Evaluating the Accuracy of Value-at-Risk Forecasts: New Multilevel Tests

Arturo Leccadito
Università della Calabria, Italy

Simona Boffelli
Università di Bergamo, Italy

Giovanni Urga*
Cass Business School, City University London, UK and
Università di Bergamo, Italy

July 9, 2013

Abstract

We propose independence and conditional coverage tests aimed at evaluating the accuracy of Value-at-Risk (VaR) forecasts from the same model at different confidence levels. The proposed procedures are multilevel tests, i.e. joint tests of several quantiles corresponding to different confidence levels. In a comprehensive Monte Carlo exercise, we document the superiority of the proposed tests with respect to existing multilevel tests. In an empirical application, we illustrate the implementation of the tests using several VaR models and daily data for 15 MSCI world indices.

Keywords: Risk Management, Value-at-Risk, Backtesting, Conditional and Unconditional Coverage Tests, Monte Carlo
JEL Classification: C12, C52, G28, G32

*Corresponding Author: Centre for Econometric Analysis, Faculty of Finance, Cass Business School, City University London, 106 Bunhill Row EC1Y 8TZ, London, UK. e-mail: G.Urga@city.ac.uk

Evaluating the Accuracy of Value-at-Risk Forecasts: New Multilevel Tests

Abstract

We propose independence and conditional coverage tests aimed at evaluating the accuracy of Value-at-Risk (VaR) forecasts from the same model at different confidence levels. The proposed procedures are multilevel tests, i.e. joint tests of several quantiles corresponding to different confidence levels. In a comprehensive Monte Carlo exercise, we document the superiority of the proposed tests with respect to existing multilevel tests. In an empirical application, we illustrate the implementation of the tests using several VaR models and daily data for 15 MSCI world indices.

Keywords: Risk Management, Value-at-Risk, Backtesting, Conditional and Unconditional Coverage Tests, Monte Carlo

JEL Classification: C12, C52, G28, G32

1 Introduction

Financial risk is related to the possibility that financial loss (or gain) due to unforeseen changes in underlying risk factors may take place. One particular type of financial risk is market risk, i.e. the risk of loss (or gain) arising from unexpected changes in market prices or market rates. Value-at-risk (VaR) is the most commonly used tool in financial risk management and it is widely used by financial institutions to evaluate the market risk exposure of their trading portfolios. VaR is the quantile of the distribution of gains and losses over a target horizon and as such it summarizes in a single value the possible losses which could occur with a given probability in a given temporal horizon. The VaR measure has been criticized for not being subadditive and hence violating the axioms of coherency (see Artzner et al. 1999), and alternative coherent risk measures such as Expected Shortfall have been proposed. We focus on VaR only because it is the most utilized risk measure in applied works and it is commonly used in the financial industry. In this paper, we propose two novel multilevel testing procedures for VaR prediction, to evaluate the accuracy of VaR forecasts from the same model at different confidence levels.

Over the last decade a wide array of parametric (for instance RiskMetrics and GARCH models) and non-parametric (for instance Historical Simulation) statistical methods have been proposed to quantify VaR. Since financial institutions are required to hold regulatory capital based on their VaR forecasts, ex post techniques aimed at validating their measure of market risk are required. Hence, if on one hand it is relevant for banks to implement accurate VaR models, on the other hand they need to use sound statistical backtests to validate them. In essence, backtesting procedures are constructed comparing realized returns and model-generated VaR measures. Commonly used backtests for VaR models include the likelihood ratio test of Kupiec (1995), the Markov tests of Christoffersen (1998) and the duration based

test of Christoffersen and Pelletier (2004). The one of Kupiec (1995) is an unconditional coverage test in the sense that it only measures how distant the nominal coverage rate is from the proportion of violations in the sample, i.e. the number of time the ex post loss exceeds the ex ante VaR. Christoffersen (1998) proposes an independence test aimed at verifying if there is any clustering in the violation sequence. The intuition is that in a good model, a VaR violation today should be independent of whether or not yesterday's VaR was violated. Testing both the unconditional coverage and the independence hypotheses results in the so called conditional coverage test. The test of Christoffersen and Pelletier (2004) is based instead on the duration sequence, i.e. the number of observations between two consecutive violations. Authors exploit the fact that under a correct VaR model durations should have a geometric distribution with average equal to the reciprocal of the coverage probability. Further developments of duration tests in a GMM framework can be found in the recent paper of Candelon et al. (2011). Another testing procedure is the one introduced in Engle and Manganelli (2004). The authors build their dynamic quantile test on the idea that, if the violations are a martingale difference sequence, the probability of exceeding the VaR must be independent of all the past information. The test is based on a regression of the violations on their lagged values and other lagged variables available when the VaR is computed.

Denoting by K the number of coverage probabilities used in the VaR estimation, the testing procedures described above are based on $K = 1$, i.e. they are unilevel procedures. For example, when estimating 1%, 2.5% and 5% quantiles, the standard unilevel approach is to perform a separate test for each of these three quantiles. Unilevel tests are known to have small power especially when the sample considered has a realistic (small) number of observations, as confirmed by the Monte Carlo study in Berkowitz et al. (2011). In a recent

paper, Perignon and Smith (2008) propose a multilevel test based on $K > 1$ coverage probabilities. Again, when estimating 1%, 2.5% and 5% quantiles, a multilevel test is a joint test for the coverage of all three quantiles. The test of Perignon and Smith (2008) stands between the unconditional coverage test of Kupiec (1995), that compares the fraction of days with a VaR violation with the nominal coverage probability, and the test of Berkowitz (2001), that allows one to test the entire distribution (or the left tail) via the Rosenblatt transformation. A similar approach is in Diebold et al. (1998). In this paper, however, the authors present only graphical analyses for diagnosing how models fail, rather than formal testing procedures. For instance, they notice that the transformed data should be uniformly distributed. Hence, if the model does not capture fat tails, the histogram of the transformed data will have peaks near zero and one. The Kupiec (1995) test is widely used among practitioners but displays low power when applied to financial datasets. The Berkowitz (2001) methodology, though more powerful, is not used in practice since banks are only willing to disclose one-day ahead VaR estimates and not the entire profit/loss distribution. Multilevel tests, namely the Perignon and Smith (2008) procedure, represent instead an optimal compromise between the two cases above: they show good performance in terms of power, while requiring only a limited information disclosure from banks or financial institutions. Furthermore, multilevel tests are useful and appealing firstly because it is common for quantiles to be estimated for two or more different confidence levels, and, secondly, because they represent a more efficient and statistically more powerful alternative with respect to separate unilevel tests. However, the Perignon and Smith (2008) procedure does not allow to test for the presence of clusters of VaR exceptions, which we allow in the multilevel tests we propose in this paper. Tests designed to detect whether the VaR violations are independent and the average number of violations is correct are called conditional coverage tests. The importance

of conditional coverage backtesting procedures is confirmed by the study of Berkowitz et al. (2011), who, using desk-level profit/loss data from four business lines in a large international commercial bank and Historical Simulation VaR estimates, document the presence of severe clustering in VaR violations for two of the four business lines. To the best of our knowledge, the only existing conditional coverage backtesting procedure valid in the multilevel context is the one of Hurlin and Tokpavi (2006). The testing procedure the authors propose is based on a multivariate extension of the Box and Pierce (1970) test, which is used to jointly test the absence of autocorrelation in the vector of violations for various coverage probabilities. In this paper, instead, we consider conditional coverage testing procedures in a multilevel framework by proposing the multilevel generalization of the Markov test of Christoffersen (1998) and a Pearson-type of test based on the bivariate distributions of the total number of VaR violations in period t, N_t, and its j^{th} lag N_{t-j}. N_t is obtained by summing up the usual indicator variables corresponding to different coverage rates obtained from the unilevel tests. In an extensive Monte Carlo study we show the superiority of the proposed tests with respect to the one of Hurlin and Tokpavi.

