First proof • Typesetter: RH • 17 December 2006

Backtesting value-at-risk accuracy: a simple new test

Christophe Hurlin
LEO, University of Orléans, Rue de Blois, BP 6739, 45067 Orléans Cedex 2, France

Sessi Tokpavi*
LEO, University of Orléans, Rue de Blois, BP 6739, 45067 Orléans Cedex 2, France

This paper proposes a new test of value-at-risk (VAR) validation. Our test exploits the idea that the sequence of VAR violations (hit function) – taking value $1 - \alpha$, if there is a violation, and $-\alpha$ otherwise – for a nominal coverage rate α verifies the properties of a martingale difference if the model used to quantify risk is adequate (Berkowitz *et al.*, 2005). More precisely, we use the multivariate portmanteau statistic of Li and McLeod (1981) – extension to the multivariate framework of the test of Box and Pierce (1970) – to jointly test the absence of autocorrelation in the vector of hit sequences for various coverage rates considered as relevant for the management of extreme risks. We show that this shift to a multivariate dimension appreciably improves the power properties of the VAR validation test for reasonable sample sizes.

1 Introduction

The prudential rules of the Basle Accord allow banks to use their own models to help determine their regulatory capital requirements. The latter are related to the level of portfolio risk defined by value-at-risk (VAR), the standard measure of market risk. Many tools have been developed within the framework to assess – statistically or economically – the accuracy of banks' VAR models. Indeed, the Second Pillar in Basle II (Committee on Banking Supervision, 2004) gives special attention to the validation procedures of internal models of market risk.

There are a multitude of different methods of estimating VAR. VAR models[1] can be based on a non-parametric approach (hybrid methods, historical simulation, etc.), on a parametric approach (univariate or multivariate ARCH–GARCH models, RiskMetrics), or on a semi-parametric approach (extreme value theory, CAViaR etc.). Yet, practice generally shows that these various models lead to widely different VAR levels, for the same portfolio – see for example Beder (1995)

*Corresponding author.
We acknowledge helpful discussions and exchanges with participants at the 55th AFSE meeting in Paris, at the AEA conference on Exchange Rate and Risk Econometrics in Athens. We also thank the Editor-in-Chief, who helped to improve the quality of an earlier version of this paper, and Lionel Martinelli for his comments. Any remaining errors and inaccuracies are our own.
[1] See Jorion (2001), Engle and Manganelli (2001) and Dowd (2005) for reviews of the literature on different methods of VAR forecasting.

and Hendricks (1996). In this context, how can we evaluate statistically the precision of a VAR model?

One widely used approach to backtesting is the interval forecast approach due to Christoffersen (1998). The aim is to test the so-called "conditional coverage" hypothesis: this concerns the process of violations[2] for a given coverage rate, where a violation is the occurrence of a loss that exceeds the forecast VAR.

The conditional coverage property has two major implications, namely the unconditional coverage and independence hypotheses. The former simply means that VAR for a 5% coverage rate, for example, implies that the expected frequency of observed violations is also equal to 5%. The latter means that violations should be distributed independently. In other words, there must not be any clustering in violations: the occurrence of a loss exceeding VAR forecast does not contain information that enables one to forecast future violations.

The relevant literature suggests various tests of these two hypotheses. These include Christoffersen's test (1998), test based on the use of a Markov chain, the "hit regression" test of Engle and Manganelli (2004) based on a linear autoregressive model, and more recently the tests of Berkowitz *et al.* (2005) based on martingale difference or white noise properties.

A rather different approach is based on the whole density used to forecast VAR (Crnkovic and Drachman, 1997; Diebold *et al.*, 1998; Berkowitz, 2001). The objective is to test the whole density of profits and losses modeled *ex ante*, and so implicitly test the validity of VAR for all coverage rates between 0 and 1. However, a major weakness of this approach is that a model with superior density forecast may not necessarily meet the needs of risk managers who care much more about the tails.

In this paper, we propose a new VAR validation halfway between the "interval forecast" and "density forecast" approaches. The test looks at violations over multiple coverage rates. It is based on the white noise property of the process of violations for a given coverage rate, but also takes into account the joint validation of this hypothesis for a finite set of selected coverage rates. The statistic used is a multivariate extension of the autocorrelation test of Box and Pierce (1970) proposed by Li and McLeod (1981). This test allows us to assess a VAR model by testing the property of conditional coverage for violations processes associated with VARs specified at a variety of coverage rates (or confidence levels) and not just one. We show that the proposed test has very good power in small sample sizes, and has notably better power properties than the more traditional interval forecasts VAR validation tests.

Our paper is structured as follow: Section 2 presents the main VAR assessment tests that arise from each of the two approaches mentioned above. Section 3 presents our multivariate test statistic. Section 4 deals with the finite-sample properties of our test. The last section concludes and suggests some extensions to our test procedure.

[2] In the framework of interval forecast validation, a violation is an indicator variable which informs if the ex-post realization of the variable falls inside the predicted interval or not.

2 Existing approaches

Traditionally, the quality of the forecast of an economic variable is assessed by comparing its *ex-post* realization with the *ex-ante* forecast value. The assessment would be carried out using a statistical loss function such as the mean squared error (MSE) criterion. However, such criteria are easier to use when the *ex-post* realization of the forecast variable is observable. When the realization is unobservable, the validation exercise then requires the use of a proxy for the unobservable variable with good properties, such as low bias.[3] Since it is relatively difficult to compute such a proxy for the true unknown VAR, statistical VAR assessment criteria are generally based on tests of the two main predictions that follow from a good VAR model, namely the unconditional coverage hypothesis and the independence hypothesis.

