A review of backtesting for value at risk

DOI:
10.1080/03610926.2017.1361984

Document Version
Accepted author manuscript

A review of backtesting for value at risk

by

Y. Zhang and S. Nadarajah
School of Mathematics, University of Manchester, Manchester M13 9PL, UK

Abstract: There have been many backtesting methods proposed for Value at Risk. Yet they have rarely been applied in practice. Here, we provide a comprehensive review of all of the recent backtesting methods for VaR. This review could encourage applications and also the development of further backtesting methods.

Keywords: Coverage; Likelihood ratio; Violation

1 Introduction

Value at Risk (VaR) is the most popular measure for financial risk. VaR was proposed by Till Guldimann in the late 1980s, and at the time he was the head of global research at J.P. Morgan.

VaR has been applied extensively in the financial sector and other areas. A comprehensive review of the mathematical properties, estimation methods and applications for VaR is given in Nadarajah and Chan (2016).

Backtesting is a method that uses historical data to gauge accuracy and effectiveness. Backtesting VaR is used to compare the forecast/predicted losses from the actual calculated losses realised at the end of a fixed time horizon. The results from backtesting provide us with information on specific periods where VaR is underestimated or where the losses are greater than the original expected VaR value. These VaR values can then be recalculated if the backtesting values are not accurate, thereby helping researchers and institutions to reduce their exposure to unexpected losses.

This paper provides a comprehensive review of backtesting methods for VaR. The methods reviewed include: the simplest backtesting method, binomial distribution test, Kupiec's POF (1995) test, Kupiec's TUFF test, proportion of failures test (Haas, 2001), time between failures likelihood ratio test (Haas, 2001), scaled Crnkovic and Drachman (1996)'s method (Haas, 2001), density forecast tests (Berkowitz, 2000, 2001), Lopez's magnitude loss function (Lopez, 1998, 1999), risk map (Colletaz et al., 2013), independence test (Christoffersen, 1998), joint test, generalized Markov tests (Pajhede, 2015), the Basel Committee's traffic light test, quality control of risk measure (de la Pena et al., 2006), duration - based tests (Christoffersen and Pelletier, 2004, Berkowitz et al., 2011), Haas (2006)'s test, generalized method of moments duration - based test (Hurlin et al., 2010), conditional duration test (Christoffersen and Pelletier, 2002), Krämer and Wied (2015)'s test, Markov duration tests (Pajhede, 2015), Escanciano and Olmo (2010, 2011)'s tests, Wald statistic test (Engle and Managanelli, 2004), dynamic binary tests (Dumitrescu et al., 2012), multivariate test (Perignon and Smith, 2008), martingale difference test (Berkovitz et al., 2005), multivariate autocorrelations test (Hurlin and Tokpavi, 2006), geometric-VaR backtesting method

(Pelletier and Wei, 2015).

There have been other reviews of backtesting methods, see Campbell (2005) and Virdi (2011). But the review provided here is by far the most comprehensive, reviewing nearly thirty methods. For general references on backtesting, we refer the readers to the excellent books Schulmerich (2010), Matras (2011), Franke et al. (2012), Fitschen (2013), Gorgulho et al. (2013), Bera et al. (2015), Lichters et al. (2015) and McNeil et al. (2015).

Some recent applications of backtesting methods for VaR have been to: internal model validation in Brazil (da Silva et al., 2006); risk models in the presence of structural breaks on the Romanian and Hungarian stock markets (Daniela et al., 2014); the Romanian capital market (Iorgulescu, 2012); VaR estimation of the Prague stock market index (Kresta, 2013); VaR models in crude oil markets (Li et al., 2016); analysis of banking shares in India (Patra and Padhi, 2015); the world FX rate market (Tichy, 2012).

Section 2 reviews twenty eight backtesting methods for VaR. For each method, we give the hypotheses, test statistic, asymptotic null distribution and advantages/disadvantages. Other details like small sample properties, power, type 1 error and type 2 error can be found from the cited references. Some known software for backtesting are summarized in Section 3. A data illustration of four of the backtesting methods is given in Section 4.

We have reviewed backtesting methods for only VaR because it is the most popular and most widely used risk measure. A future work is to review backtesting methods for lesser known risk measures like expected shortfall, weighted expected shortfall, tail conditional median, expectiles, generalized quantiles, weighted expected shortfall, beyond value at risk, limited value at risk, etc.

The backtesting methods reviewed can be grouped into different categories:

- **Unconditional test methods**: Kupiec's POF (1995) test, Kupiec's TUFF test, binomial distribution test, the Basel Committee's traffic light test, the simplest backtesting method, the proportion of failures test (Haas, 2001), the scaled Crnkovic and Drachman (1996)'s method (Haas, 2001), the risk map (Colletaz et al., 2013), the quality control of risk measure (de la Pena et al., 2006) and the multivariate test (Perignon and Smith, 2008).

- **Conditional test methods**: Joint test, the time between failures likelihood ratio test (Haas, 2001), the generalized Markov tests (Pajhede, 2015), the multivariate autocorrelations test (Hurlin and Tokpavi, 2006), the dynamic binary tests (Dumitrescu et al., 2012).

- **Independence property test methods**: the independence test (Christoffersen, 1998) and Wald statistic test (Engle and Managanelli, 2004).

- **Other test approaches**: the density forecast tests (Berkowitz, 2000, 2001), Lopez's magnitude loss function (Lopez, 1998, 1999), the duration - based tests (Christoffersen and Pelletier, 2004, Berkowitz et al., 2011), Haas (2006)'s test, the generalized method of moments duration - based test (Hurlin et al., 2010), the conditional duration test (Christoffersen and Pelletier, 2002), Krämer and Wied (2015)'s test, the Markov duration tests (Pajhede, 2015), the martingale difference test (Berkovitz et al., 2005), the geometric-VaR backtesting method (Pelletier and Wei, 2015) and Escanciano and Olmo (2010, 2011)'s tests.

Unconditional methods work through counting the number of violations and comparing them with confidence levels. If the violations are within the statistical limits, the model is accepted, otherwise it is rejected. The unconditional test methods provide a useful benchmark for assessing the accuracy of a given VaR model. However, these methods show difficulties in detecting VaR measures that systematically under report risk. The tests focus only on the unconditional coverage property of an adequate VaR measure and do not examine the extent to which the independence property is satisfied.

Independence tests assess some form of independence in a VaR measure's performance from one period to the next. One of the main drawbacks of independence tests is in detecting inadequate VaR measures. However, independence tests provide an important source of discriminatory power.

2 Backtesting methods

The following notation and terminology will be used throughout: let $\{x_t, t = 1, 2, \ldots, n\}$ denote the time series of losses or profits; let Ω_t denote the information up to time t; define VaR at time t by $\text{VaR}_t(\alpha) = F^{-1}(\alpha \mid \Omega_t)$, where $F^{-1}(\cdot \mid \Omega_t)$ denotes the quantile function of the time series at time t; let $F_t(\cdot)$ denote the cumulative distribution function of the time series at time t; define $I_{t+1}(\alpha) = 1$ if $x_{t+1} \leq -\text{VaR}_t(\alpha)$ and $I_{t+1}(\alpha) = 0$ if $x_{t+1} > -\text{VaR}_t(\alpha)$; a 'violation' is defined as an event where an observation exceeds VaR; p usually denotes the probability of violations; $I\{A\}$ denotes the indicator function; 'unconditional coverage' refers to $\Pr(I_{t+1}(\alpha) = 1) = \alpha$; 'conditional coverage' refers to the probability of $I_{t+1}(\alpha) = 1$ being equal to α conditional on the information at time t.

2.1 Simplest backtesting method

One of the most basic backtesting methods consists of counting the number of losses larger than the estimated VaR for a given period and comparing it to the expected number within a given confidence interval.

2.2 Binomial distribution test

This method is an extension of Christoffersen (1998)'s backtesting discussed later on. It states that if $\{I_t(\alpha)\}$ are independently and identically distributed and $\Pr[I_{t+1}(\alpha) = 1] = \alpha$, then the total number of violations H has a binomial distribution $B(n, \alpha)$ with mean $E(H) = n\alpha$ and $Var(H) = n\alpha(1 - \alpha)$.

If the number of observations is large enough, the central limit theorem can approximate the binomial distribution by the normal distribution. As outlined by Jorion (2001), an immediate test statistic is

$$T = \frac{H - n\alpha}{\sqrt{n\alpha(1 - \alpha)}}. \tag{1}$$

Its asymptotic null distribution is the standard normal distribution. An alternative is to

use the likelihood ratio statistic:

$$LR = -2\ln\left[(1-\alpha)^{n-H}\alpha^{H}\right] + 2\ln\left[\left(1-\frac{H}{n}\right)^{n-H}\left(\frac{H}{n}\right)^{H}\right]. \tag{2}$$

Its asymptotic null distribution is the chi-squared distribution with one degree of freedom.