The paper is organized as follows. In Section 2, we review the existing multilevel tests and we introduce the Markov and Pearson-type tests that we propose in this paper. The power properties of the proposed tests are examined in a Monte Carlo study presented in Section 3. In Section 4, the multilevel tests are applied to univariate and multivariate VaR models in a backtesting exercise involving daily returns on 15 MSCI world indices. Section 5 concludes.

2 Multilevel backtesting procedures

Denote by r_t with $t = 1, \ldots, T$ the time series of log-returns or bank revenues we are interested in for backtesting purposes. Given a coverage probability α, VaR for time $t+1$, given the information up to time t, satisfies

$$P\left(r_{t+1} \leq -\mathrm{VaR}_{t+1|t}(\alpha)|\mathcal{F}_t\right) = \alpha,$$

where $\mathcal{F}_t$ denotes the information set at time t.

Consider K different critical levels $\alpha_1 > \alpha_2 > \ldots > \alpha_K$. The associated VaRs are in opposite monotonic order, namely

$$\mathrm{VaR}_{t+1|t}(\alpha_1) < \mathrm{VaR}_{t+1|t}(\alpha_2) < \ldots < \mathrm{VaR}_{t+1|t}(\alpha_K).$$

Using the notation and set-up of Perignon and Smith (2008), for each VaR measure an indicator variable is constructed as follows

$$J_{i,t+1} = \begin{cases} 1 & \text{if } -\mathrm{VaR}_{t+1|t}(\alpha_{i+1}) < r_{t+1} \leq -\mathrm{VaR}_{t+1|t}(\alpha_i) \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

for all $i = 1, \ldots, K$. By convention, we take $\alpha_{K+1} = 0$, $\mathrm{VaR}_{t+1|t}(\alpha_{K+1}) = +\infty$, and $J_{0,t+1} = \prod_{i=1}^{K}(1 - J_{i,t+1})$. With the additional convention that $\alpha_0 = 1$, the random variables $\{J_{i,t+1}\}_{i=0,\ldots,K}$ are Bernoulli distributed with probability $\theta_i = \alpha_i - \alpha_{i+1}$ under the null that the VaR model is unconditionally accurate. Hence θ_i represents, under the null, the probability of falling in between the VaR quantiles associated to the coverage probabilities α_i and α_{i+1}.

Since in any time period only one of those variables can be equal to one, these random variables are not independent. Furthermore, each J can be expressed as

$$J_{i,t+1} = I_{i,t+1} - I_{i+1,t+1}, \quad i = 1, \ldots, K,$$

where I is the usual exception indicator:

$$I_{i,t+1} = \begin{cases} 1 & \text{if } r_{t+1} \leq -\text{VaR}_{t+1|t}(\alpha_i) \\ 0 & \text{if } r_{t+1} > -\text{VaR}_{t+1|t}(\alpha_i) \end{cases}. \tag{2}$$

Consider the time series $\{N_t\}$ such that $N_{t+1} = i$ when $J_{i,t+1} = 1$ for $i = 0, \ldots, K$. Note that since

$$N_{t+1} = \sum_{i=1}^{K} I_{i,t+1}, \tag{3}$$

N_t represents the total number of VaR violations in period t at the different coverage probabilities[1]. Under the null that the VaR model is unconditionally accurate, the first two moments of N_{t+1} are

$$\mu = \mathbb{E}(N_{t+1}) = \sum_{i=1}^{K} i \cdot \theta_i = \sum_{i=1}^{K} \alpha_i \tag{4}$$

$$\mathbb{E}(N_{t+1}^2) = \sum_{i=1}^{K} i^2 \cdot \theta_i = \sum_{i=1}^{K} \alpha_i [i^2 - (i-1)^2] = \sum_{i=1}^{K} \alpha_i (2i-1) = 2\sum_{i=1}^{K} i \cdot \alpha_i - \mu \tag{5}$$

and hence

$$\sigma^2 = \text{Var}(N_{t+1}) = 2\sum_{i=1}^{K} i \cdot \alpha_i - \mu - \mu^2.$$

[1] We assume that in any given period we can observe at most one VaR violation for each coverage probability.

Note that $\sigma^2 \neq \sum_{i=1}^{K} \text{Var}(I_{i,t+1})$, because the indicators I are not independent.

In what follows, we briefly introduce the two existing multilevel backtesting procedures, namely the Perignon and Smith (2008) and the Hurlin and Tokpavi (2006) tests.

The Perignon and Smith (2008) test. A recent approach for backtesting VaR models is proposed by Perignon and Smith (2008). This is a generalization of Kupiec (1995) unconditional test to the case of K different critical levels and its null hypothesis can simply be tested using a standard chi-square goodness-of-fit test. The multivariate unconditional coverage test of Perignon and Smith (2008) is a likelihood ratio test that the empirical $\boldsymbol{\pi}$ significantly deviates from the hypothesized $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)'$. The null is

$$H_{0,uc} : \pi_i = \theta_i, \quad i = 0, 1, \ldots, K-1.$$

Collecting in the vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_K)'$ the observed probabilities of falling in between the VaR quantiles, the probability density of N is given by

$$g(n; \boldsymbol{\pi}) = P(N_{t+1} = n; \boldsymbol{\pi}) = P(J_{n,t+1} = 1; \boldsymbol{\pi}) = \prod_{i=0}^{K} \pi_i^{J_{i,t+1}} \tag{6}$$

and hence the log-likelihood function for a sample with T observations is

$$\ell(\boldsymbol{\pi}) = \sum_{t=1}^{T} \sum_{i=0}^{K} J_{i,t} \ln(\pi_i) = \sum_{i=0}^{K} T_i \ln(\pi_i), \tag{7}$$

where $T_i = \sum_{t=1}^{T} J_{i,t}$ is the number of observations in the sample for which $N_t = i$.

Denoting by $\ell_{0,uc}$ and by $\ell_{1,uc}$ the log-likelihoods under the null and the alternative

hypotheses, the test statistic is

$$LR_{uc} = 2\left(\ell_{1,uc} - \ell_{0,uc}\right) = 2\left(\ell(\widehat{\boldsymbol{\pi}}) - \ell(\boldsymbol{\theta})\right) = 2\left(\sum_{i=0}^{K} \ln(\widehat{\pi}_i/\theta_i)^{T_i}\right) \tag{8}$$

where $\widehat{\pi}_i$ is the maximum likelihood estimator of the i-th component of $\boldsymbol{\pi}$, and it is given by $\widehat{\pi}_i = \frac{T_i}{T}$ (see the Appendix). The test statistic is asymptotically chi-square with K degrees of freedom and when $K = 1$ one recovers the unconditional coverage test developed by Kupiec (1995).

