation process should possess. First, the VaR-violations are said to have unconditional coverage (uc hereafter) if the probability of a VaR-violation is equal to p on average, i.e.,

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} I_t(p)\right] = p. \tag{2}$$

Second, VaR-violations should possess the i.i.d. property. Otherwise, the sequence $\{I_t(p)\}$ could exhibit clusters of violations. In fact, there are several potential reasons for unexpected temporal occurrences of clustered VaR-violations. On the one hand, $\{I_t(p)\}$ may not be identically distributed and $\mathbb{E}(I_t(p))$ could vary over time. On the other hand, $I_t(p)$ may not be independent of $I_{t-k}(p), \forall k \neq 0$. The hypothesis of i.i.d. VaR-violations holds true if

$$\{I_t(p)\} \overset{i.i.d.}{\sim} \text{Bern}(\tilde{p}), \forall t, \tag{3}$$

where $\tilde{p}$ is an arbitrary probability.

[1] Acknowledging this need to backtest the VaR-forecasts of a bank holistically, the Basel guidelines explicitly demand that a bank should "[...] perform separate backtests on sub-portfolios using daily data on sub-portfolios subject to specific risk." (see Basel Committee on Banking Supervision , BCBS).

[2] Note that there exist some papers that deal with VaR backtests in miscellaneous multivariate settings. However, these backtests use multivariate approaches in order to investigate a univariate time series (see, e.g., Hurlin and Tokpavi, 2007).

[3] For sake of brevity, the bootstrap approximations are positioned in the Appendix.

	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	⋯	Line m
Day 1	0	**1**	**1**	**1**	**1**	0	⋯	0
Day 2	0	0	**1**	0	1	0	⋯	0
Day 3	1	0	**1**	1	0	1	⋯	0
Day 4	0	0	**1**	0	1	0	⋯	1
Day 5	0	1	0	0	1	0	⋯	0
⋮	⋮							⋮
Day n	0	0	1	0	1	0	⋯	0

Fig. 1. Multivariate Value-at-Risk hit matrix. The figure presents a stylized matrix of Value-at-Risk (VaR) violations for m business lines, banks, or sub-portfolios and evaluations for n days. If the realized return in business line i on day j exceeds the corresponding VaR-forecast, the respective entry in the hit matrix is one, and zero otherwise. Stylized clusters of VaR-violations in time (third column) and across business lines (first row) are highlighted.

Finally, the *uc* and i.i.d. properties are combined via $\mathbb{E}[I_t(p) - p|\Omega_{t-1}] = 0$ to the property of conditional coverage (cc hereafter). In detail, a sequence of VaR-forecasts is defined to have correct cc if

$$\{I_t(p)\} \overset{i.i.d.}{\sim} \text{Bern}(p), \forall t. \tag{4}$$

Note that in most related studies in the literature, the uc property is defined slightly differently than it is done in this paper. Moreover, the full i.i.d. hypothesis is not discussed at all, with almost all papers concentrating on the independence property of VaR-violations (see, e.g., Christoffersen, 1998).[4]

At this point, we extend our analysis of VaR-violations to a multivariate setting. To this end, we assume that an m-dimensional time series $\{Y_{t,i}\}_{t=1,i=1}^{n,m}$ of returns exists as well as m sequences of VaR forecasts, $VaR_{t,i|t-1}(p_i)$. We then define the indicator variable $I_{t,i}(p_i)$ as

$$I_{t,i}(p_i) = \begin{cases} 0, & \text{if } Y_{t,i} \geq -VaR_{t,i|t-1}(p_i); \\ 1, & \text{if } Y_{t,i} < -VaR_{t,i|t-1}(p_i). \end{cases} \tag{5}$$

Here, p_i is the VaR coverage probability for sub-portfolio i. Note that p_i is explicitly allowed to vary among different sub-portfolios and we do not need to assume particular values of $p_i, i = 1, \ldots, m$. In each column, the resulting matrix contains information for a single business line, bank or sub-portfolio (corresponding to the 1-dimensional case), while each row represents a single trading day. In Fig. 1, we illustrate a stylized matrix of VaR-violations across time and business lines.

As can easily be seen from Fig. 1, clusters of VaR-violations can occur both across time and across sub-portfolios/business lines/banks. Clusters across time indicate a misspecified VaR model, while clusters across sub-portfolios/business lines/banks indicate low potential for diversification or considerable systemic risk in the banking sector, respectively.

With this preliminary work, we start to define the desirable properties of VaR-violations in the multivariate case. For the uc hypothesis and most uc tests, an extension of the univariate to the multivariate case is straightforward. To this end, one simply needs to study the hit sequences of several business lines simultaneously and stack the series together. As doing so effectively increases the sample size, we expect the tests to have more power than in the univariate setting.

However, in this paper, we are interested in the multivariate distribution of VaR-violations and hence neglect this simple issue. In the present context, the VaR-violations should ideally exhibit no clusters, i.e., neither in time (rows) nor across business lines (columns). Thus, the matrix of VaR-violations should fulfill the following multivariate independence hypothesis:

$$I_{t,i}(p_i) \text{ is independent of } I_{t-k,j}(p_j), \forall t, i, j \text{ and } \forall k \geq 0. \tag{6}$$

Note, as property (6) is very restrictive, the VaR model is not necessarily wrongly calibrated if property 6 is not fulfilled. This is due to the fact that heaped violations in one row (trading day) are, though undesirable, no indicator for an incorrect VaR model. However, it may provide important information concerning diversification, aggregation of risks, and systemic risk in the banking sector. Nevertheless, it is also natural to consider the less restrictive hypothesis

$$I_{t,i}(p_i) \text{ is independent of } I_{t-k,j}(p_j), \forall t, i, j \text{ and } \forall k > 0. \tag{7}$$

Property (7) implies that no information concerning VaR-violations available to the risk manager at the time the VaR is estimated is helpful in forecasting a VaR-violation. Thus, as stated in Berkowitz et al. (2011), past observations from the hit sequence of one business line do not help to predict violations of this or any other business line if the VaR model is correctly specified. In particular, property (7) postulates that lagged violations are not correlated. However, correlations within one row (trading day) are explicitly allowed.

As in the univariate case, one can also define the cc-property in the multivariate setting. Here, properties (6) and (7) are modified to

$$\{I_{t,i}(p_i)\} \overset{i.i.d.}{\sim} \text{Bern}(p_i), \forall t, i. \tag{8}$$

and

$$\mathbb{E}(I_{t,i}(p_i)) = p_i \text{ and } I_{t,i}(p_i) \text{ is independent of } \times I_{t-k,j}(p_j), \quad \forall t, i, j \text{ and } \forall k > 0. \tag{9}$$

Again, property (8) is more restrictive than (9) as correlations within one row (trading day) are not allowed. For properties (6)–(9), we propose χ^2-tests in Section 2.3.

As stated in Ziggel et al. (2014), clusters of VaR-violations could also be caused by other reasons than simply correlation between the violations. To be more precise, the probability of obtaining a VaR-violation may change over time. For example, the risk model could not be suited to incorporate changes from calm market phases to highly volatile bear markets or financial crises, and vice versa. This would in turn lead to clustered VaR-violations regardless of the question whether the violations are independent over time or not. In Section (2.2), we consider CUSUM-tests for such instationarities. To be more precise, we consider the row sums

$$r_t := \sum_{i=1}^{m} I_{t,i}(p_i) \tag{10}$$

and test whether $\mathbb{E}(r_t)$ is constant over time (stationarity hypothesis). More precisely, we test for non-constant expectations caused by changes in $\mathbb{E}(I_{t,i}(p_i))$, resulting in the following hypothesis

$$\mathbb{E}\left[\sum_{i=1}^{m} I_{t,i}(p_i)\right] = c, \quad \forall t, \tag{11}$$

where c is an arbitrary constant. In order to define the cc-property, hypothesis (11) is modified to

$$\mathbb{E}\left[\sum_{i=1}^{m} I_{t,i}(p_i)\right] = \sum_{i=1}^{m} p_i, \quad \forall t. \tag{12}$$

Table 1 summarizes all stated hypotheses and comments on the question which user should be most interested in a risk model having the respective property.

[4] See Ziggel et al. (2014) for a critical discussion of previous treatments of the uc and the independence properties in the literature.

Table 1
Summary of all stated hypotheses and test statistics.