2.3 Kupiec's POF (1995) test (proportion of failures)

Kupiec's POF test was the earliest proposed VaR backtest. The test is concerned with whether or not the reported VaR is violated more or less than 100α percent of the time. The Kupiec test statistic is

$$POF = 2\ln\left[\left(\frac{1-\widehat{\alpha}}{1-\alpha}\right)^{n-I(\alpha)}\left(\frac{\widehat{\alpha}}{\alpha}\right)^{I(\alpha)}\right], \tag{3}$$

where

$$\widehat{\alpha} = \frac{1}{n}I(\alpha)$$

and

$$I(\alpha) = \sum_{t=1}^{n} I_t(\alpha),$$

where n denotes the number of observations. The test statistic reveals that if the proportion of VaR violations, $100\widehat{\alpha}$ percent, is exactly equal to 100α percent then the POF test takes the value zero, indicating no evidence of any inadequacy in the underlying VaR model. As the proportion of VaR violations differs from 100α percent, the POF test statistic grows indicating mounting evidence that the proposed VaR model either systematically understates or overstates the portfolio's underlying level of risk.

Dowd (2002) provided a bootstrap version (Efron and Tibshirani, 1994) of Kupiec's test. One advantage of Kupiec's POF (1995) test is that it is simple to implement and use. It is statistically weak if a sample size consistent with a framework of one year is used. Also, the test only considers the frequency of losses and not the time when they occur. Therefore, it may fail to reject a model that produces clustered violations.

2.4 Kupiec's TUFF test (Time until first failure)

Based on the same assumptions as the POF test, Kupiec's TUFF test (LR test) measures the time until the first violation. The null hypothesis is

$$H_0 : p = \frac{1}{\nu},$$

where ν denotes the time until the first violation. The test statistic can be defined as

$$TUFF = -2\ln\left[\frac{p(1-p)^{\nu-1}}{\widehat{p}\,(1-\widehat{p})^{\nu-1}}\right],$$

where $\widehat{p}$ is an estimator of p. Both POF and TUFF are asymptotically chi-squared distributed with one degree of freedom. If the value of the TUFF statistic exceeds the critical value of the chi-square distribution, we reject the null hypothesis.

This test is ideal as a simple preliminary to the POF-test when there is no larger set of data available. However, this test only considers the number of violations but ignores the time dynamics of violations. In addition, the test also exhibits a low power in identifying poor VaR models.

2.5 Proportion of failures test (Haas, 2001)

Suppose there are x violations in a sample of size n. Then the maximum likelihood estimator of p is $\widehat{p} = x/n$ and its variance is

$$\frac{\widehat{p}\,(1-\widehat{p})}{n}.$$

An approximate $100(1-\alpha)$ percent confidence interval for p is:

$$\left(\widehat{p} - z_{\alpha/2}\frac{\widehat{p}\,(1-\widehat{p})}{n}, \widehat{p} + z_{\alpha/2}\frac{\widehat{p}\,(1-\widehat{p})}{n}\right).$$

If p lies within this interval, we can consider VaR to be a good model. Otherwise, we can try to evaluate what is the true confidence level rendered by the model.

2.6 Time between failures likelihood ratio test (Haas, 2001)

Haas (2001) proposed a test for the null hypothesis that the violations are independent of each other. He suggested the likelihood ratio statistic

$$LR = \sum_{i=2}^{m}\left[-2\ln\left(\frac{p(1-p)^{v_i-1}}{\widehat{p}\,(1-\widehat{p})^{v_i-1}}\right)\right] - 2\ln\left[\frac{p(1-p)^{v-1}}{\widehat{p}\,(1-\widehat{p})^{v-1}}\right],$$

where v_i denotes the time gap between the $(i-1)$th violation and ith violation, and m denotes the number of violations. The statistic is asymptotically chi-squared distributed with m degrees of freedom. If the LR value exceeds the critical value of the chi-squared distribution, we reject the null hypothesis.

An advantage of this test is that it is very robust, since it can identify both problems with dependencies and the number of violations.

2.7 Scaled Crnkovic and Drachman (1996)'s method (Haas, 2001)

Suppose there are n backtesting points, falling within r disjoint intervals. Let $p_1, p_2, \ldots, p_r$ denote the corresponding hitting probabilities. Haas (2001) suggested a test for

$$H_0 : (p_1, p_2, \ldots, p_r) = \left(p_1^0, p_2^0, \ldots, p_r^0\right)$$

versus

$$H_a : (p_1, p_2, \ldots, p_r) \neq \left(p_1^0, p_2^0, \ldots, p_r^0\right).$$

The test statistic is

$$Q = \sum_{i=1}^{r} \frac{\left(y_i - np_i^0\right)^2}{np_i^0} = \sum_{i=1}^{r} \frac{y_i^2}{np_i^0} - n,$$

where $y_1, y_2, \ldots, y_r$ denote the number of backtesting points falling within the r intervals. The asymptotic null distribution of Q is the chi-squared distribution with $r-1$ degrees of freedom.

This method is a new extension to Crnkovic and Drachman (1996)'s test and it no longer judges a continuous distribution, but a discrete one when testing. This method provides a new graphical analysis over the entire VaR model with or without focusing on the tails of the data. However, a limitation of this test occurs when analysing the entire VaR model, as it can easily be confused by dependencies within the data.

2.8 Density forecast tests (Berkowitz, 2000, 2001)

Let $Y_t = F_{t-1}(x_t)$ and $Z_t = \Phi^{-1}(Y_t)$ for $t = 1, 2, \ldots, n$. According to Berkowitz (2000, 2001), the test of validity of VaR model corresponds to Y_t being a random sample from a uniform $[0, 1]$ distribution. The test of independence corresponds to Z_t begin a random sample from a standard normal distribution versus an alternative such as

$$Z_t = \mu + \rho_1 Z_{t-1} + \cdots + \rho_n Z_{t-n} + \gamma_1 Z_{t-1}^2 + \cdots + \gamma_m Z_{t-m}^2 + \mu_1.$$

Both these tests can be performed by the likelihood ratio principle.

This method pays specific attention to the left tail of the distribution and shows that it has particular merit in the backtesting of risk models when the left tail contains the largest losses.

2.9 Lopez's magnitude loss function (Lopez, 1998, 1999)

In contrast to the other tests, Lopez (1998, 1999) proposed to examine the distance between the observed returns and the forecasted VaR (α) when a violation occurs. He introduced a loss function representing the total number of violations and their squared distance from the corresponding VaR:

$$L_t\left(x_t, \mathrm{VaR}_t(\alpha)\right) = \begin{cases} 1 + \left[x_t - \mathrm{VaR}_t(\alpha)\right]^2, & x_t \leq \mathrm{VaR}_t, \\ 0, & x_t > \mathrm{VaR}_t. \end{cases}$$

Given a data set on profits / losses, the sample average is

$$\widehat{L} = \frac{1}{n} \sum_{t=1}^{n} L_t\left(x_t, \mathrm{VaR}_t(\alpha)\right).$$

The distribution of $\widehat{L}$ can be simulated and thus a rejection rule can be determined.

This test is ideal for determining and comparing whether one VaR model provides a better risk assessment than another competing VaR model. Therefore, this test is noted to be better at comparing competing VaR models rather than judging the accuracy of a

single model. Also, the test is very flexible and can address specific concerns that may be of interest when analysing the performance of a VaR model. However, this flexibility comes at a price of needing an increase in the information burden associated with assessing the accuracy of VaR model. Lopez (1998, 1999) recognized some disadvantages with backtesting based on the loss function. According to the nature of backtesting, we cannot determine the accuracy of the model by this method.

2.10 Risk map (Colletaz et al., 2013)

The risk map presents the backtesting results for a given risk model graphically. Colletaz et al. (2013) proposed to test the number of VaR violations and super violations through the following null hypothesis:

$$H_0 : E\left[I_t(\alpha)\right] = \alpha \text{ and } E\left[I_t\left(\alpha'\right)\right] = \alpha',$$

where $0 < \alpha' < \alpha < 1$. The test statistic is

$$LR = -2\ln\left[\frac{(1-\alpha)^{N_0}\left(\alpha-\alpha'\right)^{N_1}\left(\alpha'\right)^{N_2}}{\left(\frac{N_0}{n}\right)^{N_0}\left(\frac{N_1}{n}\right)^{N_1}\left(\frac{N_2}{n}\right)^{N_2}}\right],$$

where N_2 denotes the number of violations for $\text{VaR}_{\alpha'}$, N_1 denotes the number of occurrences of a loss between -VaR_{α} and $\text{VaR}_{\alpha'}$, and N_0 denotes the number of occurrences of a loss lower than -VaR_{α}. The asymptotic null distribution of the statistic is the chi-squared distribution with two degrees of freedom.

This test is very simple to implement and use. It can be applied to any tail risk model and is a very effective model for the banking industry to use. The framework enables easy validation of market risk, credit risk or operational risk estimates.

2.11 Independence test (Christoffersen, 1998)

This test is also known as the Markov test and it examines the independence property, that is, the test examines if the probability of VaR violation on any given day depends on the outcome of the previous day. The likelihood ratio principle is used for testing.