Perignon and Smith (2008) stress the importance of multilevel testing procedures for VaR prediction. However, their procedure is only an unconditional coverage test, which is not capable of detecting clustering in the sequence of VaR violations. Hurlin and Tokpavi (2006), instead, propose a multilevel conditional coverage testing procedures.

The Hurlin and Tokpavi (2006) test. Hurlin and Tokpavi jointly test the absence of autocorrelation and cross-correlation in the vector of hit sequences for various coverage rates. Their null hypothesis is

$$H_0 : \mathbb{E}[(I_{h,t} - \alpha_h)(I_{k,t-j} - \alpha_k)] = 0, \quad \forall j = 1, \ldots, m, \ \forall h, k = 1, \ldots, K.$$

Hence, under the null all the autocorrelations from order 1 to the maximum lag length m in the hit sequences are zero. The authors propose using the multivariate portmanteau statistic of Li and McLeod (1981), which is a multivariate extension of the Box and Pierce (1970) test. The elements of the hits covariance matrix at the lag j can be estimated by

$$\hat{\gamma}_j^{hk} = \frac{1}{T-j} \sum_{t=j+1}^{T} (I_{h,t} - \alpha_h)(I_{k,t-j} - \alpha_k).$$

The test statistic is

$$Q_m = T(T+2)\sum_{j=1}^{m}\frac{1}{T-j}\text{vec}(\boldsymbol{R}_j)'\left(\boldsymbol{R}_0^{-1}\otimes\boldsymbol{R}_0^{-1}\right)\text{vec}(\boldsymbol{R}_j)$$

where $\boldsymbol{R}_j$ is the cross-correlation matrix whose element of position (h,k) is

$$R_j^{hk} = \frac{\hat{\gamma}_j^{hk}}{\sqrt{\hat{\gamma}_0^{hh}\hat{\gamma}_0^{kk}}} \quad h,k = 1,\ldots,K.$$

Li and McLeod (1981) show that the test statistic is asymptotically chi-square with mK^2 degrees of freedom. Hurlin and Tokpavi (2006) suggest selecting the lag length $m \in \{1,2,3,4,5\}$. This choice is motivated by a simulation study in which the distance between the observed and the theoretical chi-square distribution is evaluated by means of the Kolmogorov-Smirnov test.

A drawback of the Hurlin and Tokpavi (2006) test is that Q_m cannot be calculated if the matrix $\boldsymbol{R}_0$ is singular. This is likely to happen with coverage rates that are very close to each other. Indeed, in this case, it is more likely the $\boldsymbol{R}_0$ matrix will have several identical columns. This happens especially in small samples, because they are characterized by the same occurrences of violations at 1% and at 1.5%, say.

2.1 New multilevel tests

In this paper, we propose two novel multilevel testing procedures designed to test the conditional coverage hypothesis.

2.1.1 Markov tests

The first test we propose is a generalization of the Christoffersen (1998) independence test to the multilevel case.

Consider the following transition matrix

$$\mathbf{\Pi} = [\pi_{i,j}]_{i,j=0,\ldots,K}\,, \tag{9}$$

where

$$\pi_{i,j} = P(J_{j,t+1} = 1 | J_{i,t} = 1).$$

Under the null hypothesis of independence, all rows in the matrix $\mathbf{\Pi}$ are the same, i.e.

$$H_{0,ind} : \pi_{0,j} = \pi_{1,j} = \ldots = \pi_{K,j}, \quad \text{for } j = 0, \ldots, K-1. \tag{10}$$

The intuition is that the return has equal probability of being in interval j in period $t+1$, regardless of which of the $K+1$ intervals the return lies in period t.

Note that in the above formulation the column index j runs from 0 to $K-1$ given that $\pi_{i,K} = 1 - \sum_{j=0}^{K-1} \pi_{i,j}$ for each $i = 0, \ldots, K$.

If we assume that transitions are described by matrix (9), the log-likelihood is

$$\ell(\mathbf{\Pi}) = \sum_{\substack{0 \le i \le K \\ 0 \le j \le K}} T_{i,j} \ln(\pi_{i,j}). \tag{11}$$

where $T_{i,j}$ denotes the number of observations in the sample of N_t values with a j following an i, with $i, j = 0, \ldots, K$.

The null hypothesis in (10) can be tested using a likelihood ratio test, that we construct as

follows. Denoting by $\ell_{0,ind}$ and by $\ell_{1,ind}$ the log-likelihoods under the null and the alternative hypotheses, respectively, the test statistic is

$$LR_{ind} = 2\left(\ell_{1,ind} - \ell_{0,ind}\right) = 2\left(\ell(\widehat{\mathbf{\Pi}}) - \ell(\widehat{\boldsymbol{\pi}})\right) = 2\left(\sum_{\substack{0\leq i\leq K\\ 0\leq j\leq K}} T_{i,j}\ln(\widehat{\pi}_{i,j}) - \sum_{i=0}^{K} T_i \ln(\widehat{\pi}_i)\right) \quad (12)$$

where $\widehat{\pi}_i = \frac{T_i}{T}$ is, as already mentioned, the maximum likelihood estimator of the i-th component of $\boldsymbol{\pi}$, and $\widehat{\pi}_{i,j} = \frac{T_{i,j}}{T_i}$ is the maximum likelihood estimator of the (i,j)-element of matrix $\mathbf{\Pi}$ (see the Appendix for the derivation of this result). Under (10) the test statistic is asymptotically chi-square with K^2 degrees of freedom, given that there are K^2+K free parameters under the alternative and K free parameters under the null hypothesis.

We now turn to the conditional coverage test. Note that conditional coverage and independence are not the same concept given that conditional coverage entails testing also whether the average number of violations is correct. In this case, under the null hypothesis all rows in the matrix $\mathbf{\Pi}$ are the same and equal to the vector $(\theta_0, \ldots, \theta_K)$, i.e.

$$H_{0,cc} : \pi_{0,j} = \pi_{1,j} = \ldots = \pi_{K,j} = \theta_j, \quad \text{for } j = 0, \ldots, K-1. \quad (13)$$

Denoting by $\ell_{0,cc}$ and by $\ell_{1,cc}$ the log-likelihoods under the null and the alternative hypotheses respectively, the likelihood ratio test statistic is

$$LR_{cc} = 2\left(\ell_{1,cc} - \ell_{0,cc}\right) = 2\left(\ell(\widehat{\mathbf{\Pi}}) - \ell(\boldsymbol{\theta})\right) = 2\left(\sum_{\substack{0\leq i\leq K\\ 0\leq j\leq K}} T_{i,j}\ln(\widehat{\pi}_{i,j}) - \sum_{i=0}^{K} T_i \ln(\theta_i)\right) \quad (14)$$

where $\widehat{\pi}_i$ and $\widehat{\pi}_{i,j}$ are as before. The test statistic is asymptotically chi-square with K^2+K

degrees of freedom. Note that

$$LR_{cc} = LR_{uc} + LR_{ind},$$

since $\ell_{1,cc} = \ell_{1,ind}$, $\ell_{0,cc} = \ell_{0,cc}$ and $\ell_{1,uc} = \ell_{0,ind}$.