Formally, let r_t denote the return of an asset or a portfolio at time t. Let $VAR_{t|t-1}(\alpha)$ be the *ex-ante* VAR for an α% coverage rate forecast conditionally on an information set Ω_{t-1}. If the model is adequate, then the following relation must hold:

$$\Pr\left[r_t < VAR_{t|t-1}(\alpha)\right] = \alpha \tag{1}$$

Let $I_t(\alpha)$ be the indicator variable associated with the *ex-post* observation of an α% VAR violation at time t, ie:

$$I_t(\alpha) = \begin{cases} 1 & \text{if } r_t < VAR_{t|t-1}(\alpha) \\ 0 & \text{else} \end{cases} \tag{2}$$

The problem of assessing a VAR model corresponds to the problem of assessing whether the violations sequence $\{I_t\}_{t=1}^{T}$ obeys the following two hypotheses:

- *The unconditional coverage hypothesis*: the probability of an *ex-post* loss exceeding VAR forecast must be equal to the α coverage rate:

$$\Pr[I_t(\alpha) = 1] = E[I_t(\alpha)] = \alpha \tag{3}$$

- *The independence hypothesis*: VAR violations observed at two different dates for the same coverage rate must be distributed independently. Formally, the variable $I_t(\alpha)$ associated with a VAR violation at time t for an α% coverage rate should be independent of the variable $I_{t-k}(\alpha)$, $\forall k \neq 0$. In other words, past VAR violations should not be informative about current and future violations. This property is also valid for any variable belonging to the Ω_{t-1} information set available at time $t-1$. So, for example, variable $I_t(\alpha)$ must be independent of variables such as past returns, past squared returns and past values of VAR.

[3] A fairly well-known example is that of the assessment of volatility models in which *ex-post* daily volatility is approximated by the "realized volatility", defined as the sum of squared intraday returns (Andersen *et al.*, 2003).

The first hypothesis is fully intuitive: for an $\alpha\%$ coverage level, the occurrence of losses exceeding VAR forecasts must correspond to $\alpha\%$ of the total number of periods for which the VAR is forecast. For a 5% VAR, used as a reference measure over 1000 periods, the expected number of violations is equal to 50. If the number of violations is significantly higher or lower than 50, then the VAR model fails the test. Thus, tests based on this hypothesis aim to verify whether the observed frequency of violations (or frequency of losses that exceed VAR) is sufficiently close to the predicted coverage rate. The most common test is the basic frequency (or binomial) test of Kupiec (1995).

However, the unconditional coverage property does not give any information about the temporal dependence of violations. The independence property of violations is nevertheless an essential property, because it is related to the ability of a VAR model to accurately model the higher order dynamics of returns. In fact, a model which does not satisfy the independence property can lead to clusterings of violations (for a given period), even if it has the correct average number of violations (ie, and satisfies unconditional coverage). So, there must be no dependence in the violations sequence, whatever the coverage rates considered. Indeed, the pioneering work of Berkowitz and O' Brien (2002) shows that the VAR models used by six big American commercial banks tend not only to be very conservative as regards risk – ie, they tend to overestimate the banks' VARs – but also to lead to violations clusters which highlight their inability to forecast changes in volatility. To quote from their study:

> Two important findings are that, unconditionally, the VAR estimates tend to be conservative relative to the 99th percentile of [the distribution of profit and loss]. However at times, losses can substantially exceed the VAR, and such events tend to be clustered. This suggests that the banks' models, besides a tendency toward conservatism, have difficulty forecasting changes in the volatility of profit and loss. (Berkowitz and O'Brien, 2002, p. 1094).

It is important to note that these two VAR properties are distinct and, of course, if a VAR measure does not satisfy both of these two hypotheses, it must be considered as not valid (Christoffersen, 1998). Let us finally note that these two hypotheses (of unconditional coverage and independence) are satisfied when the VAR violation process is a martingale difference (Berkowitz *et al.*, 2005), as indicated in the following:

$$E\left[I_t(\alpha) - \alpha \,\middle|\, \Omega_{t-1} \right] = 0 \tag{4}$$

where the information set Ω_{t-1} can include not only past VAR violations defined for the $\alpha\%$ coverage rate, ie, $\{I_{t-1}(\alpha), I_{t-2}(\alpha), \dots\}$ but also any variable Z_{t-k} known at time $t-1$, such as past VAR levels, returns and squared returns. The martingale difference property implies that for all $Z_{t-k} \in \Omega_{t-1}$, $\mathrm{E}[(I_t(\alpha) - \alpha) \otimes Z_{t-k}] = 0$, and if $I_{t-k}(\beta) \in \Omega_{t-1}$, then

$$E\left\{\left[I_t(\alpha)-\alpha\right]\left[I_{t-k}(\beta)-\beta\right]\right\}=0 \quad \forall(\alpha,\beta) \quad \forall k\neq 0 \tag{5}$$

Here we find the independence property, whereas the law of iterated expectations[4] leads to the unconditional coverage property.

To date, there are two major categories of conditional coverage test in the literature. The first category includes tests in the spirit of Christoffersen (1998): these include the dynamic quantile (DQ) test of Engle and Manganelli (2004), the white noise test of Berkowitz *et al.* (2005), etc. These tests aim to test separately or jointly the unconditional coverage and independence hypotheses, for a given coverage rate. The VAR violations included in the information set Ω_{t-1} are the only ones related to the reference coverage rate α, ie, $I_{t-k}(\alpha)$. As noted above, this category of tests corresponds to tests of "interval forecasts" because the relevant tests can be used to evaluate the "goodness" of a particular sequence of interval forecasts.