Suppose we have data on portfolio returns for n days. Each day we set the indicator value as follows:

$$I_t = \begin{cases} 0, & \text{if VaR is not violated,} \\ 1, & \text{otherwise.} \end{cases}$$

We define $N_{i,j}$, $i = 0, 1$, $j = 0, 1$ as the number of days in which state j occurred on one day while state i occurred on the previous day. For example, $N_{1,0}$ is the number of days for which $I_n = 0$ and $I_{n-1} = 1$. The following 2×2 contingency table represents all possible outcomes:

	$I_{n-1}=0$	$I_{n-1}=1$	
$I_n = 0$	N_{00}	N_{10}	N_{00}+N_{10}
$I_n = 1$	N_{01}	N_{11}	N_{01}+N_{11}
	N_{00}+N_{01}	N_{10}+N_{11}	N

Let π_0 be the conditional probability of 01 occurring if the previous day was 0. Let π_1 be the conditional probability of 11 occurring if the previous day was 1. It follows that $\pi_0 = \frac{N_{01}}{N_{00}+N_{01}}$ and $\pi_1 = \frac{N_{11}}{N_{10}+N_{11}}$. Let $\pi = \pi_0 + \pi_1$.

The test statistic for independence of violations is

$$LR = -2\ln\left[(1-\pi)^{N_{00}+N_{01}}\pi^{N_{01}+N_{11}}\right] + 2\ln\left[(1-\pi_0)^{N_{00}}\,\pi_0^{N_{01}}1-\pi_1{}^{N_{10}}\pi_1^{N_{11}}\right].$$

The asymptotic null distribution is the chi-squared distribution with one degree of freedom.

Under the null hypothesis, the probabilities should be equal, i.e. $\pi_0 = \pi_1$. Therefore, if the proportions differ greatly from each other, we know that the VaR model is unreliable and should be checked.

Haas (2001) argues that this test is too weak to produce feasible results. Also, it has limited power against general forms of time dependence in violations.

2.12 Joint test

The joint test is a combination of the independence statistic with Kupiec's POF - test. The test not only measures the correct failure rate but also the independence of violations. The test statistic is

$$LR = LR_{ind} + POF,$$

where

$$LR_{Ind} = -2\ln\left[(1-\pi)^{N_{00}+N_{01}}\pi^{N_{01}+N_{11}}\right] + 2\ln\left[(1-\pi_0)^{N_{00}}\,\pi_0^{N_{01}}1-\pi_1{}^{N_{10}}\pi_1^{N_{11}}\right]$$

and POF is as defined in (3). The asymptotic null distribution is the chi-squared distribution with $n+1$ degrees of freedom.

An accurate VaR measure must display both the unconditional coverage and independence properties. Therefore this joint test is ideal. However, it has a reduced ability to detect a VaR measure which only violates one of the two properties. If one of the two properties is satisfied, the joint test finds it more difficult to detect the inadequacy of the VaR measure.

2.13 Generalized Markov tests (Pajhede, 2015)

Pajhede (2015) supposed that the distribution of $I_t(\alpha)$ conditional on $\{I_{t-1}(\alpha), I_{t-2}(\alpha), \ldots, I_{t-k}(\alpha)\}$ is a Bernoulli distribution with parameter

$$p_t(\theta) = J_{t-1}p_E + (1-J_{t-1})\,p_S,$$

where p_E is an *excited probability*, p_S is a *steady probability*, and

$$J_{t-1} = I\left\{\sum_{i=1}^{k} I_{t-1} > 0\right\}.$$

Under independence, $p_E = p_S = \phi$ say. Under conditional coverage, $p_t(\theta) = \alpha$ for all t.

Pajhede (2015) developed the following statistics for testing independence and conditional coverage:

$$\begin{aligned} LR_1 &= -2\Big[\ln\left(1-\widehat{\phi}\right)(T_{0,0}+T_{1,0})+\ln\widehat{\phi}\,(T_{0,1}+T_{1,1})-\ln(1-\widehat{p_S})\,T_{0,0} \\ &\quad -\ln\widehat{p_S}T_{0,1}-\ln(1-\widehat{p_E})\,T_{1,0}-\ln\widehat{p_E}T_{1,1}\Big] \end{aligned}$$

and

$$\begin{aligned} LR_2 &= -2\Big[\ln(1-\alpha)(T_{0,0}+T_{1,0})+\ln\alpha\,(T_{0,1}+T_{1,1})-\ln(1-\widehat{p_S})\,T_{0,0} \\ &\quad -\ln\widehat{p_S}T_{0,1}-\ln(1-\widehat{p_E})\,T_{1,0}-\ln\widehat{p_E}T_{1,1}\Big], \end{aligned}$$

respectively, where $\left(\widehat{\phi},\widehat{p_E},\widehat{p_S}\right)$ are maximum likelihood estimates of (ϕ,p_E,p_S) and

$$\begin{aligned} T_{1,1} &= \sum_{i=1}^{n} I_t J_{t-1}, \quad T_{0,1} = \sum_{i=1}^{n} I_t\,(1-J_{t-1}), \\ T_{1,0} &= \sum_{i=1}^{n} (1-I_t)\,J_{t-1}, \quad T_{0,0} = \sum_{i=1}^{n} (1-I_t)\,(1-J_{t-1}). \end{aligned}$$

Pajhede (2015) established that the asymptotic null distributions of LR_1 and LR_2 are chi-squared distributions with one and two degrees of freedom, respectively.

This test allows for a higher or kth order dependence than the independence test (Christoffersen, 1998).

2.14 The Basel Committee's traffic light test

The traffic light approach was proposed by the Basel Committee on Banking Supervision (1996, 2006, 2011). Its purpose was to outline a framework for backtesting VaR to be used by financial institutions. The following methodology is commonly used by many banks to test their internal models.

Let x denote the number of violations in the previous 250 trading days. If $x \leq 4$ then VaR model is considered accurate. If $5 \leq x \leq 9$ then the model may be assumed accurate or inaccurate. The committee will require further tests to check the model. If $x \geq 10$ then VaR model is considered to be problematic and a new model is needed. These three are referred to as the green zone, yellow zone and red zone, respectively.

This test is very simple and is best used as a preliminary check for the accuracy of VaR. The drawback of this test is that it does not allow us to evaluate the suitability of a VaR model as it does not take into account the independence of violations. Also, the test does not have the ability to allow us to compare different models.

2.15 Quality control of risk measure (de la Pena et al., 2006)

Let p denote the probability of a violation, p_0 its value when VaR model is correct, and p_1 its value when VaR model is incorrect.

de la Pena et al. (2006) considered testing $H_0 : p > p_1$ versus $H_a : p \leq p_0$. A simple test with significance level α is to reject the null hypothesis if

$$p_1 < \widehat{p} - z_\alpha \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}},$$

where $\widehat{p}$ denotes an estimate of p.

By analogy to the Basel supervisory framework, de la Pena et al. (2006) also stated the following:

- VaR model is certified as correct if

$$\widehat{p} - z_{0.05} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} < p_0 < 1;$$

- the validity of VaR model is questioned if

$$p_0 < \widehat{p} - z_{0.05} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$

  and

$$\widehat{p} - z_{0.01} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} < p_0 < 1;$$

- VaR model is rejected if

$$p_0 < \widehat{p} - z_{0.01} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}.$$

These three are also referred to as the green zone, yellow zone and red zone, respectively.

These methods, when compared with the Wald statistic test (Engle and Managanelli, 2004), have more power in finite samples. Most importantly, this approach can be easily computed through standard quantile regression software.

2.16 Duration - based tests (Christoffersen and Pelletier, 2004; Berkowitz et al., 2011)

The main idea behind duration based tests is that clustering of violations will result in a large number of relatively short and long no hit durations, corresponding to market turbulence and market calm.

Denote the number of days between two VaR(α) violations by

$$d_i = t_i - t_{i-1},$$

where t_i denotes the day of violation i.

Suppose the d_i's follow the exponential distribution specified by the probability density function

$$f(d) = p \exp(-pd) \tag{4}$$

for $d > 0$ and $p > 0$. Then, the null hypothesis that the risk model is correctly specified corresponds to $p = \alpha$ (Christoffersen and Pelletier, 2004).

Suppose now that the d_i's follow the Weibull distribution specified by the probability density function

$$f(d) = bp^b d^{b-1} \exp\left[-(pd)^b\right]$$

for $d > 0$, $b > 0$ and $p > 0$. Then, the null hypothesis that the risk model is correctly specified corresponds to $b = 1$ and $p = \alpha$ (Berkowitz et al., 2011). The null hypothesis of independence of violations corresponds to $b = 1$ (Berkowitz et al., 2011).

All of these tests can be based on the likelihood ratio principle. The implementation of these tests is very straightforward and provides a clear interpretation of parameters. However, these tests are not very popular among practitioners. The main drawback of these tests is that they have relatively small power for realistic sample sizes (Haas, 2007) and normally one would struggle in computing the standard LR duration based statistics.