The feasible versions of (12) and (14) require replacing (11) and (7) with

$$\ell(\mathbf{\Pi}) = \sum_{\substack{i,j \in \{0,\ldots,K\}: \\ T_{i,j}>0}} T_{i,j} \ln(\pi_{i,j}) \quad \text{and} \quad \ell(\boldsymbol{\pi}) = \sum_{\substack{i \in \{0,\ldots,K\}: \\ T_i>0}} T_i \ln(\pi_i)$$

respectively, because in empirical applications we can have cases where $T_{i,j} = 0$ or even $T_i = 0$ for some i and j.

2.1.2 Pearson's χ^2 tests

The Markov test is powerful only against the first-order Markov alternative. In this section we propose a new test powerful against more general alternatives.

Consider the bivariate distribution

$$p_{N_t,N_{t-j}}(x,y) = P(N_t = x, N_{t-j} = y).$$

Under the null of the conditional coverage test, it holds that

$$p_{N_t,N_{t-j}}(x,y) = P(N_t = x)P(N_{t-j} = y) = \theta_x\theta_y \quad \forall x,y.$$

Denote by $T^{(j)}_{x,y}$ the number of observations in the sample for which $N_t = x$ and $N_{t-j} = y$. Note that $T^{(1)}_{x,y}$ coincides with $T_{x,y}$ of Section 2.1.1.

The proposed test statistic for a sample of T observations is

$$X_m = \sum_{j=1}^{m} X^{(j)}, \tag{15}$$

where

$$X^{(j)} = \sum_{x,y} \frac{(T_{x,y}^{(j)} - (T-j)\theta_x\theta_y)^2}{(T-j)\theta_x\theta_y}. \tag{16}$$

The test is designed to detect whether the average number of violations at the different rates is correct and to check for the independence in N_t with respect to its lags up to m [2]. While the asymptotic distribution of (16) is chi-square, the distribution of the test statistic (15) is not standard even for large samples because it is the sum of dependent chi-square random variables[3]. In order to calculate critical values, we use the Monte Carlo testing technique of Dufour (2006) which consists of the following three steps:

Step 1: generate under the null M time series of i.i.d. variables N_t each of length T;

Step 2: for each replica $j = 1, \ldots, M$, calculate the test statistic (15) whose value we denote by $X_{m,j}$;

Step 3: compute the p-value as

$$\hat{p}_M(X_{m,0}) = \frac{M \times \hat{G}_M(X_{m,0}) + 1}{M+1} \tag{17}$$

where $\hat{G}_M(X_{m,0}) = \frac{1}{M}\sum_{j=1}^{M} I(X_{m,0} < X_{m,j})$, $I(\cdot)$ is the indicator function, and $X_{m,0}$ is the

[2] We provide some guidance for the choice of m in Section 3 where we report the results of the Monte Carlo exercise.

[3] An alternative formulation to (16) is based on the likelihood ratio test statistic

$$X^{(j)} = 2\sum_{x,y} T_{x,y}^{(j)} \log\left(\frac{T_{x,y}^{(j)}/(T-j)}{\theta_x\theta_y}\right).$$

test statistic calculated from the original sample.

The test statistic, however, can only take on a countable number of distinct values. Consequently, the test value obtained from the sample, $X_{m,0}$, could coincide with some of the values obtained from simulating under the null hypothesis. The following tie-breaking procedure is used in these cases: for each test statistic, $X_{m,j}$, $j = 0, \ldots, M$, we draw an independent standard uniform random variate, U_j. The Monte Carlo p-value we calculated, $\tilde{p}_M(X_{m,0})$, is obtained replacing in (17) $\hat{G}_M(X_{m,0})$ with

$$\tilde{G}_M(X_{m,0}) = 1 - \frac{1}{M}\sum_{j=1}^{M} I(X_{m,0} \geq X_{m,j}) + \frac{1}{M}\sum_{j=1}^{M} I(X_{m,0} = X_{m,j}) \times I(U_0 \leq U_j).$$

Note that under the null, the hit sequence is generated by i.i.d. variables N_t, with distribution completely described by the K probability levels $\alpha_1, \alpha_2, \ldots, \alpha_K$. Hence, we do not have nuisance parameters under the null hypothesis. The validity of the above procedure is confirmed by Proposition 2.4 of Dufour (2006), which shows that, under the above construction it holds that

$$P\left(\tilde{p}_M(X_{m,0}) \leq p\right) = \frac{[p(M+1)]}{M+1} \quad \text{for } p \in [0,1],$$

where $[x]$ is the largest integer less than or equal to x.

One advantage of using Monte Carlo testing instead of bootstrap procedures is that the former guarantees consistency even when some parameters are on the boundary of the parameter space.

2.2 Numerical example

In this section, we report a numerical exercise in which we show that our test procedures, based on (14) and (15), perform better than the test of Perignon and Smith (2008) which, of course, is not designed to detect clusters in VaR violations. Let us consider three probability levels, $\alpha_1 = 5\%$, $\alpha_2 = 2.5\%$, and $\alpha_3 = 1\%$ and assume that, in a sample of dimension 500, VaR(1%) is violated 8 times, VaR(2.5%) 11 times, and VaR(5%) 21 times. Regarding the p-values for the Kupiec tests, the only difference with the previous example is for the 2.5% level for which we now find p-value(2.5%) =0.7129. For the multilevel test, where $T_0 = 479$, $T_1 = 10$ $T_2 = 3$, and $T_3 = 8$, we find that $LR_{uc} = 5.5930$, with a p-value of 0.1332 which clearly means the null is not rejected. Suppose, however, that some sort of clustering is present in the violation sequence. For instance, suppose that the first 8 returns are less than $-$VaR(1%), the subsequent 3 returns fall between $-$VaR(1%) and $-$VaR(2.5%), the subsequent 10 returns fall between $-$VaR(2.5%) and $-$VaR(5%), and finally the remaining 479 returns are larger than $-$VaR(5%). Of course the value of the Perignon and Smith (2008) test does not change. Using the first of the two new tests proposed in this paper, i.e. the Markov test described in section 2.1.1, we find $LR_{cc} = 203.45$, with a p-value of the order of 10^{-6}. If instead we apply the the second of the two new tests, i.e. Pearson's χ^2 tests of section 2.1.2, we find $X_1 = 1301.84$, $X_5 = 4242.97$ and $X_{10} = 6251.19$. Again, in all the three cases the p-value is of the order of 10^{-6}, which strongly rejects the null. For completeness, the values of the Hurlin and Tokpavi tests are $Q_1 = 1095.28$, $Q_5 = 13925.44$, and $Q_{10} = 6297.31$. In all the three cases the null is strongly rejected. □

The above numerical example is just a simple illustration of the limits of both the unilevel testing procedures and of the multilevel unconditional coverage test. A robust comparison requires instead a comprehensive Monte Carlo exercise, which we report in the next section.

3 A Monte Carlo evaluation of the testing procedures

In this section, we study the performance of the multilevel tests proposed in this paper via a Monte Carlo exercise.

3.1 Monte Carlo design

A short description of the Monte Carlo design follows. First, 10000 time series suitable to describe financial returns are generated according to the following GARCH(1,1) model with Student-t innovations:

$$h_{t+1} = \omega + \alpha e_t^2 + \beta h_t, \tag{18}$$

with $e_t = \sqrt{h_t} u_t$, $u_t \sim t(6.5)$, i.e. the distribution of the innovation is Student-t with 6.5 degrees of freedom. We use the same parameters as in the Monte Carlo experiments of Perignon and Smith (2008), i.e. $\hat{\omega} = 0.05$, $\hat{\alpha} = 0.05$ and $\hat{\beta} = 0.9$.

