Backtesting Value-at-Risk: A Generalized Markov Framework*

Thor Pajhede†

October 20, 2015

Abstract

Testing the validity of Value-at-Risk (VaR) forecasts, or backtesting, is an integral part of modern market risk management and regulation. This is often done by applying independence and coverage tests developed in Christoffersen (1998) to so-called hit-sequences derived from VaR forecasts and realized losses. However, as pointed out in the literature, see Christoffersen (2004), these aforementioned tests suffer from low rejection frequencies, or (empirical) power, when applied to hit-sequences derived from simulations matching empirical stylized characteristics of return data. One key observation of the studies is that non-Markovian behavior in the hit-sequences may cause the observed lower power performance. To allow for non-Markovian behavior, we propose to generalize the backtest framework for Value-at-Risk forecasts, by extending the original first order dependence of Christoffersen (1998) to allow for a higher, or k'th, order dependence. We provide closed form expressions for the tests as well as asymptotic theory. Not only do the generalized tests have power against k'th order dependence by definition, but also included simulations indicate improved power performance when replicating the aforementioned studies.

Keywords: Value-at-Risk, Backtesting, Risk Management, Markov Chain, Duration-based test, quantile, likelihood ratio, maximum likelihood.

JEL codes: C12, C15, C52, C32.

*The author acknowledge support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation. We thank Peter Christoffersen, Anders Rahbek and Heino Bohn Nielsen for helpful comments. The Matlab code used to implement the Backtests in this paper as well as several others is free to use, and will be made available from econ.ku.dk/pajhede/backtest.

†Department of Economics, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark. CREATES, University of Aarhus. E-mail: thor.nielsen@econ.ku.dk

1 Introduction

Since its introduction in the 90s Value-at-Risk (VaR), as measured by the p'th quantile of a forecasted distribution of losses, has become widely used when reporting aggregate market risk. This again has prompted a rich literature on validation of VaR forecasts, so-called backtesting, as much applied empirically by regulatory authorities, academics and financial institutions. See Campbell (2007) for a review of the backtesting procedures and an economic motivation for the backtesting criteria.

The leading reference on backtesting is Christoffersen (1998), wherein the evaluation of accurate VaR forecasts was first formalized. Specifically it was shown that the occurrences of losses beyond a specified VaR level, termed *violations* or *hits*, should occur independently and with a constant probability matching the p'th quantile. Based on this, the widely applied *conditional coverage* and *independence* tests were proposed. However, as documented in Christoffersen (2004) and Berkowitz et al. (2011) the tests have low empirical power in simulation studies matching empirical stylized facts of returns data.

To address this we propose to derive tests in a more general setting than the original framework of Christoffersen (1998). Specifically, we propose tests within a general backtest framework extending the underlying Markovian model of Christoffersen (1998) to allow for higher, or k'th, order dependence. Within the quite general k'th order dependence model, we consider two structures, or specifications: one which we label as the generalized Markov specification, and the other as the generalized Markov duration specification. Preceding the details given in Section 2.2, the generalized Markov specification can be viewed as similar to the extension of autoregressive models from order one to order k when testing for white noise, while the Markov duration specification mimic duration modeling approaches to backtesting of Christoffersen (2004), Haas (2006) and Wei and Pelletier (2014).

We provide asymptotic theory and closed form expressions for the implied tests for conditional coverage and independence within these generalized specifications. Moreover, simulations illustrate that the new generalized tests solve some of the leading issues with regards to low empirical power.

Note in this respect, that by definition the proposed tests will have power against higher order dependence, and in particular so when compared to the tests derived in the Markovian framework. That the tests seem to perform well in empirically stylized simulations is additional reason to prefer these.

The rest of the paper is organized as follows. Section 2 sets out the backtesting criteria i.e. Unconditional Coverage, Independence, and Conditional Coverage. Subsection 2.1 reviews the popular classic Markov backtests due to Christoffersen (1998) and Kupiec (1995). Subsection 2.2 introduces our new framework. We consider two specifications from this framework, the generalized Markov and the Markov duration specifications. From these we derive tests of unconditional coverage, independence and conditional coverage. Section 3 examines the power and size properties of the various tests using a simulation framework. Section 4 concludes.

2 Hit-sequence Based Backtesting

Let R_t denote the realization of a return of an asset or a portfolio of assets at time t. The *ex ante* VaR for time t and *coverage rate* p, denoted as $\mathrm{VaR}_{t|t-1}(p)$, conditional on all information, $\mathcal{F}_{t-1}$, available at time $t-1$ (for example past returns and macroeconomic indicators) is defined as the p'th conditional quantile of the distribution of R_t:

$$P(R_t < \mathrm{VaR}_{t|t-1}(p)|\mathcal{F}_{t-1}) = p, \qquad t = 1, ..., T.$$

Typically the coverage rate used is 1% or 5%. Several parametric (for example GARCH models) and non-parametric (for example *Historical Simulation*) methods are used to forecast $\mathrm{VaR}_{t|t-1}(p)$, see McNeil et al. (2005).

Backtesting is the procedure of comparing *realized* losses to the *forecasted* VaR. To implement backtesting of a VaR forecast, we follow Christoffersen (1998) in defining the *hit-sequence*, $\{I_t\}_{t=1}^{T}$, as follows:

Definition 1. The hit-sequence, $\{I_t\}_{t=1}^{T}$, for a sequence of VaR forecast, $\left\{\mathrm{VaR}_{t|t-1}(p)\right\}_{t=1}^{T}$, is defined as,

$$I_t := 1\left(R_t < \mathrm{VaR}_{t|t-1}(p)\right), \qquad t = 1, ...T \tag{2.1}$$

Where $1(\cdot)$ is the indicator function. Thus, the hit-sequence is by construction a binary time series indicating whether a loss at time t greater than the VaR, termed a *violation* or a *hit*, was realized.

A VaR forecast is valid, in the sense of actually having forecasted the desired quantile, only if the associated hit-sequence satisfies the following criteria due to Christoffersen (1998):

- **The unconditional coverage criteria** The unconditional probability of a violation must be exactly equal to the coverage rate p:

$$H_{UC}: \ P(I_t = 1) = p$$

- **The independence criteria**: The conditional probability of a violation must be constant:

$$H_{Ind}: \ P(I_t = 1|\mathcal{F}_{t-1}) = P(I_t = 1)$$

Combining these criteria we obtain the conditional coverage criteria:

- **The conditional coverage criteria**: The probability of a violation must be constant and equal to the coverage rate:

$$H_{CC}: \ P(I_t = 1|\mathcal{F}_{t-1}) = P(I_t = 1) = p$$

It follows, see Christoffersen (1998), that the hit-sequence of a valid VaR forecast, is in fact a sequence of i.i.d. Bernoulli distributed variables:

$$I_t \underset{i.i.d.}{\sim} Bernoulli(p), \qquad t = 1, ..., T. \tag{2.2}$$

The classic Markov framework of Christoffersen (1998) models the hit-sequence of (2.2) as a first order Markov chain. As detailed in the following subsection 2.1 this allows testing of both the unconditional coverage and independence criteria using likelihood-ratio tests. Furthermore, these tests have closed form expressions, standard asymptotics and are easy to implement. However, as previously mentioned in the introduction, the tests have also been found to suffer from low power when dependence is not Markovian.

In subsection 2.2, we extend the classic Markov framework to allow for higher, or k'th order, dependence. We detail how our approach preserves all of the aforementioned advantages of the classic Markov testing, but also have power against more general forms of dependence.

2.1 Classic Markov Testing

The first backtest by Kupiec (1995), models the hit-sequence as an i.i.d. Bernoulli sequence with an unknown probability parameter $\pi_1 \in]0,1[$, that is:

$$I_t \underset{i.i.d.}{\sim} Bernoulli(\pi_1), \qquad t = 1, ..., T \tag{2.3}$$

The likelihood for the Bernoulli sequence (2.3) is given by $\mathcal{L}_T(\pi_1) = \pi_1^{T_1}(1-\pi_1)^{T_0}$ where $T_1 = \sum_{t=1}^{T} I_t$, $T_0 = T - T_1$ and the maximum likelihood (ML) estimate of π_1 is given by $\hat{\pi}_1 = T_1/T$.

From this a likelihood-ratio test of the restriction $H_{UC}: \hat{\pi}_1 = p$, corresponding to the criteria of unconditional coverage, can be constructed in the usual way. It follows that the likelihood-ratio statistic, under the hypothesis stated in the parenthesis, for unconditional coverage satisfies, as $T \to \infty$,

$$Q_{UC}(\pi_1 = p) = -2log\left(\frac{p^{T_1}(1-p)^{T_0}}{\hat{\pi}_1^{T_1}(1-\hat{\pi}_1)^{T_0}}\right) \xrightarrow{d} \chi^2(1). \tag{2.4}$$

This test is often termed the proportion of failures (PF) test. Because the model from which the test was derived, see equation (2.3), does not allow for any dependence structure in the hit-sequence it is clear that the test is unsuited to detect dependence in the hit-sequence.

The need to also test the independence criteria led Christoffersen (1998) to develop the Markov tests of independence and conditional coverage. To do so it was proposed to model the conditional distribution of I_t given I_{t-1}, $I_t|I_{t-1}$ as a first order Markov chain. We write this first order Markov chain as

$$I_t|I_{t-1} \underset{i.i.d}{\sim} Bernoulli(p_t(\theta)),$$

with transition probability,

$$p_t(\theta) = I_{t-1}\pi_{11} + (1 - I_{t-1})\pi_{01}, \quad \theta = (\pi_{11}, \pi_{01})' \in]0, 1[^2.$$

Here π_{ij} is the probability of observing i on day $t-1$ being followed by observing j on day t for $i, j = 0, 1$. Equivalently this may be expressed in terms of the transition probability matrix given by

$$\Pi = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}. \tag{2.5}$$

In terms of Π, independence is implied by the restriction $H_{Ind}: \ \pi_{01} = \pi_{11}$ while the combined hypothesis of conditional coverage can be tested by the additional restriction $H_{CC}: \ \pi_{01} = \pi_{11} = p$.