Tested property of violations	H_0	Proposed test	Test statistic & Asymptotic distribution	Main test user(s)
Ind. of all violations Eq. (6)	$\mathbb{E}((I_{t,i}(p_i)-\tilde{p}_i)(I_{t+l,j}(p_j)-\tilde{p}_j))=0$	Ind-m-Test	$C'_{s,n}(\hat{\Sigma}_s^{ind})^{-1}C_{s,n}$ $\chi^2_{\lvert A_s\rvert}$	Regulators Portfolio managers
Ind. of lagged violations Eq. (7)	$\mathbb{E}((I_{t,i}(p_i)-\tilde{p}_i)(I_{t+l,j}(p_j)-\tilde{p}_j))=0$	Ind-m-Test	$C'_{s,n}(\hat{\Sigma}_s^{ind})^{-1}C_{s,n}$ $\chi^2_{\lvert A_s\rvert}$	Risk managers
Ind. of all violations Constant expectations Unconditional coverage Eq. (8)	$\mathbb{E}(f(i,j,l,t))=0$	Ind-m-cc-test	$B'_{s,n}(\hat{\Sigma}_s^{cc})^{-1}B_{s,n}$ $\chi^2_{\lvert A_s\rvert}$	Regulators Portfolio managers
Ind. of lagged violations Constant expectations Unconditional coverage Eq. (9)	$\mathbb{E}(f(i,j,l,t))=0$	Ind-m-cc-test	$B'_{s,n}(\hat{\Sigma}_s^{cc})^{-1}B_{s,n}$ $\chi^2_{\lvert A_s\rvert}$	Risk managers
Constant expectations Eq. (11)	$\mathbb{E}(r_1)=\ldots=\mathbb{E}(r_n)=c$	Stat-m-test	$\max_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left\lvert\sum_{t=1}^{j}r_t-\frac{j}{n}\sum_{t=1}^{n}r_t\right\rvert$ $\sup_{s\in[0,1]}\lvert B(s)\rvert$	Risk managers
Constant expectations Unconditional coverage Eq. (12)	$\mathbb{E}(r_1)=\ldots=\mathbb{E}(r_n)=\sum_{i=1}^{m}p_i$	Stat-m-cc-test	$\max_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left\lvert\sum_{t=1}^{j}r_t-j\sum_{i=1}^{m}p_i\right\rvert$ $\sup_{s\in[0,1]}\lvert W(s)\rvert$	Risk managers

This table summarized the different properties that are tested within the present article together with the corresponding test statistics. Note, the Ind-m-test and the Ind-m-cc-test can be used for several properties depending on the choice of A_s. In addition, the last column of the table comments on which stakehold group should be most interested in a risk model having the respective property that is being tested.

2.2. CUSUM-tests for non-constant expectations

In this subsection, we propose a backtest for non-constant expectations. The formal test problem which corresponds to property (11) is given by

$$H_0^s:\mathbb{E}(r_1)=\ldots=\mathbb{E}(r_n)\ \text{vs.}\ H_1^s:\neg H_0^s,$$

with the row sums $r_1,\ldots,r_n$ being defined as in Eq. (10). While the specific expectations are arbitrary in this test problem, this is different in the test problem which corresponds to property (12):

$$H_0^{s-cc}:\mathbb{E}(r_1)=\ldots=\mathbb{E}(r_n)=\sum_{i=1}^{m}p_i\ \text{vs.}\ H_1^s:\neg H_0^s.$$

Before introducing the test statistics, we impose the following assumption:

Assumption 1. Let r_i be defined as before. Then, we assume

1. $r_1,\ldots,r_n$ are independent.
2. $Var(r_1)=\ldots=Var(r_n)$.

If the VaR model is correctly specified, Assumption 1.1 is a reasonable consequence. Assumption 1.2 may be violated if the cross-sectional dependence between $I_{t,1},\ldots,I_{t,m}$ is not constant over time. We will discuss this issue in detail below. Under Assumption (1) and if either H_0^s or H_0^{s-cc} holds, the row sums fulfill a functional central limit theorem, i.e., the process $(V_n, n\in\mathbb{N})$ with

$$V_n(s)=\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor sn\rfloor}(r_t-\mathbb{E}(r_t)),\ s\in[0,1],$$

converges to a Brownian motion. Then, a suitable test statistic for H_0^{s-cc} is given by $RC_{cc,n}:=D^{-1}C_{cc}$ (RC for "row CUSUM") with

$$C_{cc}:=\max_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{j}r_t-j\sum_{i=1}^{m}p_i\right|.$$

Here, D^2 is the usual variance estimator for independent observations, $D^2=\frac{1}{n}\sum_{t=1}^{n}(r_t-\bar{r})^2$ with $\bar{r}=\frac{1}{n}\sum_{t=1}^{n}r_t$. Then, by means of the continuous mapping theorem we immediately obtain

Theorem 2. *Under H_0^{s-cc} and (Assumption 1), for $n\to\infty$, it holds that $RC_{cc,n}\to_d\sup_{s\in[0,1]}|W(s)|$, where W is a standard Brownian motion.*

With this preliminary work, we get the

Stat-m-cc-test. Reject H_0^{s-cc} whenever $RC_n>q_{1-\alpha,BM}$, where $q_{1-\alpha,BM}$ is the $1-\alpha$-quantile of the distribution of $\sup_{s\in[0,1]}|W(s)|$. The 0.95-quantile is given by 2.241.

For testing H_0^s, we do not consider any fixed values of p_i, but we use the test statistic $RC_{stat,n}:=D^{-1}C_{stat}$ with

$$C_{stat}:=\max_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{j}r_t-\frac{j}{n}\sum_{t=1}^{n}r_t\right|.$$

Then, by means of the continuous mapping theorem, we immediately obtain

Theorem 3. *Under H_0^s and Assumption (1), for $n\to\infty$, it holds that $RC_{stat,n}\to_d\sup_{s\in[0,1]}|B(s)|$, where B is a standard Brownian bridge.*

With this preparatory work, we get the

Stat-m-test. Reject H_0^s whenever $RC_n>q_{1-\alpha,KS}$, where $q_{1-\alpha,KS}$ is the $1-\alpha$-quantile of the Kolmogorov–Smirnov-distribution. The 0.95-quantile is given by 1.358.

It can be shown that both tests are consistent, e.g., if, under the alternative,

$$\mathbb{E}(r_1)=\ldots=\mathbb{E}(r_{\lfloor kn\rfloor})\neq\mathbb{E}(r_{\lfloor kn\rfloor+1})=\ldots=\mathbb{E}(r_n)$$

holds for a $k\in(0,1)$. In this case, it is possible to estimate the location of a change point by the argmax estimator

$$C_{cc}:=\text{argmax}_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{j}r_t-j\sum_{i=1}^{m}p_i\right|$$

or

$$C_{stat}:=\text{argmax}_{j=1,\ldots,n}\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{j}r_t-\frac{j}{n}\sum_{t=1}^{n}r_t\right|,$$

(see Aue and Horváth, 2013).

However, the empirical size is not close to the nominal size if there is either weak serial dependence within the $(r_t, t=1,\ldots,n)$ (such as α-mixing under appropriate conditions as described in e.g. Billingsley, 1968) and/or if the $Var(r_t)$ are not constant over time.[5] The test is not consistent in these cases. For

[5] In fact, it is even desirable that serial dependence is detected because this is a potential reason for clustering. On the other hand, it is desirable to

the case of weak serial dependence, this is an immediate consequence of Slutzky's theorem. However, for this problem, we will present a new χ^2-test for cross-sectional and serial dependence in Section 2.3.

On the other hand, we would like to be robust against non-constant variances. This issue is discussed in detail in Zhou (2013). In particular, Zhou (2013) explicitly derives the limit distribution of a general CUSUM-statistic under the assumption of piecewise local stationarity. Thus, there is a bootstrap approximation available which potentially makes the CUSUM-test more robust to changes in variances.[6] Details can be found in Appendix A.2.1.

2.3. χ^2-tests for cross-sectional and serial dependence

In this subsection, we propose a framework that can be used for testing independence as well as the corresponding cc-hypothesis taking into account arbitrary time lags and business lines. This test is somewhat similar to the test proposed by Danciulescu (2010). The main difference is that we explicitly allow for estimated violation probabilities in each business line and that we make use of explicit expressions for a certain covariance matrix.