2.17 Haas (2006)'s test

Haas (2006) supposed that the time in days between two violations follows a discrete Weibull distribution specified by the probability mass function

$$p(d) = \exp\left[-a^b(d-1)^b\right] - \exp\left[a^b d^b\right]$$

for $d = 1, 2, \ldots$, $a > 0$ and $b > 0$. The null hypothesis of the correct conditional probability α corresponds to $b = 1$ and $a = -\ln(1-\alpha)$. The null hypothesis of independence corresponds to $b = 1$. These hypotheses can be tested by the likelihood ratio test.

2.18 Generalized method of moments (GMM) duration - based test (Hurlin et al., 2010)

Let $d_1, d_2, \ldots, d_m$ be durations observed between two successive violations associated with α percent VaR forecasts. Hurlin et al. (2010) showed that the null hypothesis of conditional coverage (that is, $E\left[I_t(\alpha) \mid \Omega_{t-1}\right] = \alpha$) is equivalent to

$$H_0 : E\left[M\left(d_i; \alpha\right)\right] = 0$$

for $i = 1, 2, \ldots, m$, where $M\left(d_i; \beta\right) = \left[M_1(d; \beta), M_2(d; \beta), \ldots, M_p(d; \beta)\right]^T$ and $M_j(d; \beta)$ are orthonormal polynomials defined by the recurrence relation

$$M_{j+1}\left(d; \beta\right) = \frac{(1-\beta)(2j+1) + \beta(j-d+1)}{(j+1)\sqrt{1-\beta}} M_j(d; \beta) - \frac{1}{j+1} M_{j-1}(d; \beta)$$

with initial conditions $M_{-1}(d; \beta) = 0$ and $M_0(d; \beta) = 1$. The test statistic is

$$\left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} M\left(d_i; \alpha\right)\right)^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} M\left(d_i; \alpha\right)\right).$$

Its asymptotic null distribution is the chi-squared distribution with p degrees of freedom.

Hurlin et al. (2010) also showed that the null hypothesis of independence of violations with d_i geometrically distributed with rate parameter β is equivalent to

$$H_0 : E\left[M\left(d_i;\beta\right)\right] = 0$$

for $i = 1, 2, \ldots, m$. The test statistic is

$$\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m} M\left(d_i;\beta\right)\right)^T \left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m} M\left(d_i;\beta\right)\right).$$

Its asymptotic null distribution is the chi-squared distribution with p degrees of freedom.

This test is extremely easy to implement, and requires only a few constraints to improve the feasibility of the backtesting tests. When using Monte-Carlo simulations, this test shows that it has relatively high power properties. Overall, this test improves the feasibility and the power of traditional duration based tests. Also, the GMM statistics can be numerically computed for any sample size. A similar test based on GMMs is given in Bontemps (2014).

2.19 Conditional duration test (Christoffersen and Pelletier, 2002)

Christoffersen and Pelletier (2002) proposed the conditional duration model $E\left[d_i\right] = w + \beta d_{i-1}$ with an underlying unit exponential distribution. The null hypothesis of independent no-hit durations corresponds to $\beta = 0$. The null hypothesis of unconditional duration allowing for dependence corresponds to $\frac{w}{1-\beta} = \frac{1}{\alpha}$. These can be tested by the likelihood ratio principle.

This test has better power properties than the other listed tests and the size of the test is easily controlled through finite sample critical values.

2.20 Krämer and Wied (2015)'s test

Suppose there are m violations with time gaps $d_1, \ldots, d_m$ in a series of n observations. Krämer and Wied (2015) suggested testing the independence hypothesis using the statistic

$$T = \sqrt{n}\left[\frac{1}{2m^2\overline{d}}\sum_{i,j=1}^{m}\left(d_i - d_j\right) - \frac{1-\frac{m}{n}}{2-\frac{m}{n}}\right],$$

where $\overline{d}$ denotes the sample mean. The hypothesis can be rejected for large values of T, and critical values can be determined by simulation.

This test exhibits higher power than others against many deviations from independence of VaR violations. In addition, the test rejects for large values of the Gini coefficient of durations between VaR violations.

2.21 Markov duration tests (Pajhede, 2015)

Pajhede (2015) supposed that the distribution of $I_t(\alpha)$ conditional on $\{I_{t-1}(\alpha), I_{t-2}(\alpha), \ldots, I_{t-k}(\alpha)\}$ is a Bernoulli distribution with parameter

$$p_t(\theta) = J(1)_{t-1}p_{E_1} + \cdots + J(k)_{t-1}p_{E_k} + \left[1 - \sum_{i=1}^{k} J(i)_{t-1}\right] p_S,$$

where

$$J(1)_{t-1} = I\left\{I_{t-1} = 1\right\}, \ldots, J(k)_{t-1} = I\left\{I_{t-1} = 0, \ldots, I_{t-k} = 1\right\},$$

$$p_{E_1} = \Pr\left(I_t = 1 \mid I_{t-1} = 1\right), \ldots, p_{E_k} = \Pr\left(I_t = 1 \mid I_{t-1} = 0, \ldots, I_{t-k} = 1\right),$$

and $p_S = \Pr\left(I_t = 1 \mid I_{t-1} = 0, \ldots, I_{t-k} = 0\right)$. Under independence, $p_{E_1} = \cdots = p_{E_k} = p_S = \phi$ say. Under conditional coverage, $p_t(\theta) = \alpha$ for all t.

Pajhede (2015) developed the following statistics for testing independence and conditional coverage:

$$\begin{aligned} LR_1 &= -2\Big[\ln\left(1-\widehat{\phi}\right)(T_{0,0}+T_{1,0})\ln\widehat{\phi}\,(T_{0,1}+T_{1,1}) - \ln\left(1-\widehat{p_S}\right)T_{0,0} \\ &\quad - \ln\widehat{p_S}T_{0,1} - \sum_{i=1}^{k}\ln\left(1-\widehat{p_{E_i}}\right)T_{1,0}(i) - \sum_{i=1}^{k}\ln\widehat{p_{E_i}}T_{1,1}(i)\Big] \end{aligned}$$

and

$$\begin{aligned} LR_2 &= -2\Big[\ln\left(1-\alpha\right)(T_{0,0}+T_{1,0})\ln\alpha\,(T_{0,1}+T_{1,1}) - \ln\left(1-\widehat{p_S}\right)T_{0,0} \\ &\quad - \ln\widehat{p_S}T_{0,1} - \sum_{i=1}^{k}\ln\left(1-\widehat{p_{E_i}}\right)T_{1,0}(i) - \sum_{i=1}^{k}\ln\widehat{p_{E_i}}T_{1,1}(i)\Big], \end{aligned}$$

respectively, where $\left(\widehat{\phi}, \widehat{p_{E_1}}, \ldots, \widehat{p_{E_k}}, \widehat{p_S}\right)$ are maximum likelihood estimates of $(\phi, p_{E_1}, \ldots, p_{E_k}, p_S)$ and $T_{1,1}, T_{0,1}, T_{1,0}, T_{0,0}$ are as defined in Section 2.13. Pajhede (2015) established that the asymptotic null distributions of LR_1 and LR_2 are chi-squared distributions with $k-1$ and k degrees of freedom, respectively.

Pajhede (2015) found evidence of improved size properties for the generalized Markov test compared to the original Markov test of Christoffersen (1998), but worse size properties for the Markov duration test.

2.22 Escanciano and Olmo (2010, 2011)'s tests

Escanciano and Olmo (2010) considered a test of the null hypothesis $E\left[I_t(\alpha;\theta) \mid \Omega_{t-1}\right] = \alpha$, where θ are parameters of the VaR model. Let R denote the first few observations in the sample used to estimate the parameters in the first forecast and let $P = n - R$ denote the number of predictions. The statistic for the test is

$$\frac{1}{\sqrt{P}}\sum_{t=R+1}^{n}\left[I_t\left(\alpha;\widehat{\theta}_t\right) - \alpha\right],$$

where $\widehat{\theta}_t$ denotes an estimate of θ at time t. The asymptotic null distribution is normal with zero mean and a certain variance, see Corollary 1 in Escanciano and Olmo (2010).

Escanciano and Olmo (2011) considered tests for joint and marginal independence of $\{I_t(\alpha;\theta)\}$. The respective statistics are

$$\frac{1}{\sqrt{P-j}}\sum_{t=R+j+1}^{n}\left[I_t\left(\alpha;\widehat{\theta}_{t-1}\right) - \alpha\right]\left[I_{t-j}\left(\alpha;\widehat{\theta}_{t-j-1}\right) - \alpha\right]$$

and

$$\frac{1}{\sqrt{P-j}}\sum_{t=R+j+1}^{n}\left\{I_t\left(\alpha;\widehat{\theta}_{t-1}\right)-E\left[I_t\left(\alpha;\widehat{\theta}_{t-1}\right)\right]\right\}\left\{I_{t-j}\left(\alpha;\widehat{\theta}_{t-j-1}\right)-E\left[I_{t-j}\left(\alpha;\widehat{\theta}_{t-j-1}\right)\right]\right\}.$$

The asymptotic null distributions of both statistics are normal with zero means and certain variances, see Theorem 2 in Escanciano and Olmo (2011).