By contrast, the second category consists of the "density forecast" tests, which test the conditional coverage property for all possible coverage rates. The idea is to assess the whole of the forecast profits and losses distribution, without restricting ourselves to the forecasts of any particular quantile or VAR.

We now present the main tests to be considered. We restrict our attention to the tests of conditional coverage hypothesis, even if they can be used to test separately unconditional coverage and independence properties. For example, the DQ test can be adapted to test only the independence hypothesis.

2.1 Interval forecast approach

When we seek to jointly test the unconditional coverage and independence hypotheses, we encounter the problem of specifying the form of the dependence of the $I_t(\alpha)$ processes under the alternative hypothesis. Various solutions to this problem have been proposed.

Christoffersen (1998) supposes that, under the alternative hypothesis of VAR inefficiency, the process of violations $I_t(\alpha)$ can be modeled as a Markov chain whose matrix of transition probabilities is defined by

$$\Pi=\begin{pmatrix}\pi_{00} & \pi_{01}\\ \pi_{10} & \pi_{11}\end{pmatrix} \tag{6}$$

where $\pi_{ij}=\Pr[I_t(\alpha)=j \mid I_{t-1}(\alpha)=i]$. This Markov chain postulates the existence of a memory of order one in the $I_t(\alpha)$ process: the probability of having a violation (resp. not having one) for the current period depends on the occurrence or not of a violation in the previous period. The null hypothesis of conditional coverage is then defined by the following equality:

[4] Indeed, the null conditional moment $E[I_t(\alpha)-\alpha \mid \Omega_{t-1}]=0$ implies the null unconditional moment $E[I_t(\alpha)-\alpha]=0$ and so the equality $E[I_t(\alpha)]=\alpha$.

$$H_0: \quad \Pi = \Pi_\alpha = \begin{pmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{pmatrix} \tag{7}$$

If we accept the null hypothesis, then we accept the unconditional coverage hypothesis. Whatever the state of the system in $t-1$, the probability of having a violation at time t is equal to α, the coverage rate, ie $\pi_t = \Pr[I_t(\alpha) = 1] = \alpha$. Furthermore, the probability of a violation at time t is independent of the state in $t-1$. A simple likelihood ratio statistic, denoted LR_{CC}, then allows us to test the null hypothesis of conditional coverage. Under H_0, Christoffersen shows that

$$LR_{CC} = -2\left\{\ln L\left[\Pi_\alpha, I_t(\alpha), \ldots, I_T(\alpha)\right] - \ln L\left[\hat{\Pi}, I_t(\alpha), \ldots, I_T(\alpha)\right]\right\} \underset{T\to\infty}{\overset{d}{\to}} \chi^2(2) \tag{8}$$

where $\hat{\Pi}$ is the maximum likelihood estimator of the transition matrix under the alternative hypothesis:

$$\hat{\Pi} = \begin{pmatrix} \dfrac{n_{00}}{n_{00}+n_{01}} & \dfrac{n_{01}}{n_{00}+n_{01}} \\ \dfrac{n_{10}}{n_{10}+n_{11}} & \dfrac{n_{11}}{n_{10}+n_{11}} \end{pmatrix} \tag{9}$$

where n_{ij} is the number of times we have $I_t(\alpha) = j$ and $I_{t-1}(\alpha) = i$. $\ln[\hat{\Pi}, I_t(\alpha), \ldots, I_t(\alpha)]$ is the log-likelihood of the sequence $I_t(\alpha)$ associated with $\hat{\Pi}$:

$$\ln L\left[\hat{\Pi}_\alpha, I_t(\alpha), \ldots, I_T(\alpha)\right] = (1-\hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1-\hat{\pi}_{11})^{n_{10}} \hat{\pi}_{11}^{n_{11}} \tag{10}$$

Let us note that the log-likelihood under the null hypothesis is equal to

$$\ln L\left[\Pi_\alpha, I_t(\alpha), \ldots, I_T(\alpha)\right] = (1-\alpha)^{n_0} \alpha^{n_1} \tag{11}$$

where $n_0 = n_{00} + n_{10}$ and $n_1 = n_{01} + n_{11}$.

While this test is easy to use, it is rather limited for two main reasons. Firstly, independence is tested against a very particular form of alternative dependence structure that does not take into account dependences of order higher than one. Moreover, the use of a Markov chain makes it possible only to measure the influence of past violations $I_t(\alpha)$ and not the influence of any other exogenous variable.

2.1.1 DQ test of Engle and Manganelli (2004)

These latter two drawbacks are overcome by a test recently proposed by Engle and Manganelli (2004). They suggest using a linear regression model linking

current violations to past violations. Let $Hit_t(\alpha) = I_t(\alpha) - \alpha$ be the de-meaned process on α associated with $I_t(\alpha)$:

$$Hit_t(\alpha) = \begin{cases} 1-\alpha & \text{if } r_t < VAR_{t|t-1}(\alpha) \\ -\alpha & \text{else} \end{cases} \tag{12}$$