Denote with A the set of all triples $(i, j, l), i, j = 1, \dots, m, l = 0, 1, \dots,$ where (i, j) describes a pair of sub-portfolios and l the lag of interest. We consider an arbitrary subset $A_s \subseteq A$ that has to be chosen by the analyst. This choice allows us to verify parts of property (7) and/or property (6). In fact, to verify property (7) as a whole, A_s would have to consist of all triples with $i \leq j$ and $l \geq 1$, while it would have to consist of all triples with $i \leq j$ and $l \geq 0$ for property (6). As the set A would become too large then, further restrictions are necessary. The convention is that we consider lags up to a fixed upper bound K, e.g. $K = 5$, corresponding to one week. Moreover, in the following we separate the cases of serial dependence and cross-sectional dependence, whereas serial dependence approximates property (7) and cross-sectional dependence property (6). With the serial dependence application, one can test for clusters in individual banks; with the cross-sectional dependence application, one can test for diversification effects. Formally, this means that we only consider $i = j$ in the first and $i < j$ and $l = 0$ in the second case. For $K = 1$, we have $A_s = \{(i, i, 1)\}, i = 1, \dots, m,$ in the first and $A_s = \{(i, j, 0)\}, i < j, i, j = 1, \dots, m,$ in the second case. We focus on these two specifications in the following. Of course, other combinations are also possible.[7]

The test problem which corresponds to (7) and (6) is given by

$$H_0^{m-ind} : \mathbb{E}((I_{t,i}(p_i) - \tilde{p}_i)(I_{t+l,j}(p_j) - \tilde{p}_j)) = 0 \text{ for } (i, j, l) \in A_s \text{ and } t = 1, \dots, n - l \text{ and some} \tilde{p}_i := \mathbb{E}(I_{t,i}(p_i)), \tilde{p}_j := \mathbb{E}(I_{t,j}(p_j)) \text{ vs. } H_1^{m-ind} : \neg H_0^{m-ind}.$$

In the above test problem, the expectations of $I_{t,i}(p_i)$ and $I_{t,j}(p_j)$ are arbitrary. If one is also interested in testing for them (i.e., for the correct number of VaR-violations), one can consider a modified test problem for the cc-hypothesis. With $f(i, j, l, t) := (I_{t,i}(p_i) - p_i)(I_{t+l,j}(p_j) - p_j)$ and the desired VaR coverage probabilities p_i and p_j,

$$H_0^{m-cc} : \mathbb{E}(f(i, j, l, t)) = 0 \text{ for } (i, j, l) \in A_s \text{ and } t = 1, \dots, n - l \text{ vs. } H_1^{m-cc} : \neg H_0^{m-cc}.$$

First, we consider the cc-test which is based on the vector

$$B_{s,n} := \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n-l} (I_{t,i}(p_i) - p_i)(I_{t+l,j}(p_j) - p_j) \right)_{(i,j,l) \in A_s}$$

whose dimension is equal to the amount of elements in the set A_s. Under the assumption that the VaR model is correct, one obtains by definition of the covariance that the vectors $f(i, j, l, 1)_{(i,j,l) \in A_s}, \dots, f(i, j, l, n - l)_{(i,j,l) \in A_s}$ are uncorrelated. Moreover, we impose the following assumption:

Assumption 4. Let the notation be as before. Then, $Cov(f(i, j, l, 1)_{(i,j,l) \in A_s}) = \dots = Cov(f(i, j, l, n - l)_{(i,j,l) \in A_s}) =: \Sigma_s$, where Σ_s is a positive definite matrix.

Assumption (4) contains a higher-order stationarity assumption, as well as a regularity assumption on the matrix

$$\Sigma_s = (Cov((I_{1,i_1}(p_{i_1}) - p_{i_1})(I_{1+l_1,j_1}(p_{j_1}) - p_{j_1}), (I_{1,i_2}(p_{i_2}) - p_{i_2})(I_{1+l_2,j_2}(p_{j_2}) - p_{j_2})))_{(i_1,j_1,l_1),(i_2,j_2,l_2) \in A_s}.$$

This matrix can easily be calculated for each given set A_s under H_0^{m-cc}. If, e.g., $m = 2$ and $A_s = \{(1, 2, 0)\}$, it holds that $\Sigma_s = p_1 p_2 - p_1^2 p_2 - p_1 p_2^2 + p_1^2 p_2^2$. If $A_s = \{(1, 1, 1), (2, 2, 1)\}$, it holds

$$\Sigma_s = \begin{pmatrix} p_1^2 - 2p_1^3 + p_1^4 & \rho_{12}^2 \\ \rho_{12}^2 & p_2^2 - 2p_2^3 + p_2^4 \end{pmatrix} \tag{13}$$

with $\rho_{12} = Cov(I_{t,1}(p_1), I_{t,2}(p_2))$. In these situations, Σ_s is for example positive definite for $0 < p_1 = p_2 < 1$ and $\rho_{12} = 0$.

In general, under Assumption (9) and in the situation in which it holds for all triples (i, j, l) that $i \leq j$ and $l \geq 1$, the matrix Σ_s consists of the entries

$$Cov(I_{1,i_1}(p_{i_1}), I_{1,i_2}(p_{i_2}))Cov(I_{1+l_1,j_1}(p_{j_1}), I_{1+l_2,j_2}(p_{j_2}))$$

in the row corresponding to the triple $(i_1, j_1, l_1) \in A_s$ and in the column corresponding to the triple $(i_2, j_2, l_2) \in A_s$. This general expression contains (13) as a special case.

Under Assumption (8) and in the situation in which it holds for all triples (i, j, l) that $i < j$ and $l = 0$,

$$\Sigma_s = diag(p_i p_j - p_i p_j^2 - p_i^2 p_j + p_i^2 p_j^2)_{(i,j,0) \in A_s}.$$

Apart from using explicit expressions for Σ_s, one could use a bootstrap procedure which is described in Appendix A.2.2.

Assumption (4) could be relaxed to the case in which the matrix $\Sigma_s^* := \lim_{n \to \infty} B_{s,n}$ exists, is positive definite and can be suitably estimated. An example for this would be the case in which $Cov(f(i, j, l, t)_{(i,j,l) \in A_s})$ is piecewise constant with a finite number of breaks and positive definite in all parts, which is a special case of the PLS setting discussed in Zhou (2013) and Section 2.2. In this case, the estimators described below are consistent for Σ_s^*.

Under H_0^{m-cc} and Assumption (4), it holds that $B_{s,n} \to_d N(0, \Sigma_s)$, while this quantity diverges if, e.g., under the alternative, $\mathbb{E}(f(i, j, l, t)) = c \neq 0$ for $(i, j, l) \in A_s$. Furthermore, with the continuous mapping theorem,

$$\Sigma_s^{-1/2} B_{s,n} \to_d N(0, I_{|A_s|})$$

and

$$B'_{s,n} \Sigma_s^{-1} B_{s,n} \to_d \chi^2_{|A_s|}.$$

Therefore, a suitable test statistic for the cc-test is given by $T_{s,n}^{m-cc} := B'_{s,n} (\hat{\Sigma}_s^{cc})^{-1} B_{s,n}$. Here, $\hat{\Sigma}_s^{cc}$ is an appropriate estimator of Σ_s. For $A_s = \{(1, 1, 1), (2, 2, 1)\}$, one would replace ρ_{12} with

be robust against time-varying variances. Note that $Var(r_t) = \sum_{i=1}^m Var(I_{t,i}(p)) + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m Cov(I_{t,i}(p), I_{t,j}(p))$. So, it is possible that, under the null hypothesis, the variances of r_t might be time-varying (so that Assumption 1.2 would be violated) only because of time-varying covariances. In this situation, we would like the test to keep its size.

[6] In fact, this bootstrap approximation makes the test robust against serial correlation, too.

[7] It is also possible to use formal methods in order to determine the maximum number of lags, see Bender and Grouven (1993).

Table 2
Simulated rejection probabilities for non-constant expectations.

		Stat-m-test						Ind-m-test					
Panel A: $p = 0.01$													
δ/p		0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
	250	0.04	0.06	0.13	0.28	0.48	0.77	0.09	0.09	0.09	0.11	0.13	0.14
n	500	0.04	0.08	0.25	0.57	0.89	1.00	0.15	0.15	0.17	0.18	0.22	0.25
	1000	0.05	0.14	0.51	0.92	1.00	1.00	0.13	0.15	0.16	0.20	0.23	0.28
	2000	0.05	0.28	0.86	1.00	1.00	1.00	0.12	0.12	0.14	0.18	0.23	0.32
Panel B: $p = 0.05$													
	250	0.04	0.16	0.61	0.97	1.00	1.00	0.06	0.07	0.08	0.14	0.21	0.32
n	500	0.04	0.34	0.94	1.00	1.00	1.00	0.06	0.07	0.09	0.15	0.26	0.47
	1000	0.05	0.65	1.00	1.00	1.00	1.00	0.05	0.06	0.10	0.19	0.38	0.68
	2000	0.05	0.94	1.00	1.00	1.00	1.00	0.05	0.07	0.11	0.26	0.59	0.91

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data with non-constant expectations with $p = 0.01$ (Panel A) and $p = 0.05$ (Panel B).

$\frac{1}{T}\sum_{t=1}^{T} I_{t,1}(p_1)I_{t,2}(p_2) - p_1 p_2$ in Σ_s. Then, we get by the strong law of large numbers and Slutzky's theorem

Theorem 5. *Under* H_0^{m-cc} *and (Assumption 4), for* $n \to \infty$, $T_{s,n}^{m-cc} \to_d \chi^2_{|A_s|}$.