The likelihood for the unrestricted Markov chain $\{I_t\}_{t=1}^{T}$, with the first observation (I_0) fixed, is given by

$$\mathcal{L}_T(\pi_{01}, \pi_{11}) = (1 - \pi_{01})^{T_{00}} \pi_{01}^{T_{01}} (1 - \pi_{11})^{T_{10}} \pi_{11}^{T_{11}}.$$

Here T_{ij} indicates the number of observations of the hit-sequence where a j follows an i. Noting that $T_1 = T_{11} + T_{10}$ and $T_1 = T_{11} + T_{10}$, the ML estimates are $\hat{\pi}_{01} = T_{01}/T_0$, $\hat{\pi}_{11} = T_{11}/T_1$. It follows that the likelihood-ratio test statistic of independence, as $T \to \infty$, satisfies,

$$Q_{Ind}(\pi_{01} = \pi_{11}) = -2log\left(\frac{(1 - \hat{\pi}_1)^{T_{00}} \hat{\pi}_1^{T_{01}} (1 - \hat{\pi}_1)^{T_{10}} \hat{\pi}_1^{T_{11}}}{(1 - \hat{\pi}_{01})^{T_{00}} \hat{\pi}_{01}^{T_{01}} (1 - \hat{\pi}_{11})^{T_{10}} \hat{\pi}_{11}^{T_{11}}}\right) \xrightarrow{d} \chi^2(1).$$

Likewise, the likelihood-ratio test statistic for conditional coverage (the so-called joint test), satisfies,

$$Q_{CC}(\pi_{01} = \pi_{11} = p) = -2log\left(\frac{(1 - p)^{T_{00}} p^{T_{01}} (1 - p)^{T_{10}} p^{T_{11}}}{(1 - \hat{\pi}_{01})^{T_{00}} \hat{\pi}_{01}^{T_{01}} (1 - \hat{\pi}_{11})^{T_{10}} \hat{\pi}_{11}^{T_{11}}}\right) \xrightarrow{d} \chi^2(2).$$

The tests of Christoffersen (1998) have standard asymptotics, closed form expressions and remain popular in the applied literature. However, because they only model the hit-sequence as a first order Markov chain the ability to detect higher order dependence may be limited. Furthermore, simulation studies have shown them to have a low power in realistic settings. In the following subsection 2.2 we extend the model to allow for higher order dependence, in order to remedy the shortcomings of the classical framework but still derive tests that are easy to implement and interpret.

2.2 Generalized Markov framework

As detailed in the first part of the section, we now extend the classic Markov framework to a k'th order Markov chain. Specifically, let the Markov chain be given by,

$$I_t|\mathcal{F}_{t-1,k} \underset{i.i.d}{\sim} Bernoulli(p_t(\theta)), \quad \mathcal{F}_{t-1,k} = I_{t-1}, ..., I_{t-k}, \qquad t = 1, ..., T. \tag{2.6}$$

The transition probabilities of (2.6) are given by,

$$p_t(\theta) = P(I_t = 1|\mathcal{F}_{t-1,k}), \qquad t = 1, ..., T, \tag{2.7}$$

With θ a 2^k vector of the individual parameters, corresponding to the possible permutations of $I_{t-1}, I_{t-2}, ..., I_{t-k}$.

Equivalently, one could specify a k-tuple $\tilde{I}_t = (I_t, ..., I_{t-k+1})'$ which would then follow a Markov chain governed by a $2^k \times 2^k$ transition matrix P. Since the rows of P must sum to 1 and each state is only accessible from 2 other states, this implies that each row has two non-zero elements[1], which restricts it to the 2^k parameters also found in θ.

The likelihood for this model conditioned on k observations prior to $t = 1$ fixed, is given by,

$$\mathcal{L}_T(\theta) = \prod_{t=1}^{T} p_t(\theta)^{I_t}(1 - p_t(\theta))^{1-I_t},$$

and the log-likelihood by, $L_T(\theta) = \sum_{t=1}^{T} log(p_t(\theta))I_t + log(1 - p_t(\theta))(1 - I_t)$.

The principal motivation was to allow for dependence of order $k > 1$. However since the number of parameters increase at the geometric rate of 2^k, estimating the model quickly becomes infeasible for larger values of k. In order to have a feasible number of parameters we therefore impose parametric structures on the model of equation (2.7). Examples of such structures or restrictions are presented in the following subsections 2.2.1 and 2.2.2. The criteria of independence and conditional coverage impose further restrictions, which are used to create likelihood-ratio tests. Specifically if the restriction $p_t(\theta) = p$ holds for all t, then the Markov chain of equation (2.6) reduces to the i.i.d. Bernoulli sequence of equation (2.2).

There is no clear choice of k. A too low value might not adequately allow for the modeling of higher order dependence. While a too high k conditions on too many observations making the effective sample size small. For $k = 1$ the tests suggested in the following subsections reduce to the tests of Christoffersen (1998) described in section 2. A natural choice of k is to use 5, 10 or 20, corresponding to testing for a change in the probability of a hit in the week, two weeks or 1 month following a hit.

2.2.1 The Generalized Markov Test

In terms of the unrestricted model in (2.6), we first consider the restriction that the probability of a hit at time t, $p_t(\theta)$, is a function of only whether or not a hit has occurred in $I_{t-1}, I_{t-2}, ..., I_{t-k}$. This reduces the parameters

[1] Intuitively, if $k = 1$, one can recall that the two permutations of I_{t-1} (either 1 or 0) meant that the classical tests of Christoffersen (1998) are based on a Markov chain with 2 parameters in θ which are gathered into a 2×2 transition matrix.

of the model to two, or equivalently,

$$p_t(\theta) = J_{t-1}p_E + (1 - J_{t-1})p_S, \quad J_{t-1} := 1\left(\sum_{i=1}^{k} I_{t-1} > 0\right).$$

The bivariate parameter vector $\theta = (p_E,\ p_S)'$ belongs to the parameter space $\Theta =]0,1[^2$. Intuitively, this corresponds to an excited (p_E) and a steady (p_S) probability. Because the restricted model retains the interpretation of two categories similar to the Markov tests of Christoffersen (1998), we will refer to it as the the generalized Markov specification.

The likelihood is then given by,

$$\mathcal{L}_T(\theta) = (1 - p_S)^{T_{00}} p_S^{T_{01}} (1 - p_E)^{T_{10}} p_E^{T_{11}},$$

where T_{ij} are the counts; $T_{11} := \sum_{t=1}^{T} I_t J_{t-1}$, $T_{01} := \sum_{t=1}^{T} I_t(1 - J_{t-1})$, $T_{10} := \sum_{t=1}^{T}(1 - I_t)J_{t-1}$, $T_{00} := \sum_{t=1}^{T}(1 - I_t)(1 - J_{t-1})$. That is, T_{11} (T_{10}) is the number of hits (no hits) observed where one or more hits were observed in the preceding k observations. $T_{01}(T_{00})$ is the number of hits (no hits) observed where there was not observed a hit in the prior k observations.

This leads to the ML estimates (see Appendix A),

$$\hat{p}_S = \frac{T_{01}}{T_{01} + T_{00}} \qquad \text{and} \qquad \hat{p}_E = \frac{T_{11}}{T_{11} + T_{10}}.$$

To test the hypothesis of independence, we consider the restriction $H_{Ind}:\ p_E = p_S := \phi$, that is, whether there is a constant probability of a hit. The restricted parameter space, Θ_H, is in this case given by,

$$\Theta_H = \{\theta |\ \theta = H\phi,\ \phi \in]0,1[\},$$

where $H = (1,1)'$, with ML estimate of ϕ given by (see Appendix A)

$$\hat{\phi} = \frac{T_{01} + T_{11}}{T_{01} + T_{11} + T_{00} + T_{10}} = \frac{T_1}{T}. \tag{2.8}$$

Defining the unrestricted estimator, the estimator restricted under H_{Ind} and the estimator restricted under H_{CC} as

$$\hat{\theta} := \underset{\theta\in\Theta}{argmax}\ \mathcal{L}_T(\theta) = (\hat{p}_S, \hat{p}_E)', \quad \tilde{\theta} := \underset{\theta\in\Theta_H}{argmax}\ \mathcal{L}_T(\theta) = H\hat{\phi} \quad \text{and} \quad \theta_0 = Hp.$$

As was the case for the classic Markov tests of the previous section, the likelihood ratio test statistic of independence conveniently factorizes, see Appendix A, into tests for conditional coverage and unconditional coverage as follows (with the hypothesis of each test in parenthesis)

$$\begin{aligned} Q_{G-Ind}(\theta = H\phi) &= -2log\left(\frac{\mathcal{L}_T(\tilde{\theta})}{\mathcal{L}_T(\hat{\theta})}\right) = Q_{G-CC}(\theta = Hp) - Q_{G-UC}(H\phi = Hp) \\ &= \left(-2\left[L_T(\theta_0) - L_T(\hat{\theta})\right]\right) - \left(-2\left[L_T(\theta_0) - L_T(\tilde{\theta})\right]\right) \end{aligned}$$

Note the simple relation, $Q_{G-CC}(\theta = Hp) = Q_{G-Ind}(\theta = H\phi) + Q_{G-UC}(H\phi = Hp)$. This provides a simple way of analyzing a rejection of CC. If a rejection of conditional coverage is found one can examine if it was due to dependence, an incorrect coverage or both, using the $Q_{G-Ind}(\theta = H\phi)$ and $Q_{G-UC}(H\phi = Hp)$ tests.