We now consider the following linear regression model:

$$Hit_t(\alpha) = \delta + \sum_{k=1}^{K} \beta_k Hit_{t-k}(\alpha) + \\ + \sum_{k=1}^{K} \gamma_k g\left[Hit_{t-k}(\alpha), Hit_{t-k-1}(\alpha), \ldots, z_{t-k}, z_{t-k-1}, \ldots\right] + \varepsilon_t \tag{13}$$

where ε_t is an i.i.d. process and where $g(\cdot)$ is a function of past violations and of variables z_{t-k} from the available information set Ω_{t-1} (eg, past returns, etc.). But, whatever the variables we include, the null hypothesis of conditional coverage corresponds to the joint nullity of the coefficients β_k and γ_k and of the constant δ, ie:

$$H_0 : \delta = \beta_k = \gamma_k = 0 \qquad \forall k = 1, \ldots, K \tag{14}$$

$\beta_k = \gamma_k = 0$ reflects the independence hypothesis, whereas $\delta = 0$ reflects the unconditional coverage hypothesis. Indeed, under the null hypothesis $\mathrm{E}[Hit_t(\alpha)] = \mathrm{E}(\varepsilon_t) = 0$, which implies that $\Pr[I_t(\alpha) = 1] = \mathrm{E}[I_t(\alpha)] = \alpha$. The joint nullity test of all coefficients, including the constant, therefore corresponds to a conditional coverage test. A LR statistic or a Wald statistic can then be used to test the simultaneous nullity of these coefficients. If we now let $\Psi = (\delta, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_K)'$ be the vector of the $2K + 1$ parameters in this model and let Z be the matrix of explanatory variables of model (13), then the Wald statistic DQ_{CC} associated with a test of conditional coverage is

$$DQ_{CC} = \frac{\hat{\Psi}' Z' Z \hat{\Psi}}{\alpha(1-\alpha)} \underset{T\to\infty}{\overset{d}{\to}} \chi^2(2K+1) \tag{15}$$

2.1.2 Martingale difference test by Berkowitz et al. (2005)

Berkowitz *et al.* (2005) start from the fact that the conditional coverage hypothesis is nothing but the martingale difference hypothesis of the $Hit_t(\alpha)$ process, and several tests of the martingale difference hypothesis can then be used to test VAR models for a given coverage rate α. They particularly focus on tests based on the spectral density of $Hit_t(\alpha)$, and on the univariate Ljung–Box test applied to the $Hit(\alpha)$ sequence. For the Ljung–Box test, the statistic associated with the nullity of the first K autocorrelations of the violation process obeys

$$LB(K) = T(T+2)\sum_{i=1}^{K}\frac{\hat{r}_i^2}{T-i} \underset{T\to\infty}{\overset{d}{\to}} \chi^2(K) \tag{16}$$

where $\hat{r}_i^2$ is the ith order empirical autocorrelation of the $Hit(\alpha)$ process.

Other interval forecast tests can also be mentioned in this context. For example, tests of the duration between two violations, such as that of Christoffersen and Pelletier (2004), allow one to consider wider dependences than those specified under the Markov chain hypothesis or within the confines of the linear probability model. However, the test logic remains unchanged: we test the conditional coverage hypothesis for some single given coverage level.

2.2 Density forecast approach

The tests mentioned above deal with conditional coverage for a single coverage rate α. However, the property of conditional coverage must be valid for any coverage rate and this reasoning, pushed to the limit, leads to tests of the conditional coverage hypothesis for all possible coverage rates between zero and one. This is the basic principle exploited by tests based on the complete density forecast approach (Crnkovic and Drachman, 1997; Diebold *et al.*, 1998; Berkowitz, 2001). Assessing the forecasts of the profit/loss distribution then means we test the accuracy of interval forecasts for all levels of nominal coverage.

These tests of forecast densities use the PIT[5] transformation (probability integral transformation). Suppose that r_t is the observed return of an asset and $F_{t-1}(\cdot)$ is the corresponding distribution function. Under the null hypothesis, the PIT transformation then implies that

$$X_t = F_{t-1}(r_t) \sim \text{iid}\, U_{[0,1]} \tag{17}$$

Testing the validity of the VAR model corresponds to testing this hypothesis. As Berkowitz (2001) emphasizes:

> Therefore, if banks are required to regularly report forecast distributions $F(\cdot)$, regulators can use this probability integral transformation and then test for violations of independence and/or uniformity. Moreover, this result holds regardless of the underlying distribution of the portfolio returns, r_t, and even if the forecast model $F(\cdot)$ changes over time." (Berkowitz, 2001, p 7).

Given this general principle, different techniques can be used to test independence and/or uniformity. For example, Crnkovic and Drachman (1997) suggest using Kuiper statistics to test uniformity, whereas Diebold, Gunther and Tay (1998) suggest the use of non-parametric tests (Kolmogorov-Smirnov, Cramer–Von

[5] With this transformation, if Y_t is a random variable with distribution function $F_t(\cdot)$, the transformed random variable $X_t = F_t(Y_t)$ is uniformly distributed on the interval $[0,1]$.

Mises) to evaluate the significance of the distance between the transformed series and theoretical distribution $U(0,1)$. For his part, Berkowitz (2001) proposes a parametric test which is based on a further transformation. More particularly, if $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function of the standard normal distribution, he suggests we take the transformation:

$$Z_t = \Phi^{-1}(X_t) = \Phi^{-1}\left[F_{t-1}(r_t)\right] \sim \text{iid } N(0,1) \tag{18}$$

Berkowitz then proposes to test this prediction using a variety of likelihood ratio tests, For example, one can test $H_0 : Z_t$ i.i.d. $N(0,1)$ against an alternative of the following form:

$$H_1 : Z_t = \mu + \rho_1 Z_{t-1} + \cdots + \rho_n Z_{t-n} + \gamma_1 Z_{t-1}^2 + \cdots + \gamma_m Z_{t-m}^2 + \mu_t \tag{19}$$

In this case, the null hypothesis implies $n + m + 2$ constraints, ie, $\mu = \rho_1 = \ldots = \rho_n = \gamma_1 = \ldots = \gamma_m$ and $\sigma_{Z_t} = 1$. From various Monte Carlo simulation exercises, Berkowitz shows that the LR test associated with H_0 is a powerful model-adequacy test even with sample sizes as small as 100.