We obtain the

Ind-m-cc-test. Reject H_0^{m-cc} if $T_{s,n}^{m-cc} > q_{1-\alpha,\chi^2}$, where $q_{1-\alpha,\chi^2}$ is the $1-\alpha$-quantile of the χ^2-distribution with $|A_s|$ degrees of freedom.

For testing the ind-property, we opt for a test statistic which is based on the quantity

$$C_{s,n} := \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n-l} (I_{t,i}(p_i) - \hat{p}_i)(I_{t+l,j}(p_j) - \hat{p}_j) \right)_{(i,j,l)\in A_s}.$$

This is essentially the quantity $B_{s,n}$, whereas the summands $-p_i$ and $-p_j$ are replaced with $-\hat{p}_i$ and $-\hat{p}_j$, respectively. Here, $\hat{p}_i$ and $\hat{p}_j$ are the actual measured percentages of VaR-violations, $\hat{p}_i := \frac{1}{n}\sum_{t=1}^{n} I_{t,i}(p_i)$ and $\hat{p}_j := \frac{1}{n}\sum_{t=1}^{n} I_{t,j}(p_j)$. The test statistic is defined as $T_{s,n}^{m-ind} := C'_{s,n}(\hat{\Sigma}_s^{ind})^{-1}C_{s,n}$. Here, $\hat{\Sigma}_s^{ind}$ is an appropriate estimator of Σ_s. In contrast to $\hat{\Sigma}_s^{cc}$, within $\hat{\Sigma}_s^{ind}$, also the p_i would have to be estimated. Interestingly, the asymptotic behavior of $T_{s,n}^{m-cc}$ and $T_{s,n}^{m-ind}$ are the same, as the following theorem shows. Its proof is deferred to Appendix A.1.

Theorem 6. *Let* H_0^{m-ind} *and (Assumption 4) be true. Then, as* $n \to \infty$, $T_{s,n}^{m-ind} \to_d \chi^2_{|A_s|}$.

We obtain the

Ind-m-Test. Reject H_0^{m-ind} if $T_{s,n}^{m-ind} > q_{1-\alpha,\chi^2}$, where $q_{1-\alpha,\chi^2}$ is the $1-\alpha$-quantile of the χ^2-distribution with $|A_s|$ degrees of freedom.

Both the Ind-m-cc-Test and the Ind-m-Test are consistent if, e.g., under the alternative $\mathbb{E}(f(i,j,l,t)) = c \neq 0$ for $(i,j,l) \in A_s$ and if (for the Ind-m-Test) $\tilde{p}_i := \mathbb{E}(I_{t,i}(p_i))$, $\tilde{p}_j := \mathbb{E}(I_{t,j}(p_j))$ as well as Assumption (4) are fulfilled.

3. Simulation study

To examine the performance of our newly proposed backtests in finite samples, we perform a simulation study. Within the study, we distinguish between different kinds of controllable violations concerning Assumptions (7), (6), (11), (9), (8), and (12). We compute all rejection rates for a significance level of 5%.

3.1. Non-constant expectations

As a first step in our simulation study, we want to test if the new tests are able to detect non-constant expectations. We use the Stat-m-test as well as the Ind-m-test. With the latter, the subset A_s consists of the vectors $(i,i,1), i = 1,\dots,m$, which corresponds to a time lag of 1. We expect the CUSUM-test to clearly outperform the χ^2-test in this setting. Basically, the data generating process used throughout the whole simulation study is given by:

$$I_{t,i} = \mathbb{I}(X_{t,i} \leq q_p), \quad \forall i,t. \tag{14}$$

In the first case we consider, q_α is the α-quantile of a standard normal distribution. Moreover, $(X_{t,1},\dots,X_{t,m})$ follows a multivariate normal distribution with mean zero, marginal variances one and cross-correlation $\rho = 0$. Next, we set $m = 10$, $p = 0.01, 0.05$, $n = 250, 500, 1000, 2000$ and use 5000 Monte Carlo repetitions. Finally, we modify Eq. (14) such that

$$I_{t,i} = \begin{cases} \mathbb{I}(X_{t,i} \leq q_{p-2\delta}), & 1 \leq t \leq \frac{n}{4}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p+\delta}), & \frac{n}{4} < t \leq \frac{n}{2}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p-\delta}), & \frac{n}{2} < t \leq \frac{3n}{4}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p+2\delta}), & \frac{3n}{4} < t \leq n, \quad \forall i. \end{cases}$$

In this setting, VaR-violations are independent over time. Hence, clustering is solely based on changes of the probability of obtaining a VaR-violation. We choose $\delta = 0p$ to analyze the size of a test and $\delta = 0.1p,\ 0.2p,\ 0.3p,\ 0.4p$ and $0.5p$ for the power study.

This setting leads to variations in the probability of obtaining a VaR-violation between the four equal-sized subsamples. Consequently, the violations will occur unequally distributed. Note that the probability variations are determined in a way which ensures $\mathbb{E}(\sum_{t=1}^{n}\sum_{i=1}^{m} I_{t,i}) = n \cdot m \cdot p$. The setup of this part of the simulation study covers a realistic scenario in which VaR-models do not, or not fully, incorporate changes from calm market phases to highly volatile bear markets or financial crises, and vice versa. This in turn leads to clustered VaR-violations regardless of the question whether the data might show signs of dependence or autocorrelation. The results of the simulations are given in Table 2.

The results show that the Stat-m-test clearly outperforms the Ind-m-test with rejection probabilities being regularly higher for the Stat-m-test than for the Ind-m-test. In fact, for the larger sample sizes of $n = 1000$ and $n = 2000$ and higher values of δ, the probability of the Stat-m-test to reject the matrix of VaR-violations with non-constant expectations is close to one. In contrast, the Ind-m-test rejects H_0 only in 32% of the simulations at most for $p = 0.01$.

Table 3
Simulated rejection probabilities for cross-sectional correlation.

Panel A: $p = 0.01$											
ρ		Stat-m-test					Ind-m-test				
		0.0	0.2	0.4	0.6	0.8	0.0	0.2	0.4	0.6	0.8
	250	0.04	0.04	0.03	0.02	0.02	0.28	0.81	0.99	1.00	1.00
n	500	0.04	0.04	0.04	0.03	0.02	0.26	0.89	1.00	1.00	1.00
	1000	0.04	0.04	0.04	0.03	0.03	0.21	0.98	1.00	1.00	1.00
	2000	0.04	0.04	0.04	0.04	0.04	0.15	1.00	1.00	1.00	1.00
Panel B: $p = 0.05$											
	250	0.04	0.04	0.03	0.03	0.04	0.09	0.97	1.00	1.00	1.00
n	500	0.04	0.04	0.04	0.04	0.04	0.08	1.00	1.00	1.00	1.00
	1000	0.04	0.04	0.04	0.04	0.04	0.06	1.00	1.00	1.00	1.00
	2000	0.05	0.05	0.04	0.05	0.04	0.05	1.00	1.00	1.00	1.00

The table presents the rejection probabilities of the Stat-m-test and the Ind-m-test based on simulated data with cross-sectional correlation ρ, $p = 0.01$ (Panel A) and $p = 0.05$ (Panel B).

Table 4
Simulated rejection probabilities for serial dependence.

Panel A: $p = 0.01$											
ϕ		Stat-m-test					Ind-m-test				
		0.0	0.25	0.5	0.75	1.0	0.0	0.25	0.5	0.75	1.0
	250	0.03	0.05	0.07	0.09	0.10	0.09	0.41	0.74	0.88	0.90
n	500	0.04	0.06	0.08	0.10	0.10	0.16	0.59	0.91	0.98	0.99
	1000	0.04	0.06	0.09	0.10	0.11	0.14	0.77	0.99	1.00	1.00
	2000	0.05	0.07	0.10	0.11	0.11	0.11	0.91	1.00	1.00	1.00
Panel B: $p = 0.05$											
	250	0.04	0.08	0.12	0.15	0.15	0.07	0.73	0.99	1.00	1.00
n	500	0.05	0.08	0.12	0.16	0.17	0.06	0.92	1.00	1.00	1.00
	1000	0.05	0.09	0.14	0.17	0.17	0.05	1.00	1.00	1.00	1.00
	2000	0.04	0.10	0.15	0.17	0.18	0.05	1.00	1.00	1.00	1.00

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data with autocorrelation ϕ, cross-correlation $\rho = 0.3$, $p = 0.01$ (Panel A) and $p = 0.05$ (Panel B).