In order to capture excess skewness and kurtosis, important features of financial data especially in a period of financial turmoil, we extend our Monte Carlo analysis to the case of returns generated by a GARCH(1,1) with skew-t (see Hansen, 1994) and GED distributions (see Box and Tiao, 1992). In both cases the parameters used in the simulations are estimated from the daily S&P 500 returns over the period March 2008 to March 2012 (1000 observations).

For the GARCH(1,1) specification (18), the values of the parameters are $\hat{\omega} = 1.86 \times 10^{-6}$, $\hat{\alpha} = 0.1051$ and $\hat{\beta} = 0.8924$ in the skew-t case and $\hat{\omega} = 2.17 \times 10^{-6}$, $\hat{\alpha} = 0.1109$ and $\hat{\beta} = 0.8835$ in the GED case. Further, for the skew-t specification, the degrees of freedom are 7.2760 and the asymmetry parameter is -0.1939. For the GED distribution, the shape parameter equals 1.4021.

For each sample size $T \in \{250, 500, 1000, 2500\}$ and for each sample path, we generate 250 additional returns. Next, we compute for each sample path T out-of-sample VaR estimates based on a rolling window of length 250. VaR estimates are computed using four different models: Normal, Historical Simulation (HS), Hybrid Historical Simulation (HHS, see Boudoukh et al. 1998), and RiskMetrics (RM) with decay factor $\lambda = 0.94$. In addition to the existing multilevel tests of Perignon and Smith (PS), and Hurlin and Tokpavi (Q_m), we evaluate the performance of the novel tests proposed in the present paper, namely the Markov and the Pearson (X_m) tests. For both the Hurlin and Tokpavi and the Pearson test we choose the lag length $m \in \{1, 5, 10\}$. We compare the above multilevel tests in the case of $\alpha_1 = 5\%$, $\alpha_2 = 2.5\%$ and $\alpha_3 = 1\%$. For all tests, we use simulated critical values, based on $M = 50000$ simulations, in order to avoid small sample distortions when calculating the rejection frequencies. This is important for both the Pearson test that has a non-standard distribution and the remaining tests that have an asymptotic chi-square distribution. Hence, we report size-adjusted rejection frequencies.

3.2 Results

Tables 1–3 report, for each of the testing procedures, the proportion of times (rejection frequency) a test rejects the null that the VaR model is 'appropriate'. Since returns are generated from a GARCH model and quantiles are estimated according to a different model (Normal, HS, HHS and RM models), the higher the rejection frequency, the better is the associated test.

Table 1 reports the results for the case when returns are generated according to a GARCH model with Student-t innovations. For the Normal VaR the proposed Pearson tests based on $m = 5$ and $m = 10$ outperform both the Perignon and Smith and Hurlin and Tokpavi

test across all sample sizes. It is worth stating that the Pearson X_1 test shows the same properties of the Perignon and Smith test. The performance of the Markov test introduced in the paper is something in between the Pearson and the Perignon and Smith tests, but better than the Hurlin and Tokpavi test. The results for the Normal VaR are confirmed by those in Panel C (HHS VaR) and Panel D (RM VaR). For the case of HS VaR (Panel B) when $m = 1$ the Pearson test is better than the corresponding Q test. For $m = 5$ and $m = 10$ the Pearson tests outperform the corresponding Q tests in small sample.

Tables 2 and 3 report the size-adjusted powers of the different test for the case when returns are generated according to a GARCH model with Skew Student-t and with GED innovations, respectively.

If we compare the results in Tables 2–3 with those in Table 1, we notice a substantial increase in power mainly due to the fact that the underlying distributions (skew-t and GED) make the misspecification in VaR modelling easier to be captured by the tests.

For the case where the Normal distribution is used for computing VaR, i.e. Panel A of Tables 2 and 3, the Pearson tests we propose in this paper, based on $m = 5$ and $m = 10$, outperform both the Perignon and Smith and Hurlin and Tokpavi tests across all sample sizes. When $T = 2500$ all tests but the Hurlin and Tokpavi show the same performance. As in Table 1, the results of Panels C and D (HHS and RM VaR, respectively) of both tables are in line with those of Panel A. In Panel B, instead, the best performance is achieved by the Q_{10} statistic which is marginally better for T larger than 500, while for $T = 2500$ all tests have the same performance. It is worth noticing that the Pearson tests are always better in terms of power than all the other tests when $T = 250$, which is the most common sample size used in practice. Finally, from the Monte Carlo experiments we can draw useful information for the choice of the lag length m providing guidelines to practitioners. Indeed, especially

Table 1: Size-adjusted rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. PS denotes the Perignon and Smith (2008) test, Q_m, $m \in \{1, 5, 10\}$, is the Hurlin and Tokpavi (2006) test, Markov is the test (14) and X_m, $m \in \{1, 5, 10\}$, is the Pearson test based on (15). Returns are generated according to a GARCH model with Student-t innovations.

Panel A: Normal VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.163	0.020	0.007	0.002	0.147	0.177	0.209	0.215
$T = 500$	0.226	0.086	0.140	0.167	0.200	0.228	0.303	0.307
$T = 1000$	0.417	0.135	0.246	0.305	0.345	0.381	0.517	0.528
$T = 2500$	0.834	0.202	0.399	0.509	0.736	0.818	0.902	0.900

Panel B: HS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.058	0.016	0.004	0.001	0.097	0.136	0.156	0.157
$T = 500$	0.025	0.083	0.151	0.181	0.090	0.132	0.154	0.142
$T = 1000$	0.022	0.140	0.264	0.327	0.103	0.161	0.175	0.159
$T = 2500$	0.032	0.210	0.423	0.537	0.145	0.240	0.284	0.244

Panel C: HHS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.101	0.005	0.000	0.000	0.079	0.127	0.152	0.160
$T = 500$	0.282	0.018	0.009	0.006	0.172	0.238	0.376	0.413
$T = 1000$	0.740	0.031	0.017	0.011	0.526	0.632	0.816	0.846
$T = 2500$	0.999	0.049	0.039	0.035	0.995	0.999	1.000	1.000

Panel D: RM VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.116	0.010	0.002	0.001	0.099	0.119	0.138	0.138
$T = 500$	0.222	0.037	0.037	0.032	0.156	0.186	0.252	0.268
$T = 1000$	0.435	0.049	0.052	0.045	0.304	0.366	0.499	0.511
$T = 2500$	0.884	0.058	0.062	0.060	0.747	0.842	0.914	0.919

Table 2: Size-adjusted rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. PS denotes the Perignon and Smith (2008) test, Q_m, $m \in \{1, 5, 10\}$, is the Hurlin and Tokpavi (2006) test, Markov is the test (14) and X_m, $m \in \{1, 5, 10\}$, is the Pearson test based on (15). Returns are generated according to a GARCH model with Skew Student-t innovations.