The test statistic of independence has the following expression

$$\begin{aligned} Q_{G-Ind}(\theta = H\phi) &= -2log\left(\frac{\mathcal{L}_T(\tilde{\theta})}{\mathcal{L}_T(\hat{\theta})}\right) \\ &= -2\{log(1-\hat{\phi})(T_{00}+T_{10}) + log(\hat{\phi})(T_{01}+T_{11}) \\ &- log(1-\hat{p}_S)T_{00} - log(\hat{p}_S)T_{01} - log(1-\hat{p}_E)T_{10} - log(\hat{p}_E)T_{11}\} \end{aligned} \tag{2.9}$$

The test statistic of conditional coverage has the following expression

$$\begin{aligned} Q_{G-CC}(\theta = Hp) &= -2log\left(\frac{\mathcal{L}_T(\theta_0)}{\mathcal{L}_T(\hat{\theta})}\right) \\ &= -2\{log(1-p)(T_{00}+T_{10}) + log(p)(T_{01}+T_{11}) \\ &- log(1-\hat{p}_S)T_{00} - log(\hat{p}_S)T_{01} - log(1-\hat{p}_E)T_{10} - log(\hat{p}_E)T_{11}\} \end{aligned} \tag{2.10}$$

The test statistic for unconditional coverage, $Q_{G-UC}(H\phi = Hp)$ is by definition simply the proportion of failures test of section 2, where the first k observations are dropped from the sample.

The distribution of the generalized Markov tests of for independence, conditional coverage and unconditional coverage are asymptotically $\chi^2(1)$, $\chi^2(2)$ and $\chi^2(1)$. That is, We have the following results:

Theorem 1. *For $T \to \infty$, and under the null-hypothesis that $\{I_t\}$ is an i.i.d. Bernoulli sequence with probability parameter p,*

$$Q_{G-Ind}(\theta = H\phi) \xrightarrow{d} \chi^2(1),$$

$$Q_{G-CC}(\theta = Hp) \xrightarrow{d} \chi^2(2),$$

$$Q_{G-UC}(H\phi = Hp) \xrightarrow{d} \chi^2(1).$$

For a proof see Appendix A.

2.2.2 The Markov Duration Test

In terms of the unrestricted model in 2.6, we now consider the restriction that the probability of a hit at time t, $p_t(\theta)$, is a function of the number of observations since the last hit (the *duration*) in the preceding k lags, after which the probability is a constant. This reduces the parameters of the model to $k+1$, or equivalently,

$$p_t(\theta) = J(1)_{t-1}p_{E1} + ... + J(k)_{t-1}p_{Ek} + (1 - \sum_{i=1}^{k} J(i)_{t-1})p_S,$$

where

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

Specifically this implies $p_{E1} = P(I_t = 1|I_{t-1} = 1)$, $p_{Ek} = P(I_t = 1|I_{t-1} = 0, ..., I_{t-k} = 1)$ and $p_S = P(I_t = 1|I_{t-1} = 0, ..., I_{t-k} = 0)$. Because the restricted model is similar to the underlying models of the duration based backtests of Christoffersen (2004), Haas (2006) and Wei and Pelletier (2014) we will refer to this as the Markov duration specification.

The parameter vector $\theta = (p_{E1}, ..., p_{Ek},\ p_S)$ belongs to the parameter space $\Theta =]0,1[^{k+1}$. The Markov duration specification is less restrictive than that of the generalized Markov specification and contains it as the special case $p_{E1} = ... = p_{Ek}$. Despite being less restrictive, the specification ensures that the number of parameters in (2.6) only grows linearly with k.

The likelihood is given by,

$$\mathcal{L}_T(\theta) = (1 - p_S)^{T_{00}} p_S^{T_{01}} \prod_{i=1}^{k} (1 - p_{Ei})^{T_{10}(i)} p_{Ei}^{T_{11}(i)},$$

where $T_{10}(i) = \sum_{t=1}^{T}(1 - I_t)J(i)_{t-1}$ is the number of zeros observed after having observed a hit in I_{t-i}, but not in any I_{t-j} where $i > j$. $T_{11}(i)$ is the number of ones observed after having observed a hit I_{t-i} lags previously, but not in any I_{t-j} where $i > j$.

This leads to the ML estimates (see Appendix B),

$$\hat{p}_S = \frac{T_{01}}{T_{01} + T_{00}} \qquad \text{and} \qquad \hat{p}_{Ei} = \frac{T_{11}(i)}{T_{11}(i) + T_{10}(i)}, \quad i = 1, ..., k.$$

To test the hypothesis of independence, consider the restriction $H_{Ind}:\ p_{E1} = ... = p_{Ek} = p_S := \phi$, that is, whether there is a constant probability of a hit. The restricted parameter space is given by

$$\Theta_H = \{\theta |\ \theta = H\phi,\ \phi \in]0,1[\},$$

Where $H = (1, ..., 1)'$ is a $k \times 1$ vector and with ML estimate $\hat{\phi}$ unchanged.

Defining the unrestricted estimator, the estimator restricted under H_{Ind} and the estimator restricted under H_{CC} as

$$\hat{\theta} := \underset{\theta \in \Theta}{argmax}\ \mathcal{L}_T(\theta) = (\hat{p}_S, \hat{p}_{E1}, ..., \hat{p}_{Ek})', \quad \tilde{\theta} := \underset{\theta \in \Theta_H}{argmax}\ \mathcal{L}_T(\theta) = H\hat{\phi} \quad \text{and} \quad \theta_0 = Hp.$$

The likelihood ratio test statistic factorizes as into tests of conditional coverage and unconditional coverage as in the previous subsection (with the hypothesis of each test in parenthesis)

$$\begin{aligned} Q_{D-Ind}(\theta = H\phi) &= -2log\left(\frac{\mathcal{L}_T(\tilde{\theta})}{\mathcal{L}_T(\hat{\theta})}\right) = Q_{D-CC}(\theta = Hp) - Q_{D-UC}(H\phi = Hp) \\ &= \left(-2\left[L_T(\theta_0) - L_T(\hat{\theta})\right]\right) - \left(-2\left[L_T(\theta_0) - L_T(\tilde{\theta})\right]\right), \end{aligned}$$

We will refer to these tests as the Markov-Duration tests of independence, unconditional coverage and conditional coverage. We again have the relation between the tests that $Q_{D-CC}(\theta = Hp) = Q_{D-Ind}(\theta = H\phi) + Q_{D-UC}(H\phi = Hp)$.

Intuitively, $Q_{D-Ind}(\theta = H\phi)$ tests whether the hazard function can be reduced to a constant and $Q_{D-CC}(\theta = Hp)$ tests if that constant is exactly p. They can be viewed as duration tests, with the hazard rate being entirely free of restrictions except a truncation to a constant beyond the k'th lag.

The test statistic of independence has the following expression

$$\begin{aligned} Q_{D-Ind}(\theta = H\phi) &= -2log\left(\frac{\mathcal{L}_T(\tilde{\theta})}{\mathcal{L}_T(\hat{\theta})}\right) \qquad (2.11) \\ &= -2\Big(log(1-\hat{\phi})(T_{00}+T_{10}) \times log(\hat{\phi})(T_{01}+T_{11}) - log(1-\hat{p}_S)T_{00} - log(\hat{p}_S)T_{01} \\ &- \sum_{i=1}^{k} log(1-\hat{p}_{Ei})T_{10}(i) - \sum_{i=1}^{k} log(\hat{p}_{Ei})T_{11}(i)\Big). \end{aligned}$$

The test statistic of conditional coverage has the following expression

$$\begin{aligned} Q_{D-CC}(\theta = Hp) &= -2log\left(\frac{\mathcal{L}_T(\theta_0)}{\mathcal{L}_T(\hat{\theta})}\right) \\ &= -2(log(1-p)(T_{00}+T_{10}) \times log(p)(T_{01}+T_{11}) - log(1-\hat{p}_S)T_{00} - log(\hat{p}_S)T_{01} \\ &- \sum_{i=1}^{k} log(1-\hat{p}_{Ei})T_{10}(i) - \sum_{i=1}^{k} log(\hat{p}_{Ei})T_{01}(i)). \qquad (2.12) \end{aligned}$$

Lastly, $Q_{D-UC}(H\phi = Hp)$ is simply the proportion of failures test of section 2, where the first k observations are dropped from the calculations (equivalent to the $Q_{G-UC}(H\phi = Hp)$ test statistic).

The distribution of the Markov duration tests of for independence, conditional coverage and unconditional coverage are asymptotically χ^2 distributed. That is, we have the following results:

Theorem 2. *For $T \to \infty$, and under the null-hypothesis that $\{I_t\}$ is an i.i.d. Bernoulli sequence with probability parameter p,*

$$Q_{D-Ind}(\theta = H\phi) \xrightarrow{d} \chi^2(k-1),$$

$$Q_{D-CC}(\theta = Hp) \xrightarrow{d} \chi^2(k),$$

$$Q_{D-UC}(H\phi = Hp) \xrightarrow{d} \chi^2(1).$$

For a proof see Appendix B.

In section 3.1 we demonstrate in a simulation study that using the asymptotic distributions of Theorems 1 and 2 to calculate p-values can cause a distortion of the size. Instead the Monte Carlo Method of Dufour (2006) can be used to simulate the exact distribution under the null hypothesis and obtain valid p-values. It is the tests using the Monte Carlo Method of Dufour (2006) which should be used in practice and it is what is used in our empirical power simulations found in sections 3.2 and 3.3.

3 Simulation Study of Size and Power

In this section we conduct a simulation study to investigate the empirical size and power properties of the generalized Markov and duration tests of conditional coverage developed in section 2.2. Further, we evaluate the *empirical rejection frequency* (ERF) of the tests using a simulation setup not contained in the general model of equation (2.6), generating the returns using a GARCH model and forecasting the VaR using *historical simulation* (HS). This later simulation is commonly included in papers which develop VaR backtests and we refer to it as scenario power.