3.2. Cross-sectional dependence

In the second part of our simulation study, we want to investigate if the new tests are able to detect cross-sectional dependence within VaR-violations. Here, we expect the χ^2-test to clearly outperform the CUSUM-test. Again, we simulate random variables by

$$I_{t,i} = \mathbb{I}(X_{t,i} \leq q_p), \forall i, t,$$

where, q_α is the α-quantile of a standard normal distribution. Moreover, $(X_{t,1}, \ldots, X_{t,m})$ follows a multivariate normal distribution with mean zero, marginal variances one and cross-correlation ρ. In addition, we choose $m = 10$, $p = 0.01, 0.05$, $n = 250, 500, 1000, 2000$, and again use 5000 Monte Carlo repetitions. The cross-correlation ρ of the normally distributed random variables is set to be in {0, 0.2, 0.4, 0.6, 0.8}. Based on this setting, we analyze the Stat-m-test and the Ind-m-test with time lag 0, that means that we test for cross-sectional dependence. The subset A_s consists of the vectors $(i, j, 0), i < j, i, j = 1, \ldots, m$.

The results are given in Table 3.

As can be seen from the simulation results given in Table 3, the probability to detect a matrix of VaR-violations suffering from cross-correlation is almost always lower for the Stat-m-test than for the Ind-m-test. While the Ind-m-test has appealing power properties in almost all settings, the Stat-m-test is not able to detect cross-sectional dependence. However, the Ind-m-test suffers from size distortions for $p = 0.01$ and small sample sizes.

Next, we also consider the situation with time lag 1 which implies that we test for serial dependence in the VaR-violations. Here, $(X_{t,1}, \ldots, X_{t,m})$ follows a MA(1)-process with autocorrelation parameter $\phi \in \{0, 0.25, 0.5, 0.75, 1\}$, i.e.,

$$(X_{t,1}, \ldots, X_{t,m}) = \epsilon_t + \phi\epsilon_{t-1}$$

for a sequence of i.i.d. bivariate normally distributed vectors $\epsilon_t, t = 1, \ldots, n$, with cross-correlation set to $\rho_t = 0.3$. The subset A_s consists of the vectors $(i, i, 1), i = 1, \ldots, m$. The indicator variables are defined as $I_{t,i} = \mathbb{I}(X_{t,i} \leq q_p\sqrt{1+\phi^2})$. The results of the simulations in which both multivariate backtests are used on data with serially correlated VaR-violations are given in Table 4.

The results given in Table 4 show that the Ind-m-test again performs significantly better than the Stat-m-test.

3.3. Non-constant expectations and serial dependence

In the third part of our simulation study, we investigate the performance of our new multivariate backtests in a setting in which the data exhibit a combination of non-constant expectations and serial dependence. For this purpose, we define

$$I_{t,i} = \begin{cases} \mathbb{I}(X_{t,i} \leq q_{p-2\delta}\sqrt{1+\phi^2}), & 1 \leq t \leq \frac{n}{4}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p+\delta}\sqrt{1+\phi^2}), & \frac{n}{4} < t \leq \frac{n}{2}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p-\delta}\sqrt{1+\phi^2}), & \frac{n}{2} < t \leq \frac{3n}{4}, \quad \forall i; \\ \mathbb{I}(X_{t,i} \leq q_{p+2\delta}\sqrt{1+\phi^2}), & \frac{3n}{4} < t \leq n, \quad \forall i. \end{cases}$$

Here, $(X_{t,1}, \ldots, X_{t,m})$ follows the same MA(1) process as previously described above. Consequently, we use the same parameterization as before and investigate all parameter combinations of δ and ϕ. This setting ensures that we can draw correct conclusions concerning the characteristics of both tests in various situations. We consider the Stat-m-test and the Ind-m-test with $A_s = \{(i, i, 1)\}, i = 1, \ldots, m$.

The results are given in Tables 5 and 6.

Again, the results from the simulations show a clear picture. Except for $\phi = 0$, the Ind-m-test always performs significantly better

Table 5
Simulated rejection probabilities for non-constant expectations and serial dependence with $p = 0.01$.

Panel A: $\phi = 0$													
δ/p		Stat-m-test						Ind-m-test					
		0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
	250	0.04	0.05	0.10	0.17	0.30	0.45	0.10	0.09	0.10	0.12	0.13	0.15
n	500	0.04	0.07	0.18	0.37	0.64	0.95	0.15	0.15	0.17	0.19	0.21	0.26
	1000	0.04	0.10	0.35	0.75	0.98	1.00	0.13	0.15	0.17	0.20	0.22	0.29
	2000	0.05	0.19	0.69	0.98	1.00	1.00	0.12	0.12	0.14	0.18	0.24	0.31
Panel B: $\phi = 0.25$													
	250	0.05	0.06	0.13	0.22	0.33	0.49	0.40	0.39	0.43	0.45	0.48	0.53
n	500	0.06	0.10	0.20	0.40	0.67	0.95	0.60	0.60	0.63	0.65	0.70	0.74
	1000	0.06	0.13	0.39	0.76	0.98	1.00	0.77	0.77	0.79	0.83	0.86	0.89
	2000	0.06	0.23	0.71	0.99	1.00	1.00	0.91	0.91	0.93	0.94	0.96	0.98
Panel C: $\phi = 0.5$													
	250	0.07	0.10	0.16	0.25	0.38	0.53	0.75	0.76	0.76	0.78	0.81	0.84
n	500	0.08	0.11	0.23	0.44	0.69	0.94	0.92	0.92	0.92	0.94	0.95	0.96
	1000	0.08	0.17	0.42	0.76	0.98	1.00	0.99	0.99	0.99	0.99	1.00	1.00
	2000	0.10	0.26	0.72	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel D: $\phi = 0.75$													
	250	0.09	0.11	0.17	0.26	0.40	0.56	0.88	0.89	0.88	0.90	0.92	0.92
n	500	0.10	0.13	0.25	0.45	0.70	0.95	0.98	0.98	0.98	0.98	0.99	0.99
	1000	0.11	0.19	0.44	0.78	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.11	0.28	0.73	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel E: $\phi = 1$													
	250	0.09	0.11	0.17	0.27	0.40	0.57	0.90	0.91	0.92	0.92	0.93	0.95
n	500	0.09	0.15	0.27	0.45	0.70	0.95	0.98	0.99	0.99	0.99	0.99	0.99
	1000	0.11	0.20	0.46	0.78	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.12	0.30	0.73	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data that exhibit a combination of non-constant expectations and serial dependence with autocorrelation ϕ and cross-correlation $\rho = 0.3$.

Table 6
Simulated rejection probabilities for non-constant expectations and serial dependence with $p = 0.05$.

Panel A: $\phi = 0$													
δ/p		Stat-m-test						Ind-m-test					
		0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
	250	0.04	0.10	0.32	0.71	0.97	1.00	0.06	0.07	0.09	0.13	0.20	0.31
n	500	0.05	0.18	0.67	0.98	1.00	1.00	0.06	0.06	0.08	0.14	0.26	0.43
	1000	0.04	0.37	0.96	1.00	1.00	1.00	0.06	0.07	0.10	0.19	0.34	0.63
	2000	0.05	0.69	1.00	1.00	1.00	1.00	0.05	0.06	0.10	0.24	0.55	0.87
Panel B: $\phi = 0.25$													
	250	0.08	0.14	0.38	0.72	0.97	1.00	0.72	0.73	0.77	0.84	0.89	0.95
n	500	0.09	0.24	0.70	0.98	1.00	1.00	0.92	0.93	0.95	0.98	0.99	1.00
	1000	0.09	0.43	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.09	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel C: $\phi = 0.5$													
	250	0.12	0.20	0.41	0.74	0.96	1.00	0.99	0.99	1.00	1.00	1.00	1.00
n	500	0.13	0.29	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	0.14	0.47	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.15	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel D: $\phi = 0.75$													
	250	0.15	0.22	0.46	0.76	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n	500	0.15	0.33	0.72	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	0.17	0.50	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.18	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel E: $\phi = 1$													
	250	0.15	0.24	0.46	0.75	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n	500	0.16	0.33	0.72	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	0.18	0.52	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.19	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data that exhibit a combination of non-constant expectations and serial dependence with autocorrelation ϕ and cross-correlation $\rho = 0.3$.

than the Stat-m-test. In contrast, mean rejection probabilities are only higher for the Stat-m-test in the setting in which ϕ is set to zero. For $p = 0.01$, the size of the Ind-m-test is somewhat higher than expected.

3.4. Violations of the cc-property

Within the last setting of our simulation study, we concentrate on violations of the cc-property. To this end, we simulate data that exhibit serial dependence and also violations of the uc-property.

Table 7
Simulated rejection probabilities for violation of the cc-property with $p = 0.01$.