These tests do not consider the impact of estimation risk and therefore may use the wrong critical values to assess market risk.

2.23 Wald statistic test (Engle and Managanelli, 2004)

Engle and Managanelli (2004) used a linear regression model approach to perform backtesting on VaR. They defined a *de-meanded process* as

$$\mathrm{Hit}_t(\alpha)=I_t(\alpha)-\alpha=\begin{cases}1-\alpha, & \text{if } x_t<\mathrm{VaR}_t(\alpha),\\ -\alpha, & \text{otherwise.}\end{cases}$$

The following linear regression model is regressed over historical data and available information at $t-1$:

$$\mathrm{Hit}_t(\alpha)=\delta+\sum_{s=1}^{K}\beta_s\mathrm{Hit}_{t-s}(\alpha)+\sum_{s=1}^{K}\gamma_s g\left[\mathrm{Hit}_{t-s}(\alpha),\mathrm{Hit}_{t-s-1}(\alpha),\ldots,z_{t-s},z_{t-s-1},\ldots\right]+\eta_t,$$

where η_t are independently and identically distributed, $g(\cdot)$ is a function of past violations and of variables z_{t-k} from the available information set $_{t-1}$, and δ_s, γ_s are coefficients.

Testing the null hypothesis of conditional efficiency corresponds to testing the joint nullity of the coefficients and the constant δ:

$$H_0:\delta=\beta_s=\gamma_s=0$$

for $s=1,2,\ldots,K$. The null hypothesis states that the current VaR violations are uncorrelated with past violations. The unconditional coverage hypothesis is verified when δ is null.

The likelihood ratio test statistic is

$$LR=\frac{\widehat{\varphi}^T Z^T Z\widehat{\varphi}}{\alpha(1-\alpha)},$$

where $\widehat{\varphi}$ denotes an estimator of $\varphi=(\delta,\beta_1,\ldots,\beta_K,\gamma_1,\ldots,\gamma_K)^T$ and Z denotes a matrix of explanatory variables. The asymptotic null distribution as $n\to\infty$ is the chi-squared distribution with $2K+1$ degrees of freedom.

The linear regression model is not the most appropriate choice to infer on the parameters and consequently on the validity of VaR. Engle and Managanelli (2004) considered an extension based on a binary (probit or logit) model linking current and past violations.

2.24 Dynamic binary tests (Dumitrescu et al., 2012)

Dumitrescu et al. (2012) suggested a logistic regression model instead of a linear regression model which can test the efficiency assumption of the VaR model. They specify the conditional probability of violation at time t by

$$\Pr\left[I_t(\alpha) = 1 \mid |\Omega_{t-1}\right] = F\left(\pi_t\right),$$

where $F(\cdot)$ denotes a cumulative distribution function and π_t satisfies the following autoregressive representation

$$\pi_t = c + \sum_{i=1}^{p_1} \beta_i \pi_{t-i} + \sum_{i=1}^{p_2} \tau_i I_{t-i}(\alpha) + \sum_{i=1}^{p_3} \theta_i l\left(x_{t-i}, \phi\right) + \sum_{i=1}^{p_4} \gamma_i l\left(x_{t-i}, \phi\right) I_{t-i},$$

where $l(\cdot)$ denotes a function of a finite number of lagged values of observables, and x_t denotes a vector of explicative variables.

The corresponding log-likelihood function can be written as

$$\ln L\left(\theta; I(\alpha), Z\right) = \sum_{t=1}^{n} \left\{I_t(\alpha) \ln F\left(\pi_t\left(\theta, Z_t\right)\right) + \left[1 - I_t(\alpha)\right] \ln\left[1 - F\left(\pi_t\left(\theta, Z_t\right)\right)\right]\right\},$$

where $\theta = \left(\beta^T, \gamma^T, \psi^T, \delta^T\right)$ and Z_t denotes the vector of explanatory variables at time t.

For the null hypothesis of conditional coverage, $H_0 : \beta = 0, \tau = 0, \theta = 0, \gamma = 0, c = F^{-1}(\alpha)$, the conditional probability of a violation is

$$\Pr\left[I_t = 1 \mid \Omega_{t-1}\right] = F\left(F^{-1}(\alpha)\right) = \alpha.$$

The likelihood ratio test statistic is

$$LR = -2\left\{\ln L\left(0, F^{-1}(\alpha); I_t(\alpha), Z_t\right) - \ln L\left(\widehat{\theta}, \widehat{c}; I_t(\alpha), Z_t\right)\right\},$$

where $\widehat{\theta}$ and $\widehat{c}$ are estimates of θ and c, respectively. The asymptotic null distribution is the chi-squared distribution with degrees of freedom equal to the dimension of Z_t.

For the null hypothesis of independence, $H_0 : \beta = 0, \tau = 0, \theta = 0, \gamma = 0$, the conditional probability of a violation is

$$\Pr\left[I_t = 1 \mid \Omega_{t-1}\right] = F(c),$$

where c is now a free parameter. The likelihood ratio test statistic is similar, except now the asymptotic null distribution is the chi-squared distribution with degrees of freedom equal to one minus the dimension of Z_t.

These tests use an appropriate non-linear methodology for the binary dependent variables during backtesting and this will improve the finite sample properties of the backtesting tests. Other benefits of these tests are that they allow us to test both hypotheses of independence and conditional coverage.

2.25 Multivariate test (Perignon and Smith, 2008)

Perignon and Smith (2008) proposed a multivariate unconditional coverage test. Let $p_1 > p_2 > \cdots > p_K$ be K coverage probabilities and let $\mathrm{VaR}_t(p_1) < \mathrm{VaR}_t(p_2) < \cdots < \mathrm{VaR}_t(p_K)$ be the corresponding VaRs. Let

$$J_{i,t+1} = \begin{cases} 1, & \text{if } -\mathrm{VaR}_t(p_{i+1}) < x_{t+1} - \mathrm{VaR}_t(p_i), \\ 0, & \text{otherwise} \end{cases}$$

with the convention $p_{K+1} = 0$, $\mathrm{VaR}_t(p_{K+1}) = \infty$ and

$$J_{0,t+1} = \prod_{i=1}^{K} (1 - J_{i,t+1}).$$

Let $\theta_i = p_i - p_{i+1}$, $\theta = (\theta_1, \theta_2, \ldots, \theta_K)^T$ and $n_i = J_{i,1} + J_{i,2} + \cdots + J_{i,n}$.

A likelihood ratio test statistic for the hypothesis of unconditional coverage is

$$LR = 2\left[n_0 \ln\left(1 - 1^T\widehat{\theta}\right) + \sum_{k=1}^{K} n_i \ln \widehat{\theta}_i\right] - 2\left[n_0 \ln\left(1 - 1^T\theta\right) + \sum_{k=1}^{K} n_i \ln \theta_i\right],$$

where $\widehat{\theta}_i = n_i/n$ denotes the maximum likelihood estimate of θ_i. The asymptotic null distribution of the statistic is the chi-squared distribution with K degrees of freedom.

This backtesting method, based on multiple points on the left tail of trading revenue, gives improvements over the ability of univariate tests to reject misspecified VaR models.

2.26 Martingale difference test (Berkovitz et al., 2005)

Berkovitz et al. (2005) noted that the unconditional coverage and independence hypotheses are consequences of the martingale difference hypothesis of $\mathrm{Hit}_t(\alpha) = I_t(\alpha) - \alpha$. For the latter, Berkovitz et al. (2005) proposed a test based on the univariate Ljung-Box statistic – a statistic testing nullity of the first K autocorrelations of $\mathrm{Hit}_t(\alpha)$:

$$n(n+2)\sum_{i=1}^{K} \frac{\widehat{r}_i^2}{n-i}, \tag{5}$$

where $\widehat{r}_i$ is the empirical autocorrelation of order i of the $\mathrm{Hit}_t(\alpha)$ process. The asymptotic null distribution of the statistic as $n \to \infty$ is the chi-squared distribution with K degrees of freedom.

The martingale difference hypothesis can also be tested using the spectral density estimate

$$\widehat{f}(w) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \widehat{r}_k \exp(-\mathrm{i}kw),$$

where $\mathrm{i} = \sqrt{-1}$ denotes the complex unit. Two possible test statistics are

$$\sum_{w=0}^{\pi} \left[\frac{\widehat{f}(w)}{\widehat{\sigma}^2} - \frac{1}{\pi}\right]^2$$

and

$$\sup_{w}\left[\frac{\widehat{f}(w)}{\widehat{\sigma}^2}-\frac{1}{\pi}\right]^2,$$

where $\widehat{\sigma}^2$ denotes the variance associated with the spectral density. Under the null hypothesis, both statistics converge to known distributions, the latter to the distribution of the Kolmogorov-Smirnov statistic.