We use $k = 1$, 5, 10 and 20 lags for each of the conditional coverage tests, see equations (2.10) and (2.12) from section 2.2, where we note that for $k = 1$, the generalized Markov and generalized duration tests both reduce to the original joint test of Christoffersen (1998). We use sample sizes $T = 500$, $1,000$, $1,500$, $2,500$, $5,000$ and $N = 100,000$ replications for each sample size. For the size simulations we use $p = 1\%$, 5% and 10%, where the latter is included to illustrate the improved size properties for larger values of p. For the power simulations we use only $p = 1\%$ and 5% reflecting empirically relevant cases. We use a significance level of 5% for all simulations. In the empirical power and scenario power simulations in subsections 3.2 and 3.3 we use the Monte Carlo testing technique of Dufour (2006) (see Appendix C) to obtain tests with a size of 5%.

3.1 Empirical Size

It is a well established fact of the backtest literature that the use of asymptotic distributions critical values can create significant size distortions in existing tests, see Christoffersen (2004). To examine the size distortion of the tests developed in this paper, and to examine when the asymptotic critical values can be used, we simulate the hit-sequence, $\{I_t\}_{t=1}^{T}$, under the null hypothesis of Conditional Coverage as an independent Bernoulli sequence. Recalling equation (2.2), we simulate the hit-sequence, $\{I_t\}_{t=1}^{T}$, using the *data generating process* (DGP):

$$I_t \underset{i.i.d}{\sim} Bernoulli(p), \qquad t = 1, ..., T$$

ERFs of the generalized Markov and generalized duration tests of conditional coverage, when using the asymptotic distributions critical value are presented in Table 1. From the table it is clear that using the critical values of the asymptotic distributions can cause size distortion, especially when testing a low p or when using a small sample. In general most tests appear to be undersized when using the low $p = 1\%$. Though the generalized Markov test, of equation (2.10), is only slightly undersized for $k > 5$ and $T > 1,000$. When the higher $p = 5\%$ or 10%, is used, the size properties are generally much improved for the generalized Markov tests. Especially so for $k > 5$. The generalized duration test, of equation (2.12), has somewhat varying size properties. For the low $p = 1\%$ it is undersized while for $p = 5\%$ or 10% it is oversized, though not to high degree when $T = 5,000$ observations are used.

Because of the size distortion the empirical power and scenario power simulations in subsections 3.2 and 3.3

use the Monte Carlo testing technique of Dufour (2006) (see Appendix C) to obtain tests with a size of 5%.

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	1.08	2.31	2.69	2.46	0.51	0.13	0.02
1000	2.63	3.05	3.77	4.19	0.73	0.26	0.05
1500	3.08	3.72	4.22	5.13	0.94	0.39	0.09
2500	2.75	4.20	5.04	5.64	1.25	0.72	0.27
5000	3.34	5.50	5.29	5.36	2.12	1.64	1.20

(a) $p = 1\%$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	4.04	4.97	5.21	5.28	4.00	4.04	3.38
1000	5.56	5.34	5.09	5.22	6.94	7.75	8.79
1500	6.33	5.04	5.02	4.99	6.80	7.83	9.55
2500	5.62	4.99	5.11	5.15	6.04	6.90	8.51
5000	5.01	4.95	4.93	4.99	5.30	5.52	6.16

(b) $p = 5\%$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	5.04	5.16	5.12	5.31	6.78	7.79	8.15
1000	5.28	5.15	5.06	5.06	5.66	6.40	8.16
1500	5.12	4.93	4.96	5.01	5.33	5.73	7.42
2500	5.22	5.08	4.99	5.23	5.29	5.36	6.29
5000	5.01	4.91	5.05	5.01	5.15	5.14	5.53

(c) $p = 10\%$

Table 1: ERF when simulating under the null hypothesis of Conditional Coverage (the empirical size) and using the asymptotic distributions 95% critical value. The hit-sequences were drawn as i.i.d. $Bernoulli(p)$ sequences. The results reported are based on $100,000$ replications for each test and sample size. The test names refer to the generalized Markov and generalized Duration tests developed in this paper, the Markov-1 test is also found in Christoffersen (1998) as the joint test.

3.2 Empirical power

To evaluate the power, the probability of rejecting $\theta \in \Theta_0$ when $\theta \notin \Theta_0$, of the tests of conditional coverage we specify a *DGP* using the generalized Markov specification, which is itself a special case of the generalized duration specification. Let the Markov chain be given by,

$$I_t|\mathcal{F}_{t-1,k} \underset{i.i.d}{\sim} Bernoulli(p_t(\theta)), \quad \mathcal{F}_{t-1,k} = I_{t-1}, ..., I_{t-k}, \qquad t = 1, ..., T. \tag{3.1}$$

with transition probabilities of (3.1) given by

$$p_t(\theta) = J_{t-1}p_E + (1 - J_{t-1})p_S, \quad J_{t-1} := 1\left(\sum_{i=1}^{k} I_{t-1} > 0\right).$$

We use $p_S = 1\%$ and $p_E = 3\%$ with $k = 5$ for the first DGP, this corresponds to the hit-sequence of a VaR forecast of coverage rate 1% which is misspecified in such a way that for 5 days following a hit the actual quantile modeled is the 3% quantile. The second DGP is identical, except $p_E = 4\%$. We repeat these simulations using $k = 10$, giving a total of 4 DGPs. The resulting empirical power[2] of the backtests are presented in figure 4.1. Note that we use the Monte Carlo testing technique of Dufour (2006) rather than the critical values implied by the asymptotic distributions in evaluating the tests.

From the first row of figure 4.1 it can be seen that for $p_S = 1\%$ and $p_E = 3\%$ the attained power can be quite limited, for example, when $1,000$ observations are available the empirical power never exceeds 50%. The DGP with the lowest order of dependence, $k = 5$, also has a lower empirical power compared to the DGP with $k = 10$ order dependence. Further, the tests which correctly specify the DGP as the alternative-hypothesis attain the highest power. The second row of figure 4.1, displaying results for the $p_S = 1\%$ and $p_E = 4\%$ cases, shows markedly higher empirical power, indicating that a more incorrectly specified VaR model, as measured by the difference between the intended quantile and the actually forecasted quantile, will be easier to identify. Lastly we see that using the tests with the highest empirical power can greatly improve the empirical power compared to the joint test of Christoffersen (1998).

[2]Strictly speaking, the term empirical power is only appropriate for those tests based on models which contain the DGP as a special case, eg. for $k = 5$ the simulations indicate the empirical power of the Markov-5, Duration-5, Duration-10 and Duration-20 tests.

3.3 Scenario Power Using GARCH Returns and Historical Simulation

The scenario power simulation consists of two elements, a model with parameters matching those found in empirical studies for generating non-i.i.d. returns and a forecast method which does not produce a valid forecast. Similar to Christoffersen (2004), Haas (2006), Berkowitz et al. (2011) and Candelon et al. (2011), we thus simulate a series of returns from a GARCH model and estimate VaR using the popular *HS* method[3].

Specifically, let the returns, R_t, be generated by a $GARCH(1,1)-t(d)$ with a skew and a conditional t distribution as:

$$R_t = \sigma_t z_t \sqrt{\frac{d-2}{d}}, \tag{3.2}$$

where the conditional variance is given by,

$$\sigma_t^2 = \omega + \alpha\sigma_{t-1}^2 \left(\sqrt{\frac{d-2}{d}} z_{t-1} - \theta \right)^2 + \beta\sigma_{t-1}^2. \tag{3.3}$$

Here z_t is an i.i.d. draw from a student t-distribution with d degrees of freedom. The parameter values are similar to estimates of this GARCH model on daily S&P500 returns, see Christoffersen (2004). Specifically, we set $d = 8$ degrees of freedom with parametrization of the coefficients as $\alpha = 0.1$, $\theta = 0.5$, $\beta = 0.85$ and $\omega = 3.9683e^{-6}$. The value of ω was set to target an annual standard deviation of 0.20 and the parametrization implies a daily volatility persistence of 0.975. We use a burn-in period of $5,000$ observations for each simulation to remove traces from initialization of the process. For more details see Christoffersen (2004) which presents figures of the generated returns, estimated VaR using HS and hazard functions of the hit-sequence from a similar simulation experiment.

Forecasting $\text{VaR}_{t|t-1}(p)$ is done using HS, see equation (3.4). HS is known to be under-responsive to changes in conditional risk as it assigns an equal probability weight of $1/T_W$ to all past observations, ignoring the temporal ordering. Furthermore, the method responds asymmetrically, increasing risk (as measured by VaR) following large losses but not following large gains. See Pritsker (2006) for a thorough discussion of the problems associated with HS. HS generates a hit-sequence which violates conditional coverage, both due to dependence and an incorrect coverage rate approximately 1% point larger than p.

The forecast is found by taking the negative empirical p percentile, of a rolling window of the T_W latest returns. We set T_W to be either 250 or 500, both lengths are used so that we may evaluate the robustness of the results with respect to changes in the data generating process.

$$\text{VaR}_{t|t-1}(p) = -percentile\left(\{R_j\}_{j=t-T_W}^{t-1}, p \right), \qquad t = 1, ..., T \tag{3.4}$$

Because the forecast is slow to update to changes in volatility, this will generate clusters of violations. We then use the returns and VaR forecasts to create the hit-sequence as specified in definition 2.1, that is to say $\{I_t\}_{t=1}^T$ is

[3] Perignon and Smith (2010) find that 73% of banks that disclosed their VaR forecast method used HS.

$$I_t := 1(R_t < \mathrm{VaR}_{t|t-1}(p)), \qquad t = 1, ...T$$

The resulting ERFs of the backtests are presented in figure 4.2. Note that we use the Monte Carlo testing technique of Dufour (2006) rather than the critical values implied by the asymptotic distributions in evaluating the tests.