		Stat-m-cc-test						Ind-m-cc-test					
Panel A: $\phi = 0$													
δ/p		0.0	0.2	0.4	0.6	0.8	1.0	0.0	0.2	0.4	0.6	0.8	1.0
	250	0.02	0.08	0.24	0.47	0.69	0.86	0.24	0.30	0.38	0.42	0.48	0.53
n	500	0.05	0.12	0.45	0.81	0.95	0.99	0.27	0.36	0.44	0.59	0.68	0.75
	1000	0.05	0.28	0.79	0.99	1.00	1.00	0.12	0.23	0.36	0.52	0.67	0.80
	2000	0.05	0.55	0.98	1.00	1.00	1.00	0.15	0.29	0.46	0.62	0.76	0.86
Panel B: $\phi = 0.25$													
	250	0.03	0.10	0.27	0.49	0.69	0.85	0.65	0.73	0.78	0.82	0.85	0.85
n	500	0.06	0.15	0.47	0.78	0.94	0.99	0.73	0.84	0.90	0.94	0.96	0.96
	1000	0.07	0.29	0.79	0.98	1.00	1.00	0.78	0.91	0.97	0.99	1.00	1.00
	2000	0.07	0.55	0.98	1.00	1.00	1.00	0.91	0.97	0.99	1.00	1.00	1.00
Panel C: $\phi = 0.5$													
	250	0.05	0.13	0.29	0.52	0.71	0.84	0.89	0.92	0.93	0.94	0.93	0.92
n	500	0.08	0.17	0.47	0.78	0.93	0.98	0.95	0.97	0.99	0.99	0.98	0.97
	1000	0.09	0.31	0.77	0.97	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
	2000	0.09	0.55	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel D: $\phi = 0.75$													
	250	0.05	0.15	0.31	0.52	0.72	0.84	0.95	0.96	0.96	0.96	0.94	0.93
n	500	0.09	0.19	0.48	0.77	0.93	0.98	0.98	0.99	0.99	0.99	0.98	0.97
	1000	0.10	0.33	0.77	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.11	0.56	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel E: $\phi = 1$													
	250	0.06	0.15	0.34	0.53	0.71	0.83	0.96	0.97	0.97	0.96	0.94	0.93
n	500	0.10	0.20	0.48	0.77	0.93	0.98	0.99	0.99	0.99	0.99	0.98	0.97
	1000	0.12	0.31	0.78	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.11	0.56	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data that violate the cc-property with autocorrelation ϕ and cross-correlation $\rho = 0.3$.

We define

$$I_{t,i} = \mathbb{I}(X_{t,i} \leq q_{p+\delta}\sqrt{1+\phi^2}), \quad \forall i, t.$$

To be more precise, $(X_{t,1}, \ldots, X_{t,m})$ follows the same MA(1) process as before and δ is set to $0.2p, 0.4p, 0.6p, 0.8p$, and p, respectively. Apart from that, we use the same parameterization as before and investigate all parameter combinations of δ and ϕ. We consider the Stat-m-cc-test and the Ind-m-cc-test with $A_s = \{(i, i, 1)\}, i = 1, \ldots, m$. The results are given in Tables 7 and 8.

Just like in our simulations with data that exhibit cross-sectional dependence and non-constant expectations together with serial dependence in the matrix of VaR-violations, the Ind-m-test again performs significantly better than the Stat-m-test expect for the setting for $\phi = 0$. The Ind-m-test suffers from size distortions for $p = 0.01$ and small sample sizes.

In general, we observe that both the Ind-m-test and the Stat-m-test have sufficient power in almost all settings of our simulation study even for relatively small sample sizes of $n = 250$. Moreover, both tests also hold their nominal level in almost all simulation settings. The simulations thus underline the suitability of our newly proposed backtests for testing the adequacy of a multivariate VaR model.[8]

4. Empirical study

In this section, we perform an empirical study in which we exemplify the usefulness of our newly proposed multivariate backtests for the purpose of assessing the systemic risk of a banking sector. To be precise, we perform multivariate backtests of risk forecasts for a portfolio consisting of the stocks of the world's 20 largest banks. The motivation of this use of our multivariate backtests is straightforward. A regulator of the set of banks could be interested in testing both a) the overall fragility of the banking sector, and b) the temporal and cross-sectional clusters in severe losses at individual banks. While a bank's risk manager could be inclined to simply aggregate the risk forecasts for his individual business lines (thereby obviating the need for our multivariate backtests), the additional information our backtests provide on the sources of violations (temporal and/or cross-sectional) should be of particular interest to bank regulators.

We start our empirical analysis by retrieving the mid prices on the stocks of the 20 largest global banks available in the *Thomson Reuters Financial Datastream* database according to the banks' market capitalization on the 1st of January, 2002.[9] Our sample period runs from the 1st of February, 2003 until the 31st of December, 2014. In total, our sample consists of 3549 daily observations of the bank stocks' mid prices. All data are retrieved in $ US to mitigate a potential bias stemming from currency risk.

We then proceed by computing daily log returns for each day and each bank in our 20-dimensional sample portfolio. For each bank stock and each trading day t in the period from January 1, 2004 to December 31, 2014, we forecast the bank's Value-at-Risk ($p = 0.05$) on day t by first fitting a univariate GJR-GARCH(1,1) model of Glosten et al. (1993) with skewed t-distributed innovations to (a rolling window of) the 500 trading days preceding day t and then forecasting the estimated model's mean and conditional volatility for day t. The forecasted VaRs are then compared with the actual return on a bank's stock on the respective day to arrive at the hit matrix of VaR violations that is inserted into our multivariate backtests.

Table 9 presents summary statistics on the banks' daily stock returns as well as the yearly p-values for every applied test and the results for the argmax estimators for the CUSUM tests.

[8] All simulations were also performed with $m = 20$. The results are presented in the supplementary material as the general pattern is very similar to the case of $m = 10$.

[9] The banks in our sample are: Citigroup, HSBC Holdings (dual listing), UBS, Barclays, BNP Paribas, Mitsubishi UFJ, RBS, Crédit Agricole, Bank of America, JP Morgan Chase, Deutsche Bank, Mizuho Financial Group, ABN Amro, Société Générale, Morgan Stanley, HBOS, Banco Santander, Unicredit, and Credit Suisse.

Table 8
Simulated rejection probabilities for violation of the cc-property with $p = 0.05$.

		Stat-m-cc-Test						Ind-m-cc-Test					
Panel A: $\phi = 0$													
δ/p		0.0	0.2	0.4	0.6	0.8	1.0	0.0	0.2	0.4	0.6	0.8	1.0
	250	0.05	0.27	0.79	0.98	1.00	1.00	0.10	0.25	0.46	0.68	0.80	0.82
n	500	0.05	0.53	0.98	1.00	1.00	1.00	0.08	0.24	0.48	0.69	0.87	0.94
	1000	0.05	0.86	1.00	1.00	1.00	1.00	0.06	0.23	0.50	0.73	0.89	0.97
	2000	0.05	0.99	1.00	1.00	1.00	1.00	0.06	0.23	0.51	0.79	0.95	0.99
Panel B: $\phi = 0.25$													
	250	0.09	0.29	0.78	0.97	1.00	1.00	0.74	0.90	0.97	0.98	0.96	0.92
n	500	0.09	0.55	0.97	1.00	1.00	1.00	0.92	0.99	1.00	1.00	1.00	0.99
	1000	0.09	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.09	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel C: $\phi = 0.5$													
	250	0.12	0.33	0.75	0.97	1.00	1.00	0.99	1.00	1.00	0.99	0.97	0.95
n	500	0.12	0.55	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	1000	0.12	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.12	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel D: $\phi = 0.75$													
	250	0.14	0.35	0.75	0.95	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.95
n	500	0.14	0.56	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	1000	0.14	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.15	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel E: $\phi = 1$													
	250	0.15	0.34	0.75	0.96	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.95
n	500	0.15	0.56	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	1000	0.15	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	0.16	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The table presents the rejections probabilities of the Stat-m-test and the Ind-m-test based on simulated data that violate the cc-property with autocorrelation ϕ and cross-correlation $\rho = 0.3$.

Table 9
Summary statistics and test results for the empirical study.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
∅ Return per day	0.05%	0.03%	0.08%	−0.07%	−0.40%	0.09%	−0.04%	−0.18%	0.11%	0.09%	−0.04%
∅ Volatility per day	1.37%	1.12%	1.28%	1.72%	5.17%	4.92%	2.37%	3.16%	2.48%	1.68%	1.35%
Number of violations	220	267	295	433	404	279	223	324	221	277	334
Number of 2 subsequent violations	26	14	21	78	64	32	18	45	17	24	44
$T_{s,n}^{m-cc}$ Cross-sectional (p-value)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
$T_{s,n}^{m-ind}$ Cross-sectional (p-value)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
$T_{s,n}^{m-cc}$ Serial (p-value)	0.01%	28.86%	0.28%	0.00%	0.00%	0.00%	12.25%	0.00%	0.10%	0.00%	0.00%
$T_{s,n}^{m-ind}$ Serial (p-value)	0.00%	26.19%	1.91%	0.01%	0.00%	0.25%	3.57%	1.82%	13.57%	14.21%	0.34%
$RC_{cc,\,n}$ (p-value)	31.11%	41.84%	4.21%	0.07%	0.18%	21.89%	55.15%	6.94%	44.31%	52.72%	2.97%
$RC_{cc,\,n}$ (argmax)	31.12.	02.05.	13.06.	17.12.	24.12.	31.03.	31.12.	09.11.	31.12.	24.06.	16.12.
RC_n (p-value)	4.30%	12.95%	0.14%	22.66%	40.28%	2.41%	44.19%	18.95%	10.72%	40.54%	25.16%
RC_n (argmax)	10.05.	02.05.	13.06.	17.07.	22.01.	30.03.	07.05.	14.06.	02.08.	24.06.	24.09.