Panel A: Normal VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.550	0.047	0.041	0.029	0.532	0.499	0.559	0.563
$T = 500$	0.738	0.184	0.438	0.557	0.717	0.739	0.814	0.824
$T = 1000$	0.940	0.394	0.775	0.880	0.931	0.950	0.981	0.983
$T = 2500$	1.000	0.749	0.984	0.996	1.000	1.000	1.000	1.000

Panel B: HS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.341	0.061	0.052	0.033	0.388	0.354	0.413	0.412
$T = 500$	0.308	0.241	0.523	0.648	0.443	0.487	0.579	0.585
$T = 1000$	0.327	0.447	0.816	0.909	0.563	0.654	0.783	0.787
$T = 2500$	0.614	0.750	0.985	0.997	0.815	0.916	0.978	0.977

Panel C: HHS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.102	0.061	0.052	0.033	0.088	0.142	0.187	0.201
$T = 500$	0.241	0.241	0.523	0.648	0.164	0.236	0.374	0.393
$T = 1000$	0.608	0.447	0.816	0.909	0.438	0.539	0.739	0.768
$T = 2500$	0.991	0.750	0.985	0.997	0.967	0.986	0.996	0.997

Panel D: RM VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.300	0.009	0.002	0.000	0.296	0.364	0.482	0.499
$T = 500$	0.596	0.044	0.070	0.070	0.540	0.634	0.774	0.791
$T = 1000$	0.918	0.090	0.136	0.160	0.865	0.921	0.976	0.978
$T = 2500$	1.000	0.149	0.290	0.352	0.999	1.000	1.000	1.000

Table 3: Size-adjusted rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. PS denotes the Perignon and Smith (2008) test, Q_m, $m \in \{1, 5, 10\}$, is the Hurlin and Tokpavi (2006) test, Markov is the test (14) and X_m, $m \in \{1, 5, 10\}$, is the Pearson test based on (15). Returns are generated according to a GARCH model with GED innovations.

Panel A: Normal VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.481	0.064	0.056	0.040	0.465	0.432	0.491	0.492
$T = 500$	0.601	0.220	0.497	0.622	0.618	0.631	0.723	0.731
$T = 1000$	0.823	0.429	0.801	0.899	0.844	0.870	0.940	0.946
$T = 2500$	0.995	0.761	0.988	0.998	0.992	0.998	1.000	1.000

Panel B: HS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.328	0.063	0.057	0.038	0.381	0.357	0.419	0.424
$T = 500$	0.287	0.255	0.552	0.672	0.444	0.490	0.584	0.590
$T = 1000$	0.325	0.460	0.831	0.923	0.571	0.668	0.800	0.801
$T = 2500$	0.636	0.779	0.990	0.999	0.838	0.933	0.983	0.983

Panel C: HHS VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.115	0.007	0.001	0.000	0.104	0.169	0.233	0.247
$T = 500$	0.271	0.029	0.031	0.032	0.185	0.268	0.431	0.456
$T = 1000$	0.660	0.063	0.077	0.078	0.505	0.611	0.796	0.824
$T = 2500$	0.995	0.118	0.198	0.241	0.981	0.993	0.999	0.999

Panel D: RM VaR								
	PS	Q_1	Q_5	Q_{10}	Markov	X_1	X_5	X_{10}
$T = 250$	0.144	0.012	0.004	0.001	0.167	0.224	0.286	0.291
$T = 500$	0.239	0.066	0.098	0.109	0.242	0.322	0.430	0.438
$T = 1000$	0.480	0.109	0.182	0.211	0.437	0.548	0.707	0.712
$T = 2500$	0.919	0.169	0.333	0.413	0.830	0.931	0.973	0.974

in Tables 2–3, we notice an increase in power moving from $m = 1$ to $m = 5$, whereas the increase in power moving from $m = 5$ to $m = 10$ is only marginal. Thus we suggest using $m = 5$ in practical applications.

4 Empirical application

In this section, we use the full set of multilevel tests for a comprehensive backtesting exercise. The aim is to illustrate the implementation of the different tests, and to show how they can deliver contrasting conclusions.

We use daily returns on 15 MSCI world indices traded as iShares on the American Exchange for the following countries: US, Mexico, Canada, the UK, the Switzerland, Sweden, Spain, Italy, Germany, France, Australia, Singapore, Japan, Hong Kong and Malaysia. The exchange traded funds (ETF) do not suffer from problems related to non-synchronicity and different closing times, unlike the raw indices which trade at different times around the world. The data covers the period December 2000 to November 2011 (2750 observations).

4.1 Univariate models

First, we consider, in an out-of-sample exercise, the following univariate VaR models: HS, RM, GARCH and GARCH-t (both with and without an AR(1) model for the mean), HHS, Filtered Historical Simulation (FHS), and GJR and GJR-t (both with and without an AR(1) model for the mean). The order of all the GARCH and GJR models is (1,1). For all the VaR models, with the exception of HS, we use a rolling window of 250 observations, for both parameters estimation and VaR calculation. Consequently, we end up with 2500 VaR estimates for the period December 2001 to November 2011. In the HS case, instead, we use

an expanding window with the first one comprising the first 250 observations, obtaining the same number of VaR estimates and relative to the same period as the alternative methods.

In Table 4, we report some representative results for a limited number of countries, i.e. Australia, Germany, USA, and Singapore [4]. The table reports the p-values for the multilevel tests based on coverage probabilities 5%, 2.5%, and 1%. For Australia, the Hurlin and Tokpavi tests do not reject the null in all cases but HS, GARCH and GARCH-t in contrast to the finding from the Pearson tests that systematically reject the null. For Singapore, there is evidence of non-contradiction between tests for the case of AR(1)-GARCH-t and AR(1)-GJR-t only; in all the other cases the Q_1, Q_5 and Q_{10} accept the null whereas all other tests lead to a rejection. The same applies for USA, where however the two models dominate but in a weaker form than the Singapore case. In addition for USA, the Q_1 test systematically accepts the null with the only exception of the HHS and GJR models. For Germany, once again all tests do not reject the null only in the AR(1)-GJR-t case. The HHS model is strongly rejected by all tests with the exeption of the Hurlin and Tokpavi tests. With respect to the results we do not report in the paper, we mention that GARCH-t models perform better than GARCH models with normal innovations and the inclusion of the mean component (AR(1)-GARCH, AR(1)-GARCH-t, AR(1)-GJR and AR(1)-GJR-t models) generally improves upon GARCH models without the mean.

4.2 Multivariate models

Next, we estimate the VaR measures in a Multivariate GARCH (MGARCH) context for an equally weighted portfolio comprising the 15 securities. The motivation for including such an analysis is that modeling the dependence among returns should improve VaR forecasting.

[4]The full set of results is available from the authors upon request.

Table 4: Backtesting Results. The table reports the calculated p-values of the different multilevel tests with vector of probability levels $(1\%, 2.5\%, 5\%)$ under different VaR models.