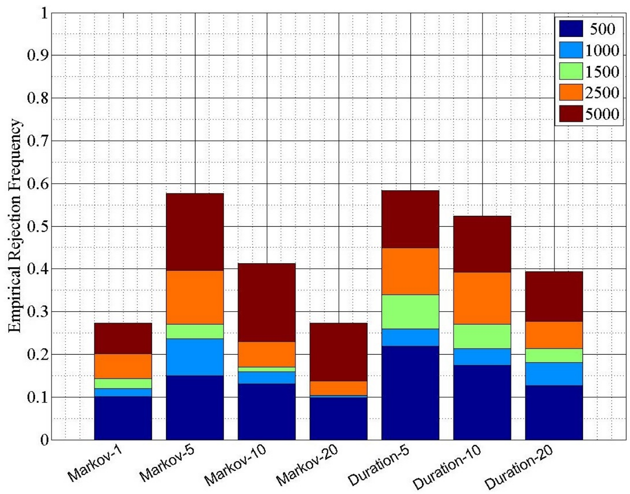
Inspecting figure 4.2, it is clear that the duration and generalized Markov tests improve the ERF compared with the original joint test. For example, when $1,000$ observations are available, using either the generalized Markov test or generalized duration tests with $k = 10$ lags will roughly double the ERF for either coverage rate. For the lower coverage rate, the Markov Duration test appears to perform slightly better than the generalized Markov test. However for the higher coverage rate the duration test can perform much worse, indicating its power is less robust (though still better than the original test). The results seem quite robust to the choice of T_W, although in general slightly better power was found when using $T_W = 500$ for all tests. This last result is as expected, since a longer window would be expected to increase the dependence in the hit-sequence.

4 Concluding Remarks

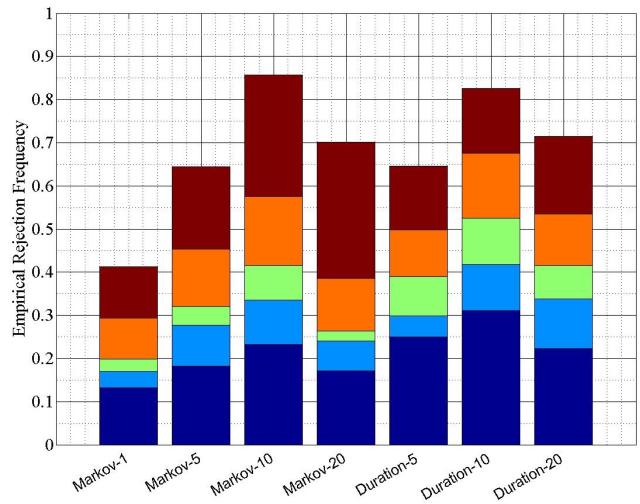
To summarize, we have introduced the generalized Markov framework for deriving Value-at-Risk backtests. Using the generalized Markov framework we suggested two specifications within this framework, the generalized Markov specification and the Markov Duration specification, inspired by the original backtests of Christoffersen (1998) and of the duration based backtests due to Christoffersen (2004), Haas (2006) and Wei and Pelletier (2014).

Based on these specifications we derived likelihood-ratio test statistics for the criteria of independence, unconditional coverage and conditional coverage. We provided closed form expressions for the tests as well as asymptotic theory. Our tests have the advantage, compared to the original tests of Christoffersen (1998). That they possess power against k'th order dependence. Furthermore, the tests of conditional coverage is equivalent to the sum of the tests for independence and unconditional coverage. This allows one to evaluate rejection of conditional coverage as being caused by either dependence, an incorrect coverage rate or both.

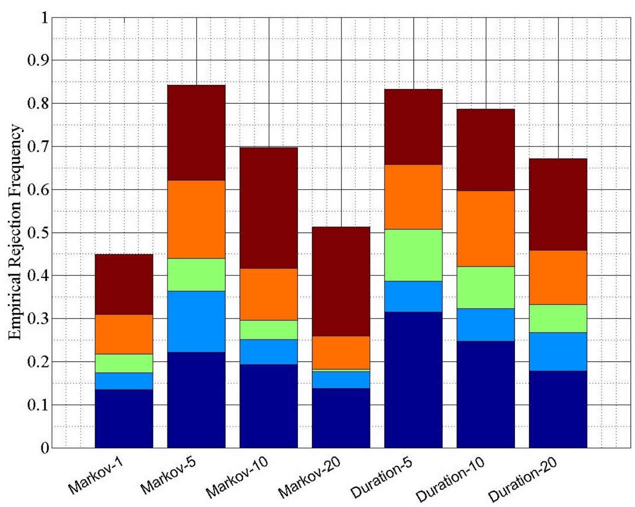
Using a simulation study we found evidence of improved size properties for the generalized Markov test compared to the original Markov test of Christoffersen (1998), though worse size properties for the Markov duration test. Simulations also indicated much improved empirical power while correcting for size distortions for the tests of conditional coverage based on either the generalized Markov or Duration specification compared to the original Markov test of Christoffersen (1998).



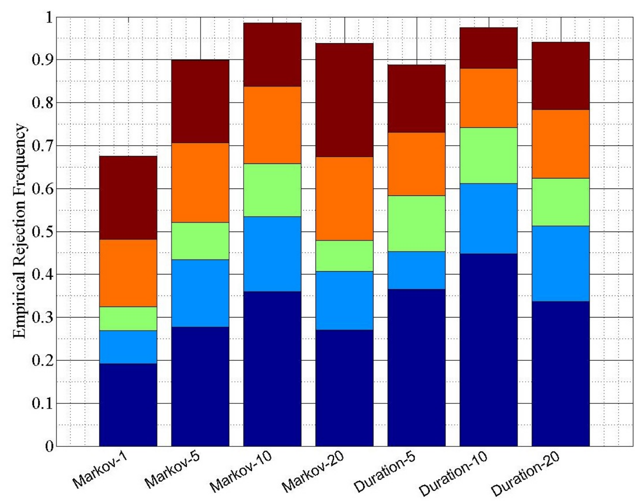
(a) $k = 5$ and $p_E = 3\%$



(b) $k = 10$ and $p_E = 3\%$

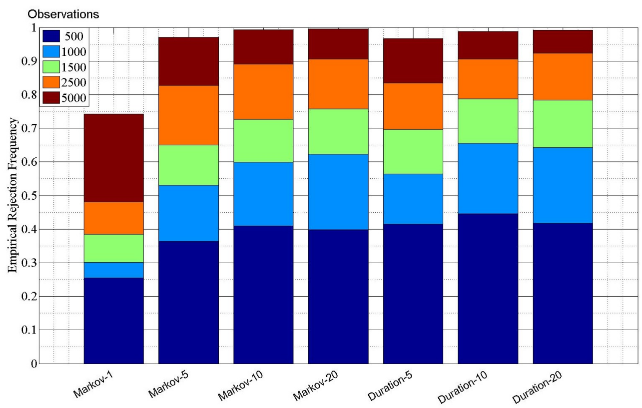


(c) $k = 5$ and $p_E = 4\%$

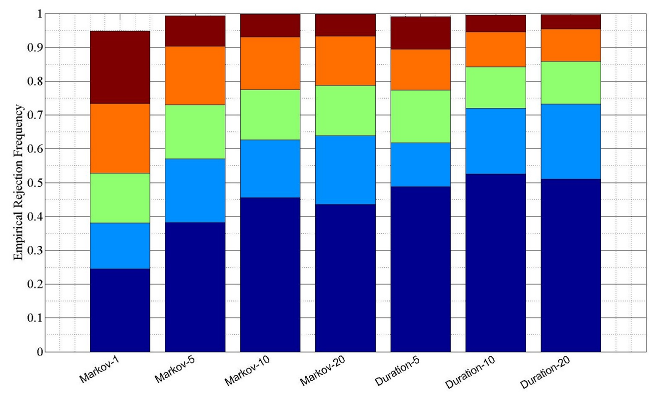


(d) $k = 10$ and $p_E = 4\%$

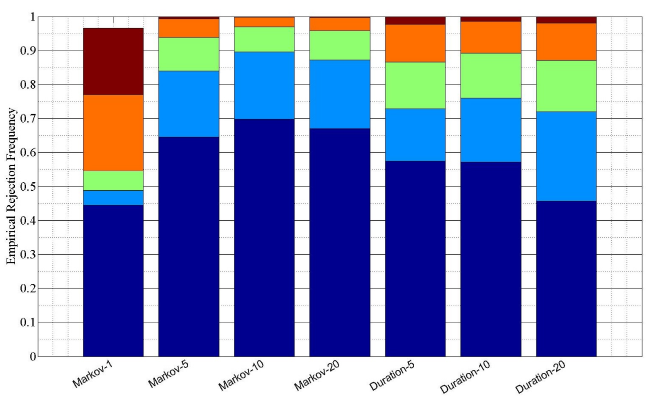
Figure 4.1: Empirical power in percent for the conditional coverage tests. The hit-sequences were simulated using the Markov chain described in section 3.2. The tests are the generalized Markov and generalized Duration tests developed in this paper, the Markov-1 test is also found in Christoffersen (1998) as the joint test. The Monte Carlo testing technique of Dufour (2006) was used to ensure a size of 5%.



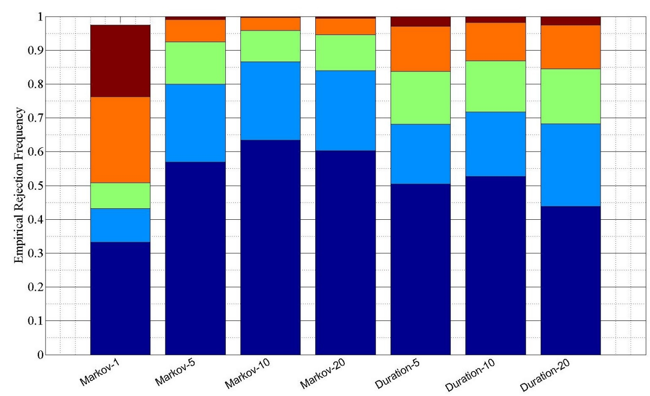
(a) $p = 1\%$ and $T_W = 500$



(b) $p = 1\%$ and $T_W = 250$



(c) $p = 5\%$ and $T_W = 500$



(d) $p = 5\%$ and $T_W = 250$

Figure 4.2: ERFs in percent for conditional coverage tests. The hit-sequences were simulated using a GARCH DGP with $VaR_{t|t-1}(p)$ estimated by historical simulation, using a rolling window of length T_W. The tests are the generalized Markov and generalized Duration tests developed in this paper, the Markov-1 test is also found in Christoffersen (1998) as the joint test. The Monte Carlo testing technique of Dufour (2006) was used to ensure a size of 5%.