The table presents some yearly summary statistics for the empirical study, including the average daily return and volatility, the number of (subsequent) violations, the p-value for every applied test and the results for the argmax estimators for the CUSUM tests. The sample consists of log returns on the mid prices of the stocks of the 20 largest global banks available in the *Thomson Reuters Financial Datastream* database according to the banks' market capitalization on the 1st of January, 2002. The banks in our sample are: Citigroup, HSBC Holdings (dual listing), UBS, Barclays, BNP Paribas, Mitsubishi UFJ, RBS, Crédit Agricole, Bank of America, JP Morgan Chase, Deutsche Bank, Mizuho Financial Group, ABN Amro, Société Générale, Morgan Stanley, HBOS, Banco Santander, Unicredit, and Credit Suisse. Our sample period runs from the 1st of February, 2003 until the 31st of December, 2014. All data are retrieved in $ US. VaR forecasts are computed using GJR-GARCH(1,1) models with skewed t-distributed innovations based on rolling windows of 500 observations and $p = 0.05$.

The test results given in Table 9 are in line with our expectation. Both cross-sectional tests (the χ^2-tests applied with $A_s = \{(i, j, 0)\}, i < j, i, j = 1, \ldots, m,$ as before) have a p-value of 0% for every year. This is due to the fact that violations often occur heaped resulting in 10 or more violations within one day. Moreover, these results illustrate and underline the contagion effects within the banking sector which could be observed during the last decade as severe losses occur at the same day and affect several banks simultaneously. Fig. 2 illustrates this phenomenon for the years 2005–2007.

In contrast to this first finding, the results of both serial tests (the χ^2-tests applied with $A_s = \{(i, i, 1)\}, i = 1, \ldots, m,$ as before) vary widely. Here, the ind-test is highly dependent on the number of subsequent violations, while the cc-tests also react to the total number of violations. This effect is highly visible for the years 2005–2007, during which the number of (subsequent) violations increases monotonously while the p-values decrease. This finding is also in line with economic developments. Following a stable and calm market phase, the financial crisis started in 2007 resulting in increased volatility and dependence, especially for banks.

Finally, the p-values of the CUSUM-tests are considerably higher on average than for the remaining tests. Here, the cc-test is much less sensitive than its cross-sectional and serial counterparts and has p-values that are lower than 1% only for the years 2007 and 2008. This finding is in line with our simulation study presented above. The CUSUM-test for non-constant expectations has to be interpreted in a slightly different way as it has higher power for clusters of violations induced by instationarities than those induced by

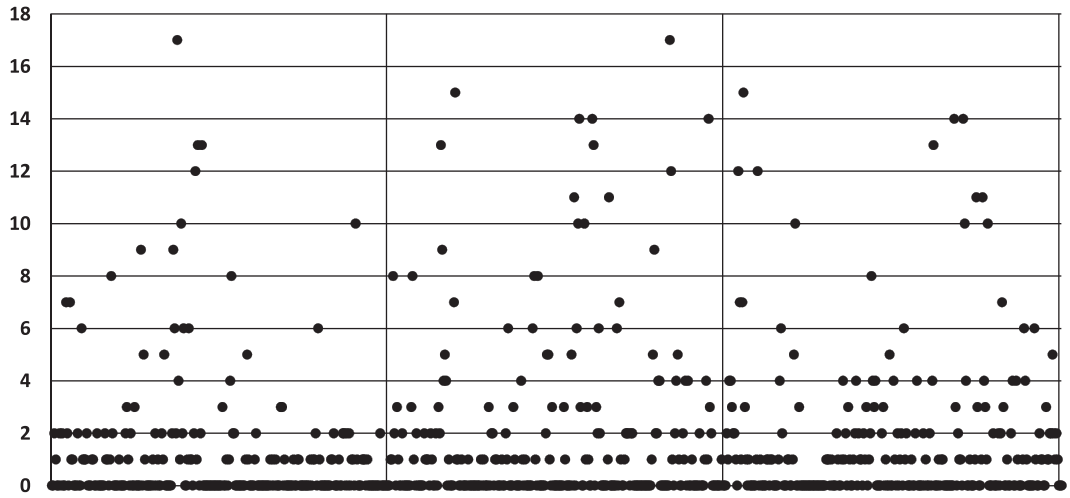

Fig. 2. Daily row sums of the multivariate Value-at-Risk hit matrix. This figure presents the daily row sums of the multivariate Value-at-Risk hit matrix for the years 2005–2007 from our empirical study. The sample consists of log returns on the mid prices of the stocks of the 20 largest global banks available in the *Thomson Reuters Financial Datastream* database according to the banks' market capitalization on the 1st of January, 2002. The banks in our sample are: Citigroup, HSBC Holdings (dual listing), UBS, Barclays, BNP Paribas, Mitsubishi UFJ, RBS, Crédit Agricole, Bank of America, JP Morgan Chase, Deutsche Bank, Mizuho Financial Group, ABN Amro, Société Générale, Morgan Stanley, HBOS, Banco Santander, Unicredit, and Credit Suisse. Our sample period runs from the 1st of February, 2003 until the 31st of December, 2014. All data are retrieved in $ US. VaR forecasts are computed using GJR-GARCH(1,1) models with skewed t-distributed innovations based on rolling windows of 500 observations and $p = 0.05$. Consequently, with $p = 0.05$ and assuming that the VaR-model is correct, the expected value is 1 for each row.

serial dependencies (for lag one). In particular, considerable instationarities lead to small p-values. So, a possible interpretation is that, in general, the cluster effects in our data set are caused by serial dependencies rather than instationarities. Nonetheless, the CUSUM-test for non-constant expectations has a p-value below 5% for 3 of the 11 investigated years. This is not negligible, particularly as the sample size is small. A special example is the year 2006 which is illustrated in Fig. 2. Here, high row sums are highly clustered, resulting in a p-value of 0.14%. Concerning the argmax estimators it is interesting to see how the clustering of violations corresponds to a change in the market condition. In some cases, there is a very strong relation. A striking example is again the year 2006. The 13th of June represents exactly the date where several of the considered stocks changed from a bearish to a bullish market state.

5. Conclusion

In this paper, we have proposed two new multivariate backtests for clusters in VaR-violations. The first test is a CUSUM-test which is based on the sums of the violations for different business lines and sub-portfolios for a single day and which attempts to detect clusters in the matrix of VaR-violations that are caused by instationarities in the mean of the violations. Second, we consider a χ^2-test for detecting clusters that are caused by cross-sectional and/or serial dependencies within the VaR-violations. Both tests are easy to implement and work without Monte Carlo simulations or bootstrap approximations, although bootstrap approximations are readily available.

In simulations, we assess the performance of our new multivariate backtests in several distinct settings in which we consider simulated data that exhibit non-constant expectations, cross-sectional dependence, and serial dependence in the VaR-violations. Moreover, we also perform simulations in which the new backtests are used to test the simulated VaR-violations for the property of conditional coverage. With the exception of the setting in which the data only exhibit non-constant expectations, the χ^2-test performs better in our simulations than the CUSUM-test. Both tests hold their nominal level and, more importantly, have considerable power for testing the conditional coverage of the matrix of VaR-violations even for relatively small sample sizes.

The multivariate backtests that we propose are intended for the use by risk managers, portfolio managers, and regulators. Especially in the last case, our backtests cannot only be used in the conventional way within individual banks, but also to backtest a whole banking sector. To this end, VaRs are estimated across time and individual banks (instead of business lines) with clusters in VaR-violations across banks indicating systemic risk in the sector. In this way, our backtests could be of significant help to regulators to forecast times of contagion in the financial system and thereby complement current endeavors to stress-test banking sectors (see, e.g., Acharya and Steffen, 2013).

Appendix

A.1. Proof of Theorem 6

First, we consider the process

$$\tilde{C}_{s,n} := \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n-l} (I_{t,i}(p_i) - \tilde{p}_i)(I_{t+l,j}(p_j) - \tilde{p}_j) \right)_{(i,j,l)\in A_s},$$

and show that $\tilde{C}_{s,n} = C_{s,n} + o_p(1)$.