		HS	RM	GARCH	GARCH-t	AR(1)-GARCH	AR(1)-GARCH-t	HHS	FHS	GJR	GJR-t	AR(1)-GJR	AR(1)-GJR-t
	PS	0.188	0.002	0.000	0.000	0.000	0.098	0.000	0.276	0.000	0.000	0.000	0.021
	Q_1	0.004	0.039	0.032	0.035	0.421	0.256	0.707	0.287	0.361	0.071	0.304	0.646
	Q_5	0.000	0.041	0.015	0.005	0.143	0.122	0.489	0.073	0.713	0.167	0.488	0.628
	Q_{10}	0.000	0.069	0.005	0.005	0.053	0.035	0.317	0.011	0.766	0.270	0.578	0.432
Australia -EWA	Markov	0.004	0.005	0.000	0.000	0.000	0.096	0.000	0.421	0.000	0.000	0.000	0.077
	X_1	0.001	0.003	0.000	0.000	0.000	0.124	0.002	0.162	0.000	0.000	0.000	0.040
	X_5	0.000	0.001	0.000	0.000	0.000	0.035	0.000	0.054	0.000	0.000	0.000	0.009
	X_{10}	0.000	0.001	0.000	0.000	0.000	0.029	0.000	0.039	0.000	0.000	0.000	0.006
	PS	0.096	0.070	0.420	0.400	0.114	0.955	0.000	0.000	0.019	0.064	0.008	0.190
	Q_1	0.001	0.038	0.087	0.079	0.181	0.083	0.422	0.013	0.005	0.011	0.152	0.256
	Q_5	0.000	0.002	0.004	0.001	0.002	0.000	0.484	0.000	0.007	0.023	0.293	0.474
	Q_{10}	0.000	0.002	0.042	0.007	0.010	0.002	0.606	0.000	0.053	0.135	0.727	0.801
Germany -EWG	Markov	0.000	0.019	0.153	0.138	0.090	0.228	0.000	0.000	0.001	0.004	0.008	0.113
	X_1	0.000	0.004	0.150	0.140	0.077	0.305	0.000	0.000	0.004	0.013	0.021	0.222
	X_5	0.000	0.002	0.064	0.034	0.011	0.072	0.000	0.000	0.006	0.026	0.009	0.219
	X_{10}	0.000	0.004	0.192	0.120	0.035	0.281	0.000	0.000	0.013	0.052	0.014	0.274
	PS	0.097	0.012	0.001	0.001	0.004	0.899	0.000	0.030	0.000	0.000	0.001	0.807
	Q_1	0.243	0.688	0.731	0.725	0.676	0.531	0.381	0.643	0.825	0.813	0.638	0.699
	Q_5	0.000	0.289	0.469	0.425	0.324	0.070	0.413	0.578	0.937	0.935	0.986	0.865
	Q_{10}	0.000	0.335	0.675	0.569	0.326	0.059	0.453	0.106	0.949	0.924	0.761	0.709
Singapore -EWS	Markov	0.065	0.039	0.006	0.008	0.016	0.322	0.000	0.052	0.003	0.004	0.003	0.537
	X_1	0.104	0.051	0.012	0.011	0.029	0.831	0.000	0.076	0.005	0.006	0.007	0.851
	X_5	0.000	0.005	0.002	0.002	0.006	0.396	0.000	0.017	0.001	0.002	0.002	0.896
	X_{10}	0.000	0.005	0.002	0.002	0.004	0.498	0.000	0.006	0.001	0.001	0.001	0.771
	PS	0.033	0.001	0.016	0.056	0.027	0.288	0.000	0.003	0.000	0.001	0.001	0.016
	Q_1	0.003	0.153	0.257	0.074	0.411	0.901	0.037	0.288	0.205	0.368	0.913	0.856
	Q_5	0.000	0.075	0.041	0.019	0.069	0.149	0.047	0.026	0.479	0.442	0.774	0.852
	Q_{10}	0.000	0.020	0.012	0.010	0.014	0.060	0.104	0.004	0.708	0.571	0.775	0.689
USA -SPY	Markov	0.000	0.001	0.057	0.032	0.069	0.652	0.000	0.031	0.000	0.003	0.006	0.078
	X_1	0.000	0.001	0.016	0.021	0.042	0.506	0.000	0.004	0.001	0.003	0.003	0.052
	X_5	0.000	0.000	0.002	0.005	0.002	0.077	0.000	0.000	0.000	0.000	0.000	0.014
	X_{10}	0.000	0.000	0.001	0.004	0.001	0.061	0.000	0.000	0.000	0.000	0.000	0.008

In all cases we estimate an Asymmetric Dynamic Conditional Correlation (ADCC) model under the multivariate normal assumption (see Cappiello et al., 2006).

In Panel A of Table 5, we report the results for the equally-weighted portfolio of the 15 countries using the univariate VaR models of Tables 4. There is evidence that the PS test favors the AR(1)-GARCH-t model while all other tests confirm the dominance of the AR(1)-GJR-t model. The HS model is consistently rejected by all the test employed. The RM and FHS models are rejected by all the testing procedures with the exception of the Markov and Hurlin and Tokpavi tests. Finally, the HHS model is always rejected with the exception of the Markov, X_1 and Hurlin and Tokpavi tests. Again, the Q_1 test does not reject any model.

Panel B of Table 5 reports the results for Gaussian MGARCH models. We consider four different models for the first stage, i.e. GJR with normal (GJR-N) or Student-t innovations (GJR-t) with and without an AR(1) process for the mean. In all the four cases we fit a GJR(1,1) model in the first stage and an ADCC(1,1,1) in the second stage. Contrary to the univariate results, the inclusion of the mean makes the results worse. The GJR-t model dominates all the other models because it is only rejected by the X_{10} test at the 5% confidence level, but not at the the 1% confidence level. Contrary to all the other testing procedures, the Hurlin and Tokpavi tests are in favour of all the four multivariate compared in the exercise.

5 Concluding remarks

In this paper, we proposed novel independence and conditional coverage tests in a multilevel setup, able to overcome the reduced power of the standard unilevel testing procedures in

Table 5: Backtesting Results for an equally weighted portfolio of the 15 securities. The table reports the calculated p-values of the different multilevel tests with vector of probability levels $(1\%, 2.5\%, 5\%)$ under different VaR models.

Panel A: Univariate Models												
	HS	RM	GARCH	GARCH-t	AR(1)-GARCH	AR(1)-GARCH-t	HHS	FHS	GJR	GJR-t	AR(1)-GJR	AR(1)-GJR-t
PS	0.003	0.015	0.015	0.140	0.014	0.543	0.026	0.016	0.002	0.039	0.002	0.249
Q_1	0.015	0.626	0.956	0.389	0.870	0.093	0.785	0.346	0.811	0.799	0.706	0.713
Q_5	0.000	0.392	0.167	0.229	0.329	0.030	0.212	0.164	0.642	0.749	0.918	0.878
Q_{10}	0.000	0.217	0.006	0.020	0.034	0.000	0.164	0.071	0.296	0.402	0.613	0.550
Markov	0.000	0.110	0.209	0.284	0.130	0.282	0.128	0.061	0.023	0.198	0.010	0.332
X_1	0.000	0.044	0.083	0.089	0.063	0.159	0.107	0.021	0.010	0.108	0.007	0.347
X_5	0.000	0.005	0.003	0.040	0.003	0.149	0.023	0.003	0.001	0.039	0.002	0.261
X_{10}	0.000	0.003	0.001	0.021	0.002	0.071	0.026	0.003	0.001	0.023	0.001	0.145

Panel B: Gaussian MGARCH Models				
	GJR N-ADCC	GJR t-ADCC	AR-GJR N-ADCC	AR-GJR t-ADCC
PS	0.024	0.186	0.012	0.038
Q_1	0.834	0.746	0.491	0.632
Q_5	0.524	0.165	0.451	0.650
Q_{10}	0.119	0.048	0.111	0.039
Markov	0.185	0.490	0.021	0.109
X_1	0.083	0.310	0.026	0.079
X_5	0.018	0.058	0.004	0.023
X_{10}	0.008	0.044	0.002	0.005

presence of small samples. In risk management analysis, it is practice to estimate quantiles for two or more different probability levels. To this purpose, using a multilevel test is intuitively more efficient, and statistically more powerful, than to use separate unilevel tests. Moreover, multilevel tests are particularly useful because they make the best use of the limited amount of information regarding the return distribution made available by banks or financial institutions in general to assess their risk exposure.