We also note here that tests for complete density forecast evaluation can be generally modified if we are interested in the distribution of tail losses rather than in a complete profits and losses distribution.[6]

3 A multivariate portmanteau statistic

We propose to expand the Berkowitz *et al.* test of conditional coverage to the multivariate case. Our test is based on a multivariate portmanteau statistic that enables us to jointly test the property of conditional coverage for a set of coverage rates.

Formally, the martingale difference hypothesis as formulated by Berkowitz *et al.* (2005), ie $E[Hit_t(\alpha) \mid \Omega_{t-1}] = 0$, implies that for a coverage rate α:

$$E\left[Hit_t(\alpha)\, Hit_{t-k}(\alpha)\right] = 0 \qquad \forall k \in N^* \tag{20}$$

As noted above, this hypothesis also implies that violations associated with different coverage rates should be independent:

$$E\left[Hit_t(\alpha)\, Hit_{t-k}(\beta)\right] = 0 \qquad \forall k \in N^* \qquad \forall (\alpha,\beta) \tag{21}$$

This latter property forms the basis of our multivariate portmanteau test. We then take into account both violations autocorrelations for a given coverage rate ($\alpha = \beta$), and also combined cross-correlations between violations obtained for different coverage rates ($\alpha \neq \beta$).

[6] See for example Berkowitz (2001), Christoffersen and Pelletier (2004) and Dowd (2005).

Let $\Theta = \{\theta_1, \ldots, \theta_m\}$ be a discrete set of m different coverage rates that are positive but less than 1. Let $Hit_t = [Hit_t(\theta_1), Hit_t(\theta_2), \ldots, Hit_t(\theta_m)]$ be the vector of violations associated with these m coverage rates at time t. The conditional coverage hypothesis then implies that

$$\mathrm{Cov}\left(Hit_t, Hit_{t-k}\right) = E\left[Hit_t\, Hit'_{t-k}\right] = V^* \delta_k \tag{22}$$

where V is a (m, m) symmetric non-zero matrix and where δ_k is a scalar:[7]

$$\delta_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{else} \end{cases} \tag{23}$$

For some chosen order $K \geq 1$, we then test the joint nullity of the autocorrelations from order 1 to K for the vector process Hit_t:

$$H_0 : \mathrm{Cov}\left(Hit_t, Hit_{t-k}\right) = V^* \delta_k \qquad \forall\, k = 1, \ldots, K \tag{24}$$

This test is nothing but a multivariate extension of common portmanteau tests (eg, Box–Pierce or Ljung–Box), and several multivariate portmanteau statistics can be used (Chitturi, 1974; Hosking, 1980; Li and McLeod, 1981). Under the null hypothesis of absence of autocorrelation in the vector Hit_t, Li and McLeod (1981) prove that

$$Q_m(K) \underset{T \to \infty}{\overset{d}{\to}} \chi^2\left(Km^2\right) \tag{25}$$

with $Q_m(K)$ equal to

$$Q_m(K) = T \sum_{k=1}^{K} \left(\mathrm{vec}\,\hat{R}_k\right)' \left(\hat{R}_0^{-1} \otimes \hat{R}_0^{-1}\right)\left(\mathrm{vec}\,\hat{R}_k\right) \tag{26}$$

and $\hat{R}_k = D\hat{C}_k D$, where D is the diagonal matrix that has as its constituents the set of inverses of the standard deviations associated with univariate processes $Hit_t(\theta_i)$ defined by $\hat{c}_{ii0}^{-0.5}$. We then take $\hat{C}_k$, the matrix of empirical covariances associated with vector Hit_t, ie:

$$\hat{C}_k = \left(\hat{c}_{ijk}\right) = \sum_{t=k+1}^{T} Hit_t\, Hit'_{t-k} \qquad \forall\, k \in N^* \tag{27}$$

[7] It is also possible to test the independence hypothesis by defining covariance as follows:

$$H_0 : \mathrm{Cov}\left(Hit_t, Hit_{t-k}\right) = E\left\{\left[Hit_t - E\left(Hit_t\right)\right]\left[Hit_{t-k} - E\left(Hit_{t-k}\right)\right]'\right\} = V^* \delta_k$$

Let us keep in mind that, indeed, under the unconditional coverage hypothesis $\mathrm{E}(Hit_{t-k}) = 0\ \forall\, k$.

It is important to note that m must be chosen carefully to avoid possible singularity of the matrix $\hat{R}_0$ because $Q_m(K)$ cannot be calculated if the matrix $\hat{R}_0$ is singular. In our simulations, we found that the probability of this matrix being singular increases when we use coverage rates that are very close to each other (for example 1% and 1.5%).[8] So, in any empirical application, one must choose coverage rates that are relatively far from each other. We consider here the set $\Theta = \{1\%, 5\%\}$ for $m = 2$, and for $m = 3$ the set $\Theta = \{1\%, 5\%, 10\%\}$, which also correspond to common coverage rates used in risk management.