2.27 Multivariate autocorrelations test (Hurlin and Tokpavi, 2006)

Hurlin and Tokpavi (2006) proposed a multivariate extension of the test of absence of autocorrelation of violations proposed by Berkowitz et al. (2005) Let $\Theta = \{\theta_1, \ldots, \theta_m\}$ be a discrete set of m different coverage rates, strictly between 0 and 1. Let $\text{Hit}_t = [\text{Hit}_t(\theta_1) : \text{Hit}_t(\theta_2) : \cdots : \text{Hit}_t(\theta_m)]$ be a vector re-grouping the violation sequences associated with these m coverage rates, at time t, $\theta_1, \theta_2, \ldots, \theta_m$. The null hypothesis corresponding to the joint nullity by the first K autocorrelations of Hit_t is equivalent to

$$H_0 : E\left[\text{Hit}_t(\theta_i)\,\text{Hit}_{t-k}(\theta_j)^T\right] = 0$$

for all $k = 1, 2, \ldots, K$, for all θ_i and all θ_j.

For testing of this hypothesis, Hurlin and Tokpavi (2006) proposed the statistic

$$T = n\sum_{k=1}^{K}\text{vec}\left(\widehat{R}_k\right)^T\left(\widehat{R}_0^{-1}\otimes\widehat{R}_0^{-1}\right)\text{vec}\left(\widehat{R}_k\right),$$

where

$$\widehat{R}_k = D\widehat{C}_k D,$$

$$\widehat{C}_k = (\widehat{c}_{i,j,k}) = \sum_{t=k+1}^{n}\text{Hit}_t\text{Hit}_{t-k}^T$$

and D is a diagonal matrix with the ith diagonal element equal to $\sqrt{\widehat{c}_{i,i,0}}$ for $i = 1, 2, \ldots, m$. The asymptotic null distribution of T as $n \to \infty$ is the chi-squared distribution with degrees of freedom equal to Km^2.

Overall the shift to a multivariate dimension significantly improves the power properties of the VaR validation test for large sample sizes and this test is very easy to implement.

2.28 Geometric-VaR backtesting method (Pelletier and Wei, 2015)

Pelletier and Wei (2015) proposed the geometric VaR test comprising of three individual hypotheses under one unified framework. The three individual tests are: a test of unconditional coverage, a test of duration independence and a test of VaR independence. The method combines the geometric test with the VaR test and specifies the following hazard function:

$$\Pr\left(I_{t_i+d} = 1 \mid \Omega_{t_i+d-1}\right) = ad^{b-1}e^{c\ \text{VaR}_{t_i+d}}$$

for $0 \leq a < 1$ and $c \geq 0$. Here, the parameter a captures the unconditional coverage and d^{b-1} describes duration dependence. Under the null hypothesis that VaR is correctly specified, duration follows a geometric distribution with parameter p, so the null corresponds to $a = p$, $b = 1$, and $c = 0$.

The three individual tests and three joint tests can be specified explicitly as follows

- Unconditional coverage test (under the assumption that $b = 1$ and $c = 0$):

$$H_0 : a = p$$
$$H_a : a \neq p;$$

- Duration independence test (under the assumption that $c = 0$):

$$H_0 : b = 1$$
$$H_a : b < 1;$$

- VaR independence test:

$$H_0 : c = 0$$
$$H_a : c > 1;$$

- Geometric test - unconditional coverage and duration independence (under the assumption that $c = 0$):

$$H_0 : a = p \text{ and } b = 1$$
$$H_a : a \neq p \text{ and } b < 1;$$

- VaR test - unconditional coverage and VaR independence (under the assumption that $b = 0$):

$$H_0 : a = p \text{ and } c = 0$$
$$H_a : a \neq p \text{ and } c > 0;$$

- Geometric - VaR test (unconditional coverage, duration independence and VaR independence):

$$H_0 : a = p, b = 1 \text{ and } c = 0$$
$$H_a : a \neq p, b < 1 \text{ and } c > 0.$$

Pelletier and Wei (2015) not only test whether VaR forecasts are misspecified in general, but also examine how they are misspecified by looking into which individual hypothesis is rejected. The test statistic is:

$$LR = LR^{UC} + LR^{Dind} + LR^{Vind},$$

where

$$LR^{UC} = -2\left[\ln L(a = p, b = 1, c = 0) - \ln L\left(\widehat{a}, b = 1, c = 0\right)\right],$$
$$LR^{Dind} = -2\left[\ln L\left(\widehat{a}, b = 1, c = 0\right) - \ln L\left(\widehat{a}, \widehat{b}, c = 0\right)\right],$$
$$LR^{Vind} = -2\left[\ln L\left(\widehat{a}, \widehat{b}, \widehat{c}\right) - \ln L\left(\widehat{a}, \widehat{b}, c = 0\right)\right],$$

where L denotes the likelihood function. The sampling distribution and the rejection rule can be determined by simulation.

The Geometric - VaR test has better power than other duration - based tests or regression - based tests, and offers a good alternative for detecting various forms of VaR misspecification. This test not only tests whether the VaR forecast is misspecified, but also helps understand how the VaR forecast is misspecified.

3 Computer software

Software packages implementing backtesting methods for VaR are widely available. Some packages available for the R software (R Development Core Team, 2016) are:

- fPortfolio due to Rmetrics Core Team, Diethelm Wuertz, Tobias Setz and Yohan Chalabi. According to the authors, this package provides an "environment for teaching "Financial Engineering and Computational Finance"".
- rmgarch and rugarch due to Alexios Ghalanos. These packages in particular implement the duration based tests due to Christoffersen and Pelletier (2004). According to the author, rmgarch provides "ARFIMA, in-mean, external regressors and various GARCH flavors, with methods for fit, forecast, simulation, inference and plotting" and rugarch "makes multivariate GARCH models including DCC, GO-GARCH and Copula-GARCH feasible".
- rcss due to Juri Hinz and Jeremy Yee. According to the authors, this package provides "the numerical treatment of optimal switching problems in a finite time setting when the state evolves as a controlled Markov chain consisting of an uncontrolled continuous component following linear dynamics and a controlled Markov chain taking values in a finite set. The reward functions are assumed to be convex and Lipschitz continuous in the continuous state. The action set is finite".
- Dowd due to Dinesh Acharya. According to the author, "Kevin Dowd's book Measuring Market Risk is a widely read book in the area of risk measurement by students and practitioners alike. As he claims, 'MATLAB' indeed might have been the most suitable language when he originally wrote the functions, but, with growing popularity of R it is not entirely valid. As Dowd's code was not intended to be error free and were mainly for reference, some functions in this package have inherited those errors. An attempt will be made in future releases to identify and correct them. Dowd's original code can be downloaded from www.kevindowd.org/measuring-market-risk/. It should be noted that Dowd offers both 'MMR2' and 'MMR1' toolboxes. Only 'MMR2' was ported to R. 'MMR2' is more recent version of 'MMR1' toolbox and they both have mostly similar function. The toolbox mainly contains different parametric and non parametric methods for measurement of market risk as well as backtesting risk measurement methods".
- PortfolioAnalytics due to Brian G. Peterson, Peter Carl, Kris Boudt, Ross Bennett, Hezky Varon, Guy Yollin and R. Douglas Martin. According to the authors, this package "gives portfolio optimization and analysis routines and graphics".

- GAS due to Leopoldo Catania, Kris Boudt and David Ardia. According to the authors, this package "implements several backtesting procedures for the Value at Risk (VaR). These are: (i) The statistical tests of Kupiec (1995), Christoffesen (1998) and Engle and Manganelli (2004), (ii) The tick loss function detailed in Gonzalez-Rivera et al. (2004), the mean and max absolute loss used by McAleer and Da Veiga (2008) and the actual over expected exceedance ratio".

- fBasics due to the Rmetrics Core Team, Diethelm Wuertz, Tobias Setz and Yohan Chalabi. According to the authors, this package serves as "the environment for teaching Financial Engineering and Computational Finance".

- MCS due to Leopoldo Catania and Mauro Bernardi. According to the authors, this package "performs the Model Confidence Set procedure of Hansen et al. (2011) for a given set of loss series belonging to several different models that should be compared".

- backtestGraphics due to David Kane, Ziqi Lu, Fan Zhang and Miller Zijie Zhu. According to the authors, this package "creates an interactive graphics interface to visualize backtest results of different financial instruments, such as equities, futures, and credit default swaps. The package does not run backtests on the given data set but displays a graphical explanation of the backtest results. Users can look at backtest graphics for different instruments, investment strategies, and portfolios. Summary statistics of different portfolio holdings are shown in the left panel, and interactive plots of profit and loss (P&L), net market value (NMV) and gross market value (GMV) are displayed in the right panel".

- backtest due to Jeff Enos, David Kane, Kyle Campbell, Daniel Gerlanc, Aaron Schwartz, Daniel Suo, Alexei Colin and Luyi Zhao. According to the authors, this package "provides facilities for exploring portfolio-based conjectures about financial instruments (stocks, bonds, swaps, options, et cetera)".

- VaR due to Talgat Daniyarov. According to the author, this package provides "a set of methods for calculation of Value at Risk (VaR)".