References

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011, December). Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science 57*(12), 2213–2227.

Billingsley, P. (1962). *statistical inference for Markov processes.* Chicago Press.

Brown, B. (1971). Martingale central limit theorems,. *The Annals of Mathematical Statistics 42*(1), 59–66.

Campbell, S. D. (2007). A review of backtesting and backtesting procedures. *Journal of Risk 9*(2), 1–18.

Candelon, B., G. Colletaz, C. Hurlin, and S. Tokpavi (2011, Spring). Backtesting Value-at-Risk: A GMM Duration-Based Test. *Journal of Financial Econometrics 9*(2), 314–343.

Christoffersen, P. (2004). Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics 2*(1), 84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–62.

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

Haas, M. (2006). Improved duration-based backtesting of value-at-risk. *Journal of Risk 8*(2), 17–38.

Jensen, S. r. T. and A. Rahbek (2004, December). Asymptotic Inference For Nonstationary Garch. *Econometric Theory 20*(06), 1203–1226.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. Finance and Economics Discussion Series 95-24, Board of Governors of the Federal Reserve System (U.S.).

Lehmann, E. (1999). *Elements of Large-Sample Theory.* Springer.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools (Princeton Series in Finance).* Princeton University Press.

Perignon, C. and D. R. Smith (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance 34*(2), 362–377.

Pritsker, M. (2006, February). The hidden dangers of historical simulation. *Journal of Banking & Finance 30*(2), 561–582.

Wei, W. and D. Pelletier (2014). The geometric-var backtesting method. Workingpaper.

A Derivation of the Generalized Markov Test Distributions

A.1 Proof of Asymptotic Distribution for the Conditional Coverage Test

The proof verifies the conditions in Lemma 1 of Jensen and Rahbek (2004) for asymptotic inference, see also Theorem 7.7.3 of Lehmann (1999) or Billingsley (1962).

I. The score of the likelihood evaluated at the true value θ_0 satisfies $\frac{1}{\sqrt{T}} S_T(\theta_0) \xrightarrow{d} N(0, \Sigma)$ as $T \to \infty$

II. The observed information of the likelihood evaluated at the true value θ_0 satisfies $\frac{1}{T} i_T(\theta_0) \xrightarrow{p} \Sigma$ as $T \to \infty$

III. $\underset{\theta \in N(\theta_0)}{Sup} \frac{1}{T} \left| \frac{\partial^3 L_T(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| \leq C_T \xrightarrow{p} C \leq \infty$ as $T \to \infty$ and where $N(\theta_0)$ is a compact neighborhood around the true value θ_0 and $\theta_i, \theta_j, \theta_k = \{p_E, p_S\}$.

Condition (I): Recalling that $J_{t-1} = 1\left(\sum_{i=1}^{k} I_{t-1} > 0\right)$, the log-likelihood conditional on first k observations fixed is given by

$$L_T(\theta) = \sum_{t=1}^{T} L_t(\theta) = \sum_{t=1}^{T} I_t log\left(J_{t-1}p_E + (1 - J_{t-1})p_S\right) + (1 - I_t) log\left(1 - (J_{t-1}p_E + (1 - J_{t-1})p_S)\right)$$

Next, the score with respect to θ is given by,

$$S_T(\theta) = \sum_{t=1}^{T} s_t(\theta) = \sum_{t=1}^{T} \frac{\partial L_t(\theta)}{\partial \theta} = \sum_{t=1}^{T} \begin{bmatrix} \frac{I_t J_{t-1}}{J_{t-1}p_E + (1-J_{t-1})p_S} - \frac{(1-I_t)J_{t-1}}{1-(J_{t-1}p_E+(1-J_{t-1})p_S)} \\ \frac{I_t(1-J_{t-1})}{J_{t-1}p_E+(1-J_{t-1})p_S} - \frac{(1-I_t)(1-J_{t-1})}{1-(J_{t-1}p_E+(1-J_{t-1})p_S)} \end{bmatrix} = \begin{bmatrix} \frac{T_{11}}{p_E} - \frac{T_{10}}{1-p_E} \\ \frac{T_{01}}{p_S} - \frac{T_{00}}{1-p_S} \end{bmatrix}$$

Here $T_{11} := \sum_{t=1}^{T} I_t J_{t-1}$, $T_{01} := \sum_{t=1}^{T} I_t(1 - J_{t-1})$, $T_{10} := \sum_{t=1}^{T}(1 - I_t)J_{t-1}$, $T_{00} := \sum_{t=1}^{T}(1 - I_t)(1 - J_{t-1})$. Recalling the definition of $\hat{\theta}$ we have that

$$\hat{\theta} = \begin{bmatrix} \frac{T_{11}}{T_{10}+T_{11}} \\ \frac{T_{01}}{T_{00}+T_{01}} \end{bmatrix} = \begin{bmatrix} \frac{T_{11}}{T_1} \\ \frac{T_{01}}{T_0} \end{bmatrix},$$

with $T_1 := T_{11} + T_{01}$, $T_0 := T_{10} + T_{00}$.

The distribution of $S_T(\theta_0)$, where θ_0 is the true value of $\theta \in \Theta_H$, can be found as

$$S_T(\theta_0) = \sum_{t=1}^{T} s_t(\theta_0) = \sum_{t=1}^{T} \frac{1}{p(1-p)}(I_t - p) \begin{bmatrix} J_{t-1} \\ 1 - J_{t-1} \end{bmatrix}.$$

Since $s_t(\theta_0)$ is a vector of martingale difference sequences with respect to $\mathcal{F}_{t-1}$, with conditional covariance matrix

$$E(s_t(\theta_0)s_t(\theta_0)') = E \begin{bmatrix} \frac{(I_t-p)^2 J_{t-1}}{p^2(1-p)^2} & 0 \\ 0 & \frac{(I_t-p)^2(1-J_{t-1})}{p^2(1-p)^2} \end{bmatrix} = \begin{bmatrix} \frac{(1-(1-p)^k)}{p(1-p)} & 0 \\ 0 & \frac{(1-p)^k}{p(1-p)} \end{bmatrix} =: \Sigma$$

and as $s_t(\theta_0)$ is stationary with finite third order moments, it follows from the martingale difference central limit theorem in Brown (1971) that as $T \to \infty$

$$\frac{1}{\sqrt{T}} S_T(\theta_0) \xrightarrow{d} \Sigma^{1/2} U$$

where $U := wlim \left(\Sigma^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{p(1-p)} (I_t - p) \begin{bmatrix} J_{t-1} \\ 1 - J_{t-1} \end{bmatrix} \right) = N(0, I_2)$ and where I_2 is the identity matrix.

Condition (II): The observed information is given by

$$i_T(\theta) := -\frac{\partial^2 L_T(\theta)}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{T_{11}}{p_E^2} + \frac{T_{10}}{(p_E-1)^2} & 0 \\ 0 & \frac{T_{01}}{p_S^2} + \frac{T_{00}}{(p_S-1)^2} \end{bmatrix}.$$

It follows that as $T \to \infty$

$$\begin{aligned} \frac{1}{T} i_T(\theta_0) &\xrightarrow{p} \begin{bmatrix} \frac{p\left(1-(1-p)^k\right)}{p^2} + \frac{(1-p)\left(1-(1-p)^k\right)}{(p-1)^2} & 0 \\ 0 & \frac{p(1-p)^k}{p^2} + \frac{(1-p)(1-p)^k}{(p-1)^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1-(1-p)^k}{p(1-p)} & 0 \\ 0 & \frac{(1-p)^k}{p(1-p)} \end{bmatrix} = \Sigma, \end{aligned}$$

By using the law of large numbers for i.i.d. observations. Observe in particular that as $T \to \infty$, $\frac{1}{T} T_{11} = \frac{1}{T} \sum_{t=1}^{T} I_t J_{t-1} \xrightarrow{p} p\left(1 - (1-p)^k\right)$, $\frac{1}{T} T_{10} = \frac{1}{T} \sum_{t=1}^{T} (1 - I_t) J_{t-1} \xrightarrow{p} (1-p)\left(1 - (1-p)^k\right)$, $\frac{1}{T} T_{00} = \frac{1}{T} \sum_{t=1}^{T} (1 - I_t)(1 - J_{t-1}) \xrightarrow{p} (1-p)(1-p)^k$ and $\frac{1}{T} T_{01} = \frac{1}{T} \sum_{t=1}^{T} I_t (1 - J_{t-1}) \xrightarrow{p} p(1-p)^k$.