TABLE 1 *P*-values of Kolmogorov-Smirnov test.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 10$	$K = 15$
				$m = 2$			
$T = 250$	0.1466	0.1917	0.4486	0.0075	0.0001	0.0000	0.0000
$T = 500$	0.2042	0.8972	0.2289	0.4000	0.4497	0.0072	0.0861
$T = 750$	0.3178	0.2149	0.5928	0.2767	0.7589	0.0279	0.6626
$T = 1{,}000$	0.7208	0.3275	0.6255	0.4023	0.0423	0.0328	0.0537
				$m = 3$			
$T = 250$	0.8347	0.4783	0.7883	0.5271	0.0209	0.0000	0.0000
$T = 500$	0.4266	0.6577	0.6782	0.5815	0.4116	0.0005	0.0012
$T = 750$	0.2711	0.8076	0.1426	0.4681	0.1558	0.1057	0.0153
$T = 1{,}000$	0.2043	0.5653	0.7909	0.1506	0.3675	0.0473	0.0013
	$K = 20$	$K = 25$	$K = 30$	$K = 35$	$K = 40$	$K = 45$	$K = 45$
				$m = 2$			
$T = 250$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$T = 500$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$T = 750$	0.1421	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000
$T = 1{,}000$	0.0671	0.4821	0.0033	0.0013	0.0000	0.0000	0.0000
				$m = 3$			
$T = 250$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$T = 500$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$T = 750$	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$T = 1{,}000$	0.0006	0.0000	0.0187	0.0000	0.0000	0.0000	0.0000

K denotes the lag order of the portmanteau test, T denotes the length of the VAR sample, and m corresponds to the number of VAR coverage rate included in the multivariate portmanteau test. For each configuration the p-value of the equivalence test (K-S test) with the chi-squared distribution is reported. *P*-values are computed with 1,000 simulations.

[8] When we consider very close coverage rates, then most likely the *Hit* matrix will have several identical columns (for example, the same occurrences of violations at 1% and at 1.5% for small samples).

The choice of K is a common traditional problem with portmanteau tests. Following Bender and Grouven (1993), we determine the choice of the acceptable values of lag order K using simulations. We set the sample size T at 250, 500, 750 and 1,000. For every couple (T,m), with $m = 2,3$, we simulate a multivariate white noise process of dimension (T,m). The statistic of Li and McLeod (1981) is then calculated for various values of K. We repeat the simulation 1,000 times for each triplet (T,m,K), and collect the series of $Q_m(K)$ statistics. For each series, the Kolmogorov-Smirnov (K-S) test is then run to assess the equivalence with the predicted theoretical distribution $\chi^2(Km^2)$. P-values of the K-S test are shown for every triplet (T,m,K) in Table 1. A global reading of these results suggests that with high values of K, there is a wide gap between the empirical distribution of statistic $Q_m(K)$ and its asymptotic distribution (ie, the null hypothesis of equivalence is rejected for large values of K). So, for these sample sizes, we suggest choosing K from $K \in \{1,2,3,4,5\}$.

4 Small-sample properties

In this section we first study the finite sample properties of our test to characterize the impact of the choice of the lag order K and m, the dimensionality of our test statistic. We also use a set of Monte Carlo experiments to compare the performance of our $Q_m(K)$ test with that of more traditional interval forecast tests (and more specifically, the LR_{CC} test of Christoffersen (1998) and the DQ_{CC} test of Engle and Manganelli (2004)).

4.1 Empirical size of the $Q_m(K)$ test

We first assess the empirical size of our new test for various values of the parameters m and K. To do so, we simulate the return distribution from an EGARCH process calibrated by Campbell (2006) which duplicates the dynamics of monthly returns for several American indexes:

$$r_{t,t-1} = v_{t,t-1} \tag{28}$$

$$v_{t,t-1} \sim N\left(0,\sigma_t^2\right) \tag{29}$$

$$\ln\left(\sigma_t^2\right) = 0.02 + 0.94\ln\left(\sigma_{t-1}^2\right) + 0.22\left|\frac{v_{t,t-1}}{\sigma_{t-1}}\right| - 0.05\frac{v_{t,t-1}}{\sigma_{t-1}} \tag{30}$$

The empirical size[9] of our test is then assessed using a VAR model in which the true dynamics of returns is known, and this ensures that the conditional coverage

[9] In traditional hypothesis testing the empirical size of a test corresponds to the Type I error rate.

hypothesis is verified. VAR is then computed for a given coverage rate from the conditional variance determined by equation (30). The VAR is then equal to:[10]

$$VAR_{t|t-1}(\alpha) = \Phi^{-1}(\alpha)\left[\exp\left(0.02 + 0.94\ln\left(\sigma_{t-1}^2\right) + 0.22\left|\frac{v_{t,t-1}}{\sigma_{t-1}}\right| - 0.05\frac{v_{t,t-1}}{\sigma_{t-1}}\right)\right]^{0.5} \tag{31}$$

Series of out-of-sample VARs are then generated for sample sizes T = 250, 500, 750, 1,000. From the VAR-violations sequences observed *ex post* (ie, the *Hit* function), the test statistic $Q_m(K)$ is computed 10,000 times. The empirical size then corresponds to the rejection frequencies of the conditional coverage hypothesis observed in these simulations. If the asymptotic distribution of our test is adequate, then these rejection frequencies should be close to the nominal size in our experiments, which is set at 10%.

Table 2 presents the empirical size of the conditional coverage test $Q_m(K)$ for various sample sizes T, lag orders K and different values of parameter m of our multivariate statistic. Remember that our test is equivalent to that of the Berkowitz *et al.* (Ljung–Box test) when m is set to one ($m = 1$). The first part of Table 2 shows the empirical size of the Ljung–Box test for $m = 1$, and the second and third parts show the empirical size of our test statistic for coverage rates of $m = 2$ and $m = 3$.