- PerformanceAnalytics due to Brian G. Peterson, Peter Carl, Kris Boudt, Ross Bennett, Joshua Ulrich, Eric Zivot, Matthieu Lestel, Kyle Balkissoon and Diethelm Wuertz. According to the authors, this package provides "a collection of econometric functions for performance and risk analysis. The package aims to aid practitioners and researchers in utilizing the latest research in analysis of non-normal return streams. In general, it is most tested on return (rather than price) data on a regular scale, but most functions will work with irregular return data as well, and increasing numbers of functions will work with P&L or price data where possible".

R is a free software and is downloadable from http://www.r-project.org

Some commercial software available for backtesting methods for VaR are:

- AmiBroker (www.amibroker.com) offering a "robust backtesting service at a relatively low price. For this reason, it's a popular choice with people who are getting started in day trading. It also allows users to make sophisticated technical charts that they can use to monitor the markets. One drawback is that you may have to pay extra for the market price quote data, depending on what securities and time periods you want to test".

- Cybertrader (www.cybertrader.com) is "Charles Schwab's product for active traders. Its Strategy Tester feature lets you test your trading idea. Then you can set it into a Strategy Ticker, which follows your strategy while the market is open, enabling you to see how your strategy performs in real time. This isn't quite the same as paper trading because it isn't testing how well you would pull the trigger".

- Investor/Rt (www.linnsoft.com) "developed by a company called Linn Software, Investor/RT allows you to develop your own tests and create your own programs. It has packages for Macintosh OS X, which makes it popular with traders who prefer Apple computers. Its users tend to be sophisticated about their trading systems and backtesting requirements; this software isn't really for beginners".

- Metastock (www.equis.com) "is designed for traders who work in stocks, although a MetaStock package is available especially for currency traders, and the regular packages include capabilities for futures and commodities traders".

- Optionvue (www.optionvue.com) "offers a range of analytical tools on the options markets. The software's BackTrader module, an add-on feature, helps you learn more about options markets, test new strategies, and examine relationships between options and the underlying stocks – really useful information for people working in equity markets".

- Tradecision (www.tradecision.com) is "a little pricier than most retail trading alternatives, but it offers more advanced capabilities, including an analysis of the strengths and weaknesses of different trading rules. It can incorporate advanced money management techniques and artificial intelligence to develop more predictions about performance in different market conditions. The system may be overkill for most new day traders, but it can come in handy for some".

- Trading Blox (www.tradingblox.com) was "developed by professional traders who needed to test their own theories and who didn't want to do a lot of programming to do it. It comes in three versions (and price levels), ranging from basic to sophisticated, and the company boasts that it works with some commercial trading firms. Of course, some of its capabilities may be more than you need when you're starting out".

- TradeStation (www.tradestation.com) is "an online broker that specializes in services for day traders. Its strategy testing service lets you specify different trading parameters, and then it shows you where these trades would have taken place in the past, using price charts. It also generates a report of the strategy, showing dollar, percentage, and win-loss performance over different time periods. It doesn't have a trade simulation feature".

Other commercial software include Deltix (www . deltixlab . com), S&P capital IQ (www . quanthouse . com), Smartquant (www . smartquant . com), Marketcetera (www . marketcetera . org), Algotrader (www . algotrader . ch), Seer trading systems (www . seertrading . com), Wealth lab (www . wealth-lab . com), Axioma (www . axioma . com), Ninja trader (www . ninjatrader . com), Right edge (www . rightedgesystems . com), Quantshare (www . quantshare . com), eSignal (www . esignal . com), Metatrader (www . metatrader4 . com), Mulitcharts (www . multicharts . com), Portfolio123 (www . portfolio123 . com), Bloodhound system (www . bloodhoundsystem . com), Quant connect

(www . quantconnect . com), Quantopian (www . quantopian . com), Alphaarchitect (www . alphaarchitect . com), Quantiacs (www . quantiacs . com), Quantpicker (www . quantpicker . com), AnalyzerX (www . analyerxl . com), ETFreplay (www . etfreplay . com) and Jamie Gritton's MI backtester (www . backtest . org).

4 Data illustration

Here, we illustrate four of the tests using a real data set. The data set is on adjusted daily closing prices for Coca Cola from the 1st of January 2001 to the 31st of December 2010. The data was obtained from the New York Stock Exchange. The prices are in United States dollars.

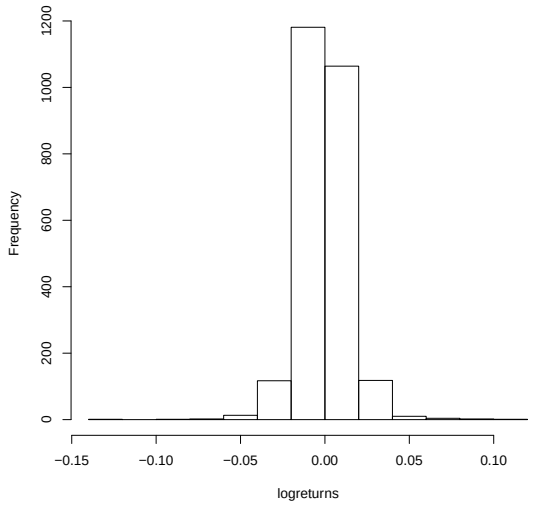

Figure 1: The histogram of the log returns.

A histogram of the log returns of the prices is shown in Figure 1. The histogram appears symmetric. We fitted the three parameter Student's t distribution with degree of freedom parameter ν, location parameter μ and scale parameter σ. We obtained the following estimates: $\widehat{\nu} = 3.122(0.225)$, $\widehat{\mu} = 0.0002(-0.0002)$ and $\widehat{\sigma} = 0.0002(0.0085)$, where the numbers in brackets are the standard errors. We also obtained: Cramer-von Misses statistic = 0.073, Anderson Darling statistic = 0.497, Kolmogorov-Smirnov test statistic = 0.013, its p-value = 0.761, value of Akaike Information Criterion = -15117.9, value of Consistent Akaike Information Criterion = -15117.9, value of Bayesian Information Criterion = -15100.41, and value of Hannan-Quinn information criterion = -15111.56. Based on the p-value, we can say that the Student's t distribution is an adequate model for the data.

We now perform the backtesting. We use the tests given by (1), (2), and the duration based tests of (4) and (5), which we shall refer to as test 1, test 2, test 3 and test 4, respectively. The p-values of these tests versus α are shown in Figures 2 to 5.

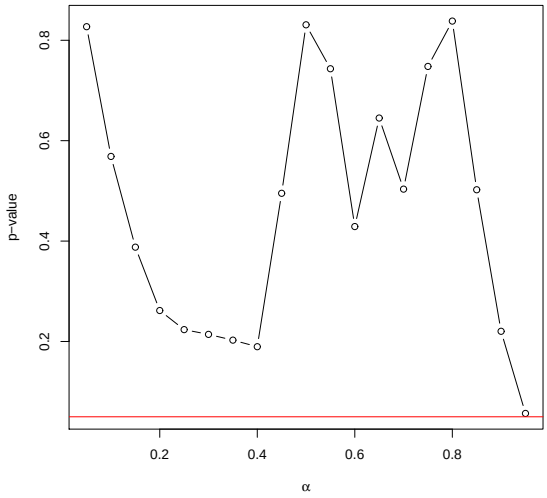

Figure 2: The p-value of test 1 versus α.

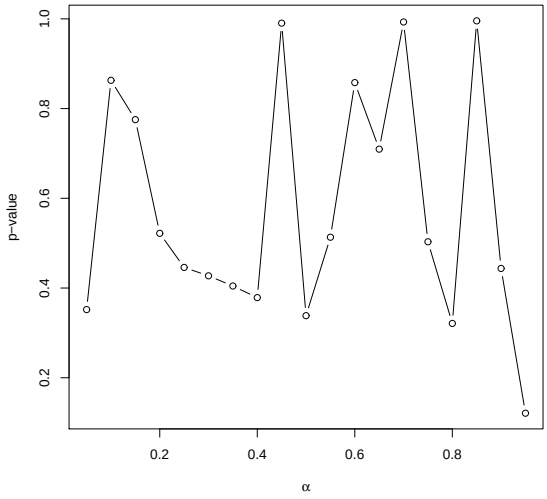

Figure 3: The p-value of test 2 versus α.

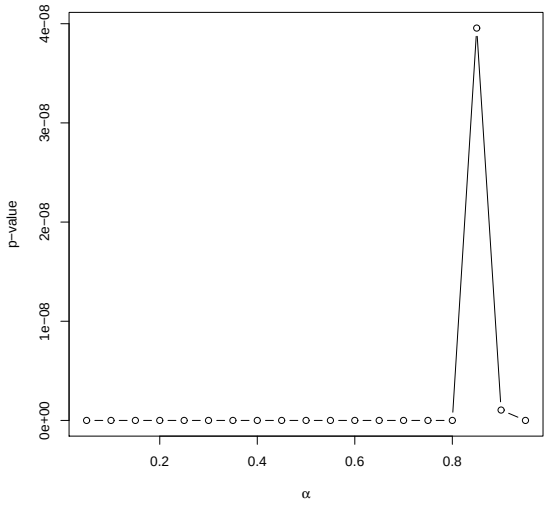

Figure 4: The p-value of test 3 versus α.