Condition (III): Define $0 < p_E^L \leq p_E \leq p_E^U < 1$ and $0 < p_S^L \leq p_S \leq p_S^U < 1$ such that $max(p_S^L, p_E^L) \leq p \leq min(p_S^U, p_E^U)$. We verify, using the above results, that since $\frac{\partial^2 L_T(\theta)}{\partial p_E \partial p_S} = 0$ it follows that

$$\frac{1}{T} \left| \frac{\partial^3 L_T(\theta)}{\partial^3 p_E} \right| = \frac{1}{T} \left| \frac{2T_{11}}{\left(p_E^L\right)^3} + \frac{2T_{10}}{\left(p_E^U - 1\right)^3} \right| \leq \frac{1}{T} \left(\frac{2T_{11}}{\left(p_E^L\right)^3} + \frac{2T_{10}}{\left(p_E^U - 1\right)^3} \right) = C_T \xrightarrow{p} c, \ for\ T \to \infty$$

$$\frac{1}{T} \left| \frac{\partial^3 L_T(\theta)}{\partial^3 p_S} \right| = \frac{1}{T} \left| \frac{2T_{01}}{\left(p_S^L\right)^3} + \frac{2T_{10}}{\left(p_S^U - 1\right)^3} \right| \leq \frac{1}{T} \left(\frac{2T_{01}}{\left(p_S^L\right)^3} + \frac{2T_{10}}{\left(p_S^U - 1\right)^3} \right) = C_T \xrightarrow{p} c, \ for\ T \to \infty$$

Having verified the conditions we can derive the Q_{G-CC} test statistics asymptotic distribution for $T \to \infty$ as

$$
\begin{aligned}
Q_{G-CC} &= -2\left(L_T(\theta_0) - L_T(\hat{\theta})\right) \\
&= (\hat{\theta}-\theta_0)^{'} i(\theta_0)(\hat{\theta}-\theta_0) + o_p(1) \\
&= \frac{1}{\sqrt{T}} S_T(\theta_0)^{'} \left(\frac{1}{T} i(\theta_0)\right)^{-1} \frac{1}{\sqrt{T}} S_T(\theta_0) + o_p(1) \\
&\xrightarrow{d} U^{'}\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}U = U^{'}U \sim \chi^2(2)
\end{aligned}
$$

A.2 Proof of Asymptotic Distribution for the Unconditional Coverage Test

The asymptotic distribution of the Q_{G-UC} test is found in the same fashion using

$$
\frac{\partial L(\theta)}{\partial p} = \frac{\partial L(\theta)}{\partial \theta}\frac{\partial \theta}{\partial p} = S_T(\theta)^{'} H, \quad -\frac{\partial^2 L(\theta)}{\partial p \partial p} = H^{'} i(\theta_0)_T H
$$

where we recall that $H = (1,1)'$and the definition of $\tilde{\theta}$, it then follows that

$$
\tilde{\theta} = \frac{T_1}{T_0 + T_1}.
$$

Then as $T \to \infty$

$$
\begin{aligned}
Q_{G-UC} &= -2\left(L_T(\theta_0) - L_T(\tilde{\theta})\right) \\
&= [\frac{1}{\sqrt{T}} S_T(\theta)^{'} H]^{'} \left(\frac{1}{T} H^{'} i_T(\theta_0) H\right)^{-1} [\frac{1}{\sqrt{T}} S_T(\theta)^{'} H] + o_p(1) \\
&\xrightarrow{d} U^{'}\Sigma^{1/2} H (H^{'}\Sigma H)^{-1} H \Sigma^{1/2} U \\
&\sim \chi^2(1)
\end{aligned}
$$

A.3 Proof of Asymptotic Distribution for the Independence Test

Using the projection $I = \Sigma^{1/2}H(H^{'}\Sigma H)^{-1}H^{'}\Sigma^{1/2} + \Sigma^{-1/2}H_{\perp}(H_{\perp}^{'}\Sigma^{-1}H_{\perp})^{-1}H_{\perp}^{'}\Sigma^{-1/2}$, where $H_{\perp}$ designates the orthogonal complement of H, we can now find the asymptotic distribution of Q_{G-Ind} as $T \to \infty$

$$
\begin{aligned}
Q_{G-Ind} &= Q_{G-CC} - Q_{G-UC} \\
&\xrightarrow{d} U^{'}U - U^{'}\Sigma^{1/2}H(H^{'}\Sigma H)^{-1}H^{'}\Sigma^{1/2}U \\
&= U^{'}\left(I - \Sigma^{1/2}H(H^{'}\Sigma H)^{-1}H^{'}\Sigma^{1/2}\right)U \\
&= U^{'}\Sigma^{-1/2}H_{\perp}(H_{\perp}^{'}\Sigma^{-1}H_{\perp})^{-1}H_{\perp}^{'}\Sigma^{-1/2}U = A^{'}A \sim \chi^2(1),
\end{aligned}
$$

where $A := (H_{\perp}^{'}\Sigma^{-1}H_{\perp})^{-1/2}H_{\perp}^{'}\Sigma^{-1/2}U$.

B Derivation of the Markov Duration Test Distributions

B.1 Proof of Asymptotic Distribution for the Conditional Coverage Test

We proceed as in the proof for Theorem 1.

Condition (I): Recalling that $J(k)_{t-1} = 1\,(I_{t-1} = 0, ..., I_{t-k} = 1)$, the log-likelihood conditional on first k observations fixed is given by

$$
\begin{aligned}
L_T(\theta) &= \sum_{t=1}^{T} I_t log\left(J(1)_{t-1}p_{E1} + ... + J(k)_{t-1}p_{Ek} + (1 - \sum_{i=1}^{k} J(i)_{t-1})p_S\right) \\
&\quad +(1 - I_t)log\left(1 - \left(J(1)_{t-1}p_{E1} + ... + J(k)_{t-1}p_{Ek} + (1 - \sum_{i=1}^{k} J(i)_{t-1})p_S\right)\right)
\end{aligned}
$$

Next, the score with respect to θ is given by,

$$
S_T(\theta) = \sum_{t=1}^{T} s_t(\theta) = \sum_{t=1}^{T} \frac{\partial L_t(\theta)}{\partial \theta} = \sum_{t=1}^{T} \begin{bmatrix} \frac{I_t J(1)_{t-1}}{p_{E1}} - \frac{(1-I_t)J(1)_{t-1}}{1-p_{E1}} \\ ... \\ \frac{I_t J(k)_{t-1}}{p_{Ek}} - \frac{(1-I_t)J(k)_{t-1}}{1-p_{Ek}} \\ \frac{I_t(1-\sum_{i=1}^{k} J(i)_{t-1})}{p_S} - \frac{(1-I_t)(1-\sum_{i=1}^{k} J(i)_{t-1})}{1-p_S)} \end{bmatrix} = \begin{bmatrix} \frac{T_{11}(1)}{p_{E1}} - \frac{T_{10}(1)}{1-p_{E1}} \\ \vdots \\ \frac{T_{11}(k)}{p_{Ek}} - \frac{T_{10}(k)}{1-p_{Ek}} \\ \frac{T_{11}}{p_S} - \frac{T_{00}}{1-p_S} \end{bmatrix}
$$

Recalling the definition of $\hat{\theta}$ we have that

$$
\hat{\theta} = \begin{bmatrix} \frac{T_{11}(1)}{T_{11}(1)+T_{10}(1)} \\ \vdots \\ \frac{T_{11}(k)}{T_{11}(k)+T_{10}(k)} \\ \frac{T_{11}}{T_{11}+T_{01}} \end{bmatrix},
$$

Recalling that $J_{t-1} \coloneqq 1\left(\sum_{i=1}^{k} I_{t-1} > 0\right)$, the distribution of $S_T(\theta_0)$, where θ_0 is the true value of $\theta \in \Theta_H$, can be found as

$$
S_T(\theta_0) = \sum_{t=1}^{T} s_t(\theta_0) = \sum_{t=1}^{T} \frac{1}{p(1-p)}(I_t - p) \begin{bmatrix} J(1)_{t-1} \\ \vdots \\ J(k)_{t-1} \\ 1 - J_{t-1} \end{bmatrix}.
$$

Since $s_t(\theta_0)$ is a vector of martingale difference sequences with respect to $\mathcal{F}_{t-1}$, with conditional covariance matrix

$$E(s_t(\theta_0)s_t(\theta_0)') = \begin{bmatrix} \frac{1}{(1-p)} & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & (1-p)^{k-2} & 0 \\ 0 & 0 & 0 & \frac{(1-p)^{k-1}}{p} \end{bmatrix} =: \Sigma$$

and as $s_t(\theta_0)$ is stationary with finite third order moments, it follows from the martingale difference central limit theorem in Brown (1971) that as $T \to \infty$

$$\frac{1}{\sqrt{T}} S_T(\theta_0) \xrightarrow{d} \Sigma^{1/2} U$$

where $U := wlim \left(\Sigma^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{p(1-p)} (I_t - p) \begin{bmatrix} J(1)_{t-1} \\ \vdots \\ J(k)_{t-1} \\ 1 - J_{t-1} \end{bmatrix} \right) = N(0, I_{k+1})$ and where I_{k+1} is the identity matrix.

Condition (II): The observed information is given by

$$i_T(\theta_0) := -\frac{\partial^2 L_T(\theta)}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{T_{11}(1)}{p_{E1}^2} + \frac{T_{10}(1)}{(1-p_{E1})^2} & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \frac{T_{11}(k)}{p_{Ek}^2} + \frac{T_{10}(k)}{(1-p_{Ek})^2} & 0 \\ 0 & 0 & 0 & \frac{T_{01}}{p_S^2} + \frac{T_{00}}{(1-p_S)^2} \end{bmatrix}$$

It follows that as $T \to \infty$

$$\frac{1}{T} i_T(\theta_0) \quad \xrightarrow{p} \quad \begin{bmatrix} \frac{1}{(1-p)} & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & (1-p)^{k-2} & 0 \\ 0 & 0 & 0 & \frac{(1-p)^{k-1}}{p} \end{bmatrix} = \Sigma$$

Here we have used the law of large numbers for i.i.d. observations, and that as $T \to \infty$, $\frac{1}{T} T_{11}(k) \to (1-p)^{k-1} p^2$ and $\frac{1}{T} T_{01}(k) \to (1-p)^k p$.