The results show that an increase in the dimensionality m leads to a slight increase in the test's empirical size, which tends to stabilize to around 14% on average for a 10% nominal size.

4.2 Finite sample power of the $Q_m(K)$ test

We propose two different ways to assess the power of our conditional coverage test. The difference is related to the VAR model considered. Indeed, to simulate the power, we must choose a VAR model that does not empirically fit the true return distribution – ie, we use two VAR models that do not fit the data well.[11] The first is an historical simulation (HS) VAR model. By definition, HS-VAR is the empirical α-quantile of past returns observed on the last *Te* periods (we set *Te* at 250):

$$VAR_{t|t-1}(\alpha) = \text{percentile}\left(\left\{r_j\right\}_{j=t-Te}^{t-1}, 100\,\alpha\right) \tag{32}$$

The second is a delta-normal VAR model, in which the VAR is given by the following equality:

[10] We begin with a starting sample of size N and the T out-of-sample VARs are computed using a recursive forecasting scheme for the volatility. Then the first forecast is formed using data from 1 to N, the second using data from 2 to $N + 1$, and so forth.

[11] Let us note that the power is the probability of rejecting the null when the latter is false.

TABLE 2 Actual sizes of LB(K) and $Q_m(K)$ tests.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
			LB(K) = {1%}		
T = 250	0.0282	0.0521	0.0628	0.0815	0.0747
T = 500	0.0455	0.0892	0.1119	0.1258	0.1196
T = 750	0.0696	0.1098	0.1258	0.1653	0.1619
T = 1,000	0.0801	0.1192	0.1456	0.1607	0.1407
			$Q_2(K)$ = {1%, 5%}		
T = 250	0.1317	0.1533	0.1645	0.1680	0.1662
T = 500	0.1181	0.1383	0.1486	0.1534	0.1536
T = 750	0.1181	0.1425	0.1500	0.1529	0.1572
T = 1,000	0.1246	0.1368	0.1430	0.1490	0.1457
			$Q_3(K)$ = {1%, 5%, 10%}		
T = 250	0.1576	0.1678	0.1566	0.1628	0.1645
T = 500	0.1335	0.1487	0.1523	0.1468	0.1547
T = 750	0.1290	0.1389	0.1418	0.1405	0.1459
T = 1,000	0.1286	0.1428	0.1323	0.1377	0.1419

For each simulation, the profit/loss distribution is generated under an EGARCH(1, 1) distribution with normal disturbances. The corresponding VAR is computed with the same EGARCH model and then satisfies the nominal coverage and independence assumptions. The empirical sizes of the Ljung–Box test and the multivariate portmanteau test correspond to the rejection frequencies of the null hypothesis obtained with 10,000 simulations. K denotes the lag order, T denotes the length of the VAR sample, and m corresponds to the number of different coverage rates included in the multivariate portmanteau test. Nominal size is 10%.

$$VAR_{t|t-1}(\alpha) = \Phi^{-1}(\alpha)\left[\text{variance}\left(\left\{r_j\right\}_{j=t-T}^{t-1}\right)\right]^{\frac{1}{2}} \tag{33}$$

The poor fits of these two VAR models are illustrated in Figures 1 and 2, where one can see at once that both models produce strongly pronounced violation clusters for both 1% VAR and 5% VAR.

To compute the power of our test, we use the methodology of Dufour (2004).[12] This enables one to compute the power of a given test, while maintaining nominal size independently of the number of replications.

Tables 3 and 4 report the power results of our test. These results are very clear-cut: whatever the lag order K or sample size T, moving from the univariate dimension to a multivariate dimension improves the power of the conditional

[12] See Christoffersen and Pelletier (2004) or Berkowitz *et al.* (2005) for a simple description of the methodology.

FIGURE 1 Simulated profits/losses and HS-VAR.

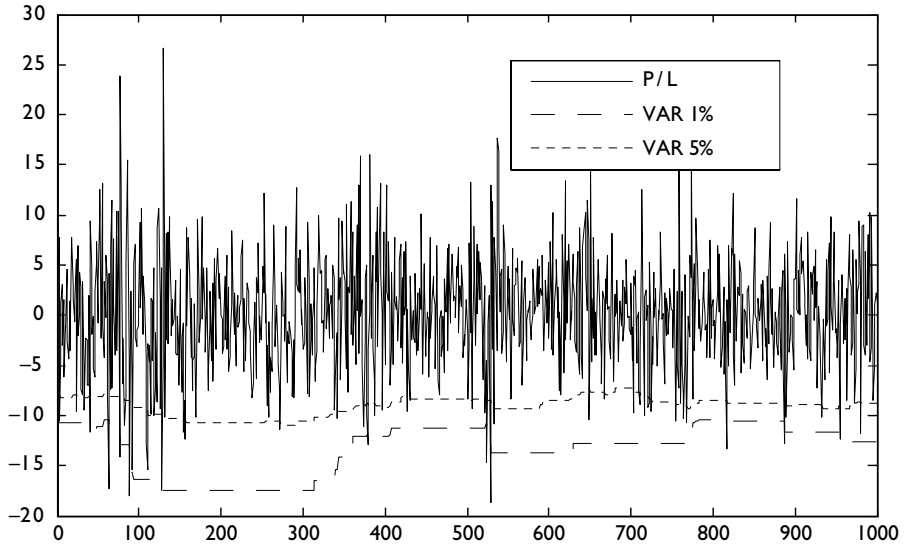

FIGURE 2 Simulated profits/losses and delta-normal VAR.