The first test we proposed is a generalization to the multilevel case of the Markov test of Christoffersen (1998), while the second test is a Pearson-type of test based on the joint distribution of the total number of VaR violations in a period and its lags. In an extensive Monte Carlo exercise, where returns were generated under alternative GARCH models with skewed and leptokurtic innovations (i.e. Student-t, skew-t and GED innovations), and where VaR were estimated using models commonly used in practice (i.e. Normal, HS, HHS and RM), the multilevel tests we proposed showed higher power than both the multilevel unconditional test of Perignon and Smith (2008) and the multilevel conditional tests of Hurlin and Tokpavi (2006). Via an empirical application using daily returns on 15 MSCI world indices, we implemented all available multilevel tests and we showed that in most cases different tests deliver different conclusions.

Appendix: Maximum likelihood estimators of π_i and $\pi_{i,j}$

Let us write the log-likelihood function (7) as

$$\ell(\boldsymbol{\pi}) = \sum_{i=0}^{K} T_i \ln(\pi_i) = T_j \ln(\pi_j) + \sum_{i \neq j} T_i \ln(\pi_i) = T_j \ln\left(1 - \sum_{h \neq j} \pi_h\right) + \sum_{i \neq j} T_i \ln(\pi_i),$$

for some $j = 0, \ldots, K$.

Setting the derivative w.r.t. π_i, $i \neq j$, equal to zero yields

$$T_j \pi_i = T_i \left(1 - \sum_{h \neq j} \pi_h \right).$$

Since the quantity in brackets is equal to π_j, the first order conditions are

$$T_j \pi_i = T_i \pi_j, \quad i = 0, \ldots, K, i \neq j.$$

Summing both sides for every $i \neq j$ yields

$$T_j \sum_{i \neq j} \pi_i = \pi_j \sum_{i \neq j} T_i \qquad \Longrightarrow \qquad T_j (1 - \pi_j) = \pi_j (T - T_j) \qquad \Longrightarrow \qquad \widehat{\pi}_j = \frac{T_j}{T}.$$

Similarly, let us write (11) as

$$\ell(\mathbf{\Pi}) = \sum_{i=0}^{K} \left[T_{i,j} \ln(\pi_{i,j}) + \sum_{h \neq j} T_{i,h} \ln(\pi_{i,h}) \right] = \sum_{i=0}^{K} \left[T_{i,j} \ln \left(1 - \sum_{h \neq j} \pi_{i,h} \right) + \sum_{h \neq j} T_{i,h} \ln(\pi_{i,h}) \right].$$

Setting the derivative w.r.t. $\pi_{i,l}$, $i = 0, \ldots, K$, $l \neq j$, equal to zero yields

$$T_{i,j} \pi_{i,l} = T_{i,l} \left(1 - \sum_{h \neq j} \pi_{i,h} \right) = T_{i,l} \pi_{i,j},$$

since $1 - \sum_{h \neq j} \pi_{i,h} = \pi_{i,j}$.

Summing over $l = 0, \ldots, K$, $l \neq j$ yields

$$T_{i,j} \sum_{l \neq j} \pi_{i,l} = \pi_{i,j} \sum_{l \neq j} T_{i,l} \qquad \Longrightarrow \qquad T_{i,j} (1 - \pi_{i,j}) = \pi_{i,j} (T_i - T_{i,j}) \qquad \Longrightarrow \qquad \widehat{\pi}_{i,j} = \frac{T_{i,j}}{T_i}.$$

Acknowledgements

We wish to thank the Editor, Dick Van Dijk, and an anonymous Referee for very useful comments and suggestions which greatly helped to improve the content and the presentation of the paper. The usual disclaimer applies. This paper was completed while Arturo Leccadito and Simona Boffelli were visiting the Centre for Econometric Analysis of Cass Business School in January-July 2012. Financial support from the Centre for Econometric Analysis and the 2012 Cass Business School Pump-Priming grant scheme is gratefully acknowledged.

References

Artzner, P., F. Delbaen, and J.-M. E. D. Heath (1999). Coherent measures of risk. *Mathematical Finance 9*(3), 203–228.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics 19*(4), 465–74.

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). Evaluating Value-at-Risk models with desk-level data. *Management Science 57*(12), 2213–2227.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (1998). The best of both worlds: A hybrid approach to calculating Value at Risk. *Risk 11*, 64–67.

Box, G. E. P. and D. A. Pierce (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association 65*(332), 1509–1526.

Box, G. E. P. and G. C. Tiao (1992). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Candelon, B., G. Colletaz, C. Hurlin, and S. Tokpavi (2011). Backtesting value-at-risk: A GMM duration-based test. *Journal of Financial Econometrics 9*(2), 314–343.

Cappiello, L., R. F. Engle, and K. Sheppard (2006). Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics 4*(4), 537–572.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–862.

Christoffersen, P. F. and D. Pelletier (2004). Backtesting Value-at-Risk: A duration-based approach. *Journal of Financial Econometrics 2*(1), 84–108.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*(4), 863–883.

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.

Engle, R. F. and S. Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*(4), 367–381.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review 35*(3), 705–730.

Hurlin, C. and S. Tokpavi (2006). Backtesting value-at-risk accuracy: a simple new test. *Journal of Risk 9*(2), 19–37.

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

Li, W. K. and A. McLeod (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society, Series B 43*(2), 231–239.

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

Arturo Leccadito (Ph.D., Bergamo) is Assistant Professor of financial mathematics at the Department of Economics, Statistics and Finance of the University of Calabria, Italy. His research interests include financial econometrics, credit risk modelling and option pricing. He has published in Quantitave Finance, Econometric Reviews and International Journal of Theoretical and Applied Finance, and others.

Simona Boffelli (MSc., Bergamo) is a PhD candidate at the University of Bergamo, Italy. Her research interests are in financial econometrics, modelling risk and cross-market correlations, asset pricing, modelling jumps and cojumps and the impact of macrofactors, macronews, bond auctions and credit rating actions on European bond markets.

Giovanni Urga (Ph.D., Oxford) is Professor of Finance and Econometrics and Director of the Centre for Econometric Analysis (CEA@Cass) at Cass Business School, London, U.K., and professor of Econometrics at the University of Bergamo, Italy. His research interests are in panel data, financial econometrics, modelling risk and cross-market correlations, asset pricing, structural breaks, modelling common stochastic trends, and credit spreads. He has published in the Journal of Econometrics, Journal of Business and Economic Statistics, Economics Letters, Econometric Theory, Oxford Bulletin of Economics and Statistics, Journal of Applied Econometrics, International Journal of Forecasting, and others. He is an Associate Editor for Empirical Economics, and has been a guest editor for the Journal of Econometrics and the Journal of Business and Economic Statistics.