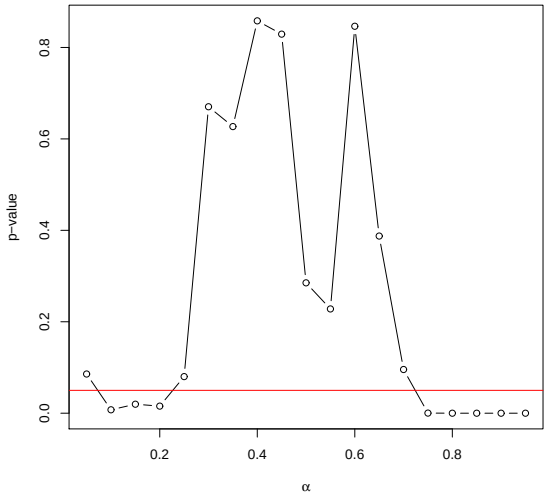

Figure 5: The p-value of test 4 versus α.

According to tests 1 and 2, the fitted model is accepted at all values of α. According to test 4, the fitted model is accepted for most values of α. However, test 3 does not accept the fitted model for any value of α. This may be due to the fitted model not capturing the temporal features well. It is hard to distinguish between tests 1, 2 and 4. But test 2 gives the largest p-values. Test 1 gives the second largest p-values and test 3 gives the smallest p-values.

Acknowledgments

The authors would like to thank the Editor and the referee for careful reading and comments which improved the paper.

References

[1] Basel Committee of Banking Supervision (1996). Supervisory Framework for the Use of "Backtesting" in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. Available at www.bis.org.

[2] Basel Committee of Banking Supervision (2006). International Convergence of Capital Measurement and Capital Standards - A Revised Framework, Comprehensive Version. Available at www.bis.org.

[3] Basel Committee on Banking Supervision (2011). Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems. URL: www.bis.org/publ/bcbs189.pdf

[4] Bera, A.K., Ivliev, S. and Lillo, F. (2015). Financial Econometrics and Empirical Market Microstructure. Springer, Cham.

[5] Berkowitz, J. (2000). Testing density forecasts, with applications to risk management. Graduate School of Management, University of California, Irvine.

[6] Berkowitz, J. (2001). Testing density forecasts with applications to risk management. Journal of Business and Economic Statistics, 19, 465-474.

[7] Berkowitz, J., Christoffersen, P. and Pelletier, D. (2011). Evaluating Value-at-Risk models with desk-level data. Management Science, 57, 2213-2227.

[8] Bontemps, C. (2014). Simple moment-based tests for Value-at-Risk models and discrete distributions. Working paper.

[9] Campbell, S. (2005). A review of backtesting and backtesting procedures. Finance and Economics Discussion Series, Federal Reserve Board, Washington, DC.

[10] Christoffersen, P. (1998). Evaluating interval forecasts. International Economic Review, 39, 841-862.

[11] Christoffersen, P. and Pelletier, D. (2002). Backtesting portfolio risk measures. Working paper.

[12] Christoffersen, P. and Pelletier, D. (2004). Backtesting Value-at-Risk: A duration-based approach. Journal of Financial Econometrics, 2, 84-108.

[13] Colletaz, G., Hurlin, C. and Perignon, C. (2013). The risk map: A new tool for validating risk models. Journal of Banking and Finance, 37, 3843-3854.

[14] Crnkovic, C. and Drachman, J. (1996). Quality control in VaR: Understanding and applying Value-at-Risk. Risk, 9, 139-143

[15] de la Pena, V.H., Rivera, R. and Ruiz-Mata, J. (2006). Quality control of risk measures: Backtesting VAR models. Journal of Risk, 9, 39-54.

[16] da Silva, A.C.R., da Silveira Barbedo, C.H., Araujo, G.S. and das Neves, M.B.E. (2006). Internal models validation in Brazil: Analysis of VaR backtesting methodologies. Revista Brasileira de Financas, 4

[17] Daniela, Z., Edina, K. and Ioan, C.M. (2014). Backtesting value at risk models in the presence of structural break on the Romanian and Hungarian stock markets. 802-810.

[18] Dowd, K. (2002). A bootstrap back-test. Risk, 93-94

[19] Dumitrescu, E.-I., Hurlin, C. and Pham, V. (2012). Backtesting Value-at-Risk: From dynamic quantile to dynamic binary tests. Finance, 33, 79-112.

[20] Efron, B. and Tibshirani, R.J. (1994). An Introduction to the Bootstrap. Chapman and Hall, New York.

[21] Engle, R.F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics, 22, 367-381.

[22] Escanciano, J.C. and Olmo, J. (2010). Backtesting parametric Value-at-Risk with estimation risk. Journal of Business and Economic Statistics, 28, 36-51.

[23] Escanciano, J.C. and Olmo, J. (2011). Robust backtesting tests for Value-at-risk models. Journal of Financial Econometrics, 9, 132-161.

[24] Fitschen, K. (2013). Building Reliable Trading Systems: Tradable Strategies That Perform As They Backtest and Meet Your Risk-Reward Goals. John Wiley and Sons, New York.

[25] Franke, J., Härdle, W. and Stahl, G. (2012). Measuring Risk in Complex Stochastic Systems. Springer Verlag, New York.

[26] Gorgulho, A.M.S.B.S., Neves, R.F.M.F. and Horta, N.C.G. (2013). Intelligent Financial Portfolio Composition Based on Evolutionary Computation Strategies. Springer, Heidelberg.

[27] Haas, M. (2001). New methods in backtesting. Working Paper, Financial Engineering Research Center. URL: www.ime.usp.br/∼rvicente/risco/haas.pdf

[28] Haas, M. (2006). Improved duration-based backtesting of value-at-risk. Journal of Risk, 8, 17-38.

[29] Hurlin, C., Colletaz, G., Tokpavi, S. and Candelon, B. (2010). Backtesting Value-at-Risk: A GMM duration based test. Journal of Financial Econometrics, 2010, 1-30.

[30] Hurlin, C. and Tokpavi, S. (2006). Backtesting Value-at-Risk accuracy: A simple new test. Journal of Risk, 9, 19-37.

[31] Iorgulescu, F. (2012). Backtesting Value-at-Risk: Case study on the Romanian capital market. Procedia - Social and Behavioral Sciences, 62, 796-800.

[32] Jorion, P. (2001). Value at Risk, second edition. McGraw-Hill, New York.

[33] Krämer, W. and Wied, D. (2015). A simple and focused backtest of value at risk1 Working paper, Technische Universität Dortmund, Germany.

[34] Kresta, A. (2013). Backtesting the filtered historical simulation for the VaR estimation of the Prague stock market index. Journal of Economics, Management and Business, 23, 15-26.

[35] Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives, 2, 73-84.

[36] Li, Y.X., Lian, J.G. and hang, H.K. (2016). Forecast and backtesting of VAR models in crude oil market. Research and Reviews: Journal of Statistics and Mathematical Sciences, 2, 131-140.

[37] Lichters, R., Stamm, R. and Gallagher, D. (2015). Modern Derivatives Pricing and Credit Exposure Analysis: Theory and Practice of CSA and XVA Pricing, Exposure Simulation and Backtesting. Springer Verlag, New York.

[38] Lopez, J.A. (1998). Testing your risk tests. Financial Survey, 18-20.

[39] Lopez, J.A. (1999). Methods for evaluating value-at-risk estimates. Economic Review: Federal Reserve Bank of San Francisco, 2, 3-17.

[40] Matras, K. (2011). Finding #1 Stocks: Screening, Backtesting and Time-Proven Strategies. John Wiley and Sons, New York.

[41] McNeil, A.J., Frey, R. and Embrechts, P. (2015). Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton, New Jersey.

[42] Nadarajah, S. and Chan, S. (2016). Estimation methods for value at risk. Chapter 12 of Extreme Events in Finance: A Handbook of Extreme Value Theory and Its Applications (edited by F. Longin), pp. 303-373. John Wiley and Sons, Chichester.

[43] Pajhede, T. (2015). Backtesting Value-at-Risk: A generalized Markov framework. Working paper, Lancaster University, UK.

[44] Patra, B. and Padhi, P. (2015). Backtesting of Value at Risk methodology: Analysis of banking shares in India. Journal of Applied Economic Research, 9, 254-277.

[45] Pelletier, D. and Wei, W. (2015). The geometric-VaR backtesting method. Journal of Financial Econometrics, 2015, 1-21.

[46] Perignon, C. and Smith, D. (2008). A new approach to comparing VaR estimation methods. Journal of Derivatives, 16, 54-66.

[47] R Development Core Team (2016). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

[48] Schulmerich, M. (2010). Real Options Valuation: The Importance of Interest Rate Modelling in Theory and Practice. Springer, Heidelberg.

[49] Tichy, T. (2012). Some findings about risk estimation and backtesting at the world FX rate market. In: Proceedings of 30th International Conference Mathematical Methods in Economics, pp. 897-902.

[50] Virdi, N. K. (2011). A review of backtesting methods for evaluating Value-at-Risk. International Review of Business Research Papers, 7, 14-24.