Condition (III): Define $0 < p_{Ek}^L \leq p_{Ek} \leq p_{Ek}^U < 1$ and $0 < p_S^L \leq p_S \leq p_S^U < 1$ such that $max(p_S^L, p_E^L) \leq p \leq min(p_S^U, p_E^U)$. Since $\frac{\partial^2 L_T(\theta)}{\partial p_E \partial p_S} = 0$ it follows that we only need the following

$$\frac{1}{T}\left|\frac{\partial^3 L_T(\theta)}{\partial^3 p_{Ek}}\right| = \tfrac{1}{T} \left|\frac{2T_{11}(k)}{\left(p_{Ek}^L\right)^3}+\frac{2T_{10}(k)}{\left(p_{Ek}^U-1\right)^3}\right| \leq \frac{1}{T}\left(\frac{2T_{11}(k)}{\left(p_{Ek}^L\right)^3}+\frac{2T_{10}(k)}{\left(p_{Ek}^U-1\right)^3}\right) = C_T \overset{p}{\to} c,\ for\ T\to\infty$$

$$\frac{1}{T}\left|\frac{\partial^3 L_T(\theta)}{\partial^3 p_S}\right| = \frac{1}{T}\left|\frac{2T_{0,1}}{\left(p_S^L\right)^3}+\frac{2T_{1,0}}{\left(p_S^U-1\right)^3}\right| \leq \frac{1}{T}\left(\frac{2T_{0,1}}{\left(p_S^L\right)^3}+\frac{2T_{1,0}}{\left(p_S^U-1\right)^3}\right) = C_T \overset{p}{\to} c,\ for\ T\to\infty$$

Having verified the conditions we can derive the Q_{D-CC} test statistics asymptotic distribution for $T\to\infty$ as

$$\begin{aligned}
Q_{D-CC} &= -2\left(L_T(\theta_0) - L_T(\hat\theta)\right)\\
&= (\hat\theta-\theta_0)' i(\theta_0)(\hat\theta-\theta_0) + o_p(1)\\
&= \frac{1}{\sqrt{T}}S_T(\theta_0)'\left(\frac{1}{T}i(\theta_0)\right)^{-1}\frac{1}{\sqrt{T}}S_T(\theta_0)+o_p(1)\\
&\overset{d}{\to} U'\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}U = U'U \sim \chi^2(k+1)
\end{aligned}$$

B.2 Proof of Asymptotic Distribution for the Unconditional Coverage Test

The asymptotic distribution of the Q_{D-UC} test is found in the same fashion using

$$\frac{\partial L(\theta)}{\partial p} = \frac{\partial L(\theta)}{\partial \theta}\frac{\partial \theta}{\partial p} = S_T(\theta)'H, \quad -\frac{\partial^2 L(\theta)}{\partial p \partial p} = H' i(\theta_0)_T H$$

where we recall that $H=(1,...,1)'$and the definition of $\tilde\theta$, it then follows that

$$\tilde\theta = \frac{T_1}{T_0+T_1}.$$

Then as $T\to\infty$

$$\begin{aligned}
Q_{D-UC} &= -2\left(L_T(\theta_0)-L_T(\tilde\theta)\right)\\
&= [\frac{1}{\sqrt{T}}S_T(\theta)'H]'\left(\frac{1}{T}H' i_T(\theta_0)H\right)^{-1}[\frac{1}{\sqrt{T}}S_T(\theta)'H]+o_p(1)\\
&\overset{d}{\to} U'\Sigma^{1/2}H(H'\Sigma H)^{-1}H\Sigma^{1/2}U\\
&\sim \chi^2(1)
\end{aligned}$$

B.3 Proof of Asymptotic Distribution for the Independence Test

Using the projection $I = \Sigma^{1/2}H(H'\Sigma H)^{-1}H'\Sigma^{1/2} + \Sigma^{-1/2}H_\perp(H_\perp'\Sigma^{-1}H_\perp)^{-1}H_\perp'\Sigma^{-1/2}$ we can now find the asymptotic distribution of Q_{D-Ind} as $T\to\infty$

$$
\begin{aligned}
Q_{D-Ind} &= Q_{D-CC} - Q_{D-UC} \\
&\xrightarrow{d} U^{'}U - U^{'}\Sigma^{1/2}H(H^{'}\Sigma H)^{-1}H^{'}\Sigma^{1/2}U,\ for\ T \to \infty \\
&= U^{'}\left(I - \Sigma^{1/2}H(H^{'}\Sigma H)^{-1}H^{'}\Sigma^{1/2}\right)U \\
&= U^{'}\Sigma^{-1/2}H_{\perp}(H_{\perp}^{'}\Sigma^{-1}H_{\perp})^{-1}H_{\perp}^{'}\Sigma^{-1/2}U = A^{'}A \sim \chi^2(1),
\end{aligned}
$$

where $A := (H_{\perp}^{'}\Sigma^{-1}H_{\perp})^{-1/2}H_{\perp}^{'}\Sigma^{-1/2}U$.

C The Monte Carlo Testing Technique Dufour (2006)

In this section we outline the Monte Carlo testing technique of Dufour (2006) used in the empirical power simulations of section C. The technique used is given by the following algorithm:

I. Generate M i.i.d. hit-sequences of length T, $\{I_t\}_{t=1}^{T}$, under the null of conditional coverage, H_{CC}, by drawing from a Bernoulli sequence, as:

$$
I_t \underset{i.i.d}{\sim} Bernoulli(p), \qquad t = 1, ..., T
$$

II. Calculate the test statistic, S_i, for each of the generated hit-sequence, $i = 1, ..., M$ and denote by S_0 the original test value. Throughout this paper we use $M = 99,999$.

III. Draw U_i for $i = 0, ..., M$ from the uniform $U(0,1)$ distribution. Calculate the p-values as

$$
\hat{p}_M(S_0) = \frac{M\hat{G}_M(S_0) + 1}{M + 1}
$$

where

$$
\hat{G}_M(S_0) = 1 - \frac{1}{M}\sum_{i=1}^{M} 1\left(S_i \leq S_0\right) + \frac{1}{M}\sum_{i=1}^{M} 1\left(S_i = S_0\right) 1\left(U_i \geq U_0\right).
$$

D Power Tables

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	8.40	10.15	8.20	7.95	14.00	10.85	9.30
1000	7.25	11.20	7.75	6.40	13.20	10.40	9.30
1500	9.45	13.05	8.35	6.00	17.65	13.35	11.05
2500	10.75	16.45	9.30	6.30	20.90	19.20	13.05
5000	13.70	21.95	14.65	11.20	27.00	23.45	15.80

(a) $p = 1\%$ and $k = 5$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	9.65	11.65	12.20	9.85	14.70	17.30	12.65
1000	8.50	13.50	14.00	11.45	13.85	18.95	15.85
1500	8.85	13.90	15.10	9.45	18.75	23.25	18.40
2500	13.30	17.90	20.65	12.60	22.95	33.80	23.85
5000	16.05	23.85	39.80	24.20	27.35	40.60	29.10

(b) $p = 1\%$ and $k = 10$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	6.25	6.75	6.15	5.45	6.25	5.35	4.55
1000	7.40	8.40	7.10	4.70	7.30	5.65	4.40
1500	5.45	10.10	7.10	5.70	6.85	5.05	5.35
2500	8.05	16.45	11.35	9.80	9.90	7.60	6.80
5000	12.75	24.40	15.20	11.00	15.85	11.60	9.00

(c) $p = 5\%$ and $k = 5$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	7.00	7.50	7.05	8.25	8.45	8.10	6.30
1000	8.55	11.20	12.10	8.00	9.45	7.20	6.45
1500	7.70	12.15	13.85	10.95	8.60	7.75	7.40
2500	13.35	18.80	22.75	14.85	10.10	10.45	8.00
5000	20.40	27.75	39.80	27.20	17.40	19.00	13.95

(d) $p = 5\%$ and $k = 10$

Table 2: Empirical power in percent for conditional coverage tests. The hit-sequences were simulated using a k'th order Markov chain specified in equation (3.1) of section 3. The tests refer to the generalized Markov and generalized Duration tests developed in this paper, the Markov-1 test is also found in Christoffersen (1998) as the joint test. The Monte Carlo testing technique of Dufour (2006) was used to ensure a size of 5%

E Scenario Power Tables

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	25.56	36.39	40.94	39.81	41.42	44.58	41.70
1000	30.16	53.09	59.92	62.32	56.45	65.52	64.33
1500	38.47	65.02	72.61	75.72	69.61	78.74	78.36
2500	48.10	82.72	89.12	90.65	83.43	90.53	92.28
5000	74.27	97.01	99.26	99.59	96.69	98.79	99.25

(a) $p = 1\%$ and $T_W = 500$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	24.47	38.26	45.61	43.63	48.82	52.54	51.05
1000	38.05	57.04	62.72	63.85	61.84	72.05	73.21
1500	52.74	73.04	77.45	78.71	77.43	84.26	85.83
2500	73.35	90.31	93.10	93.30	89.44	94.59	95.48
5000	94.80	99.27	99.81	99.87	99.06	99.56	99.65

(b) $p = 1\%$ and $T_W = 250$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	44.48	64.55	69.72	66.98	57.36	57.22	45.65
1000	48.78	83.93	89.63	87.27	72.84	75.94	72.00
1500	54.50	93.86	96.93	95.84	86.60	89.19	87.06
2500	77.05	99.28	99.80	99.65	97.63	98.57	98.08
5000	96.61	100.00	100.00	100.00	99.99	100.00	100.00

(c) $p = 5\%$ and $T_W = 500$

Sample size	Markov-1	Markov-5	Markov-10	Markov-20	Duration-5	Duration-10	Duration-20
500	33.27	56.95	63.36	60.25	50.43	52.64	43.85
1000	43.15	80.02	86.60	83.93	68.09	71.72	68.28
1500	50.77	92.50	95.88	94.60	83.77	86.83	84.47
2500	76.29	99.03	99.72	99.48	97.02	98.14	97.38
5000	97.40	100.00	100.00	100.00	99.99	100.00	100.00

(d) $p = 5\%$ and $T_W = 250$

Table 3: ERFs in percent for conditional coverage tests. The hit-sequences were simulated using a GARCH DGP with $VaR_{t|t-1}(p)$ estimated by historical simulation, using a rolling window of length T_W. The tests refer to the generalized Markov and generalized Duration tests developed in this paper, the Markov-1 test is also found in Christoffersen (1998) as the joint test. The Monte Carlo testing technique of Dufour (2006) was used to ensure a size of 5%