We define $\hat{p}_k := \frac{1}{n}\sum_{t=1}^{n} I_{t,k}(p_k)$. Then, it holds

$$\tilde{C}_{s,n} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n-l} I_{t,i}(p_i) I_{t+l,j}(p_j) - \frac{n-l}{\sqrt{n}} \hat{p}_i \tilde{p}_j - \frac{n-l}{\sqrt{n}} \hat{p}_j \tilde{p}_i + \frac{n-l}{\sqrt{n}} \tilde{p}_i \tilde{p}_j + o_p(1)$$

and

$$C_{s,n} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n-l} I_{t,i}(p_i) I_{t+l,j}(p_j) - \frac{n-l}{\sqrt{n}} \hat{p}_i \hat{p}_j - \frac{n-l}{\sqrt{n}} \hat{p}_j \hat{p}_i + \frac{n-l}{\sqrt{n}} \hat{p}_i \hat{p}_j + o_p(1)$$

such that

$$\tilde{C}_{s,n} - C_{s,n} = \frac{n-l}{\sqrt{n}}(-\hat{p}_i \tilde{p}_j + \hat{p}_i \hat{p}_j) + \frac{n-l}{\sqrt{n}}(-\hat{p}_j \tilde{p}_i + \hat{p}_j \hat{p}_i) + \frac{n-l}{\sqrt{n}}(\tilde{p}_i \tilde{p}_j - \hat{p}_i \hat{p}_j)$$

$$= \frac{n-l}{\sqrt{n}}(\tilde{p}_i\tilde{p}_j - \hat{p}_i\tilde{p}_j + \hat{p}_j\hat{p}_i - \hat{p}_j\tilde{p}_i)$$
$$= \frac{n-l}{\sqrt{n}}(\tilde{p}_j(\tilde{p}_i - \hat{p}_i) + \hat{p}_j(\hat{p}_i - \tilde{p}_i))$$
$$= \frac{n-l}{\sqrt{n}}(\hat{p}_i - \tilde{p}_i)(\hat{p}_j - \tilde{p}_j)$$
$$= O_p(1)o_p(1) = o_p(1).$$

Then, the result from the theorem follows from the fact that, by uniform integrability, one directly obtains $\Sigma_s = \lim_{n\to\infty} Cov(C_{s,n})$ and $C_{s,n} \to_d N(0, \Sigma_s)$.

A.2. Bootstrap

A.2.1. Bootstrap approximations for the CUSUM-test in Section 2.2

In order to make the tests more robust against changes in $Var(r_t)$, one can use a recently proposed approach by Zhou (2013). Here, we consider the quantity C, i.e., the test statistic without the variance estimator D^{-1}. Critical values are obtained by using a bootstrap approximation. This bootstrap is an extension of the wild bootstrap and relies on directly mimicking the behavior of the partial sum process V_n instead of mimicking the behavior of C. However, despite the theoretical relevance, some robustness checks show that the bootstrap does not seem to be necessary in the situation of moderate changes in $Var(r_t)$. Moreover, there is no power gain from the robust CUSUM-test.

A.2.2. Bootstrap approximations for the χ^2 tests in Section 2.3

To facilitate the tests' implementation in software, one can estimate the matrix Σ_s with a bootstrap approximation based on the seminal paper by Efron (1979). The bootstrap is essentially the same for testing H_0^{m-ind} and H_0^{m-cc}, respectively. We distinguish two cases, i.e., Assumptions (7)/(9) and (6)/(8). In the first one, cross-sectional dependence is allowed for, which is not true in the second one. Let B be a sufficiently large number.

Then, under Assumption (7) and given the observed matrix of VaR-violations, we generate, for $b = 1, \ldots, B$, a bootstrap sample $I_{t,i}^b, t = 1, \ldots, n, i = 1, \ldots, m$, by drawing n rows with replacement from the observed matrix. Thus, the generated bootstrap samples always fulfill Assumption (7). When testing for cross-sectional dependence (that means, if Assumption (6) holds true under the null hypothesis), the bootstrap procedure from the previous paragraph has the drawback that there is no variation within each row in the bootstrap samples. Thus, in this case a bootstrap sample $I_{t,i}^b, t = 1, \ldots, n, i = 1, \ldots, m$, is obtained in a different way. In order to keep the information concerning $p_i, i = 1, \ldots, m$, for fixed i, $I_{t,i}^b, t = 1, \ldots, n$, is obtained by drawing n values with replacement of the respective business line from the observed matrix, whereas the draws are also independent with respect to i. Then, the generated bootstrap samples always fulfill Assumption (6).

Having obtained a bootstrap sample, we calculate the vector $C_{s,n}^b$ and consider the estimator

$$\Sigma_s^B := \frac{1}{B}\sum_{b=1}^{B}(C_{s,n}^b - \bar{C}_{s,n}^B)(C_{s,n}^b - \bar{C}_{s,n}^B)'$$

with $\bar{C}_{s,n}^B := \frac{1}{B}\sum_{b=1}^{B} C_{s,n}^b$. The test statistic for H_0^{m-cc} is then given by $T_{b,n}^{m-cc} := B_{s,n}'(\Sigma_s^B)^{-1}B_{s,n}$, the one for H_0^{m-iid} is given by $T_{b,n}^{m-iid} := C_{s,n}'(\Sigma_s^B)^{-1}C_{s,n}$. Both need to be compared with the $1-\alpha$-quantile of the $\chi^2_{|A_s|}$-distribution. The validity of this approach under the null hypothesis follows from standard bootstrap theory (bootstrap central limit theorem, see Gonç alves and White, 2002, uniform integrability, see Kato, 2011, Lemma 1), the validity under the alternative follows from the fact that the generated vectors $C_{s,n}^b$ remain stochastically bounded due to the arguments given in the previous paragraph.

Simulations show that the bootstrap tests for (7) and (6) have virtually the same size and power properties as the tests based on an explicit derivation of the matrix Σ_s. While in case of (9) and (8), the bootstrap does work in the sense of accuracy under the null hypothesis and consistency under the alternative, there is some power loss compared to the case in which the matrix Σ_s is calculated directly. Under Assumption (8), a better alternative is given by drawing the $I_{t,i}^b, t = 1, \ldots, n, i = 1, \ldots, m$, independently from Bernoulli distributions with the respective $p_i, i = 1, \ldots, m$.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jbankfin.2016.07.014

References

Acharya, V.V., Steffen, S., 2013. Falling Short of Expectations? Stress-Testing the European Banking System. Working Paper

Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Aue, A., Horváth, L., 2013. Structual breaks in time series. Journal of Time Series Analysis 34, 1–16.

Basel Committee on Banking Supervision (BCBS), 2009. Revisions to the basel II market risk framework, http://www.bis.org/publ/bcbs158.pdf.

Bender, R., Grouven, U., 1993. On the choice of the number of residual autocovariances for the portmanteau test of multivariate autoregressive models. Communications in Statistics-Simulation and Computation 22, 19–32.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business and Economic Statistics 19, 465–474.

Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating value-at-risk models with desk-level data. Management Science 57, 2213–2227.

Billingsley, P., 1968. Convergence of Probability Measures. Wiley, New York.

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.

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

Christoffersen, P., Pelletier, D., 2004. Backtesting value-at-risk: a duration-based approach. Journal of Financial Econometrics 2 (1), 84–108.

Danciulescu, C., 2010. Backtesting Value-at-Risk Models: A Multivariate Approach. SSRN working paper (1591049)

Dumitrescu, E.-I., Hurlin, C., Pham, V., 2012. Backtesting value-at-risk: from dynamic quantile to dynamic binary tests. Finance 33, 79–112.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–25.

Engle, R., Manganelli, S., 2004. CAViar: conditional autoregressive value-at-risk by regression quantiles. Journal of Business and Economic Statistics 22, 367–381.

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

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of nominal excess return on stocks. Journal of Finance 48, 1779–1801.

Gonçalves, S., White, H., 2002. The bootstrap of the mean for dependent heterogeneous arrays. Econometric Theory 18, 1367–1384.

Haas, M., 2005. Improved duration-based backtesting of value-at-risk. Journal of Risk 8 (2), 17–36.

Hurlin, C., Tokpavi, S., 2007. Backtesting value-at-risk accuracy: a simple new test. Journal of Risk 9 (2), 19–37.

Kato, K., 2011. A note on moment convergence of bootstrap m-estimators. Statistics and Decisions 28, 58–71.

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

Pelletier, D., Wei, W., 2015. The geometric-var backtesting method. Journal of Financial Econometrics (forthcoming).

Zhou, Z., 2013. Heteroscedasticity and autocorrelation robust structural change detection. Journal of the American Statistical Association 108, 726–740.

Ziggel, D., Berens, T., Weiß, G.N., Wied, D., 2014. A new set of improved value-at-risk backtests. Journal of Banking and Finance 48, 29–41.