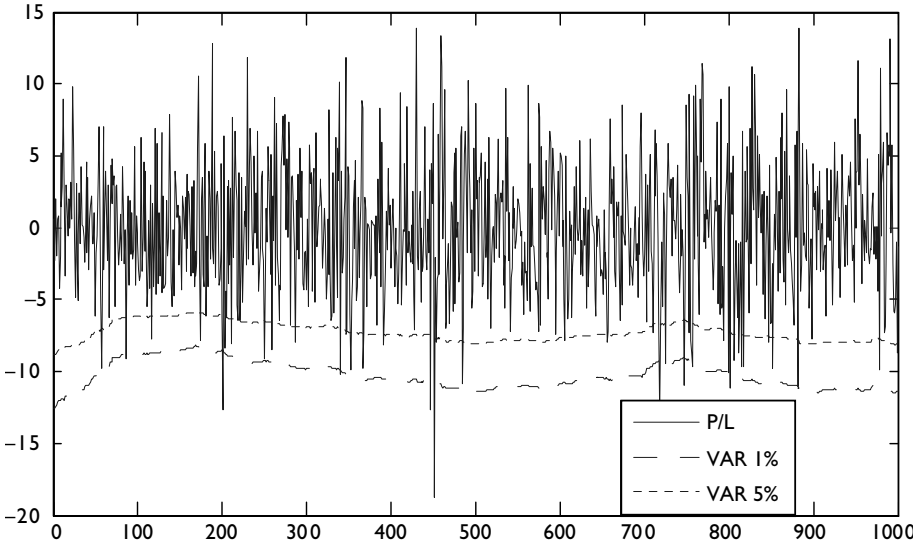

TABLE 3 Empirical power of LB(K) and $Q_m(K)$ tests. Case 1: historical simulation.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
			LB(K) = {1%}		
$T = 250$	0.2077	0.2507	0.2906	0.3261	0.3264
$T = 500$	0.2395	0.3678	0.4007	0.4240	0.4417
$T = 750$	0.3241	0.4324	0.4651	0.4760	0.4970
$T = 1{,}000$	0.3910	0.4525	0.4883	0.4948	0.5123
			$Q_2(K)$ = {1%, 5%}		
$T = 250$	0.3016	0.3644	0.4229	0.4322	0.4813
$T = 500$	0.4087	0.5050	0.5517	0.6042	0.6258
$T = 750$	0.4942	0.5958	0.6644	0.7109	0.7254
$T = 1{,}000$	0.5484	0.6604	0.7393	0.7896	0.8106
			$Q_3(K)$ = {1%, 5%, 10%}		
$T = 250$	0.3133	0.4013	0.4539	0.4937	0.5025
$T = 500$	0.4207	0.5347	0.6073	0.6532	0.6869
$T = 750$	0.5162	0.6519	0.7112	0.7478	0.7889
$T = 1{,}000$	0.5695	0.7121	0.7879	0.8321	0.8485

For each simulation, the profit/loss distribution is generated under an EGARCH(1, 1) distribution with normal disturbances. At each date, the historical simulation VAR is simply the unconditional quantile of the past 250 daily observations. It does not satisfy the independence assumption and/or the nominal coverage. The empirical power is computed using the Dufour methodology. K denotes the lag order of the portmanteau test, T denotes the length of the VAR sample, and m corresponds to the number of different coverage rates included in the multivariate portmanteau test. Nominal size is 10%.

coverage test very significantly. For example, for a sample size of 250 observations (one year of quotations) and a lag order equal to 5, our test power is about 50% for m equal to 2 or 3, but only 32% for m equal to 1.

4.3 Finite sample properties of conditional coverage tests

To finish, we now compare the power of our test with that of Christoffersen's (1998) LR_{CC} test and Engle and Manganelli's (2004) DQ_{CC} test. These results are reported in Tables 5 and 6. As these tests deal with one coverage level, we simulate the power respectively for three different coverage rates (1%, 5%, 10%). For ease of presentation, we report the power of our test only for lag order $K = 5$. We note that our test gives much better results than Christoffersen's LR_{CC} test (1998) or the DQ_{CC} test (and in the latter case the improvement is especially pronounced for the 1% VAR). Overall, the power improvement ranges from 12% to 29%. In addition to the power gain, it should be noted that the null hypothesis in our new test is more general than that of the LR_{CC} or DQ_{CC} test because we simultaneously test the VAR model for different relevant levels of coverage.

TABLE 4 Empirical power of LB(K) and $Q_m(K)$ tests. Case 2: delta-normal.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
			LB(K) = {1%}		
$T = 250$	0.3314	0.3743	0.3785	0.3891	0.3732
$T = 500$	0.3811	0.4731	0.4556	0.4504	0.4549
$T = 750$	0.4729	0.4769	0.4840	0.5038	0.5091
$T = 1{,}000$	0.5379	0.4766	0.5362	0.5447	0.5640
			$Q_2(K)$ = {1%, 5%}		
$T = 250$	0.3176	0.3502	0.4107	0.4423	0.4842
$T = 500$	0.4190	0.5065	0.5555	0.5806	0.6270
$T = 750$	0.4903	0.5846	0.6708	0.7118	0.7365
$T = 1{,}000$	0.5436	0.6826	0.7467	0.7879	0.8165
			$Q_3(K)$ = {1%, 5%, 10%}		
$T = 250$	0.3141	0.3902	0.4518	0.4801	0.5160
$T = 500$	0.4240	0.5345	0.6259	0.6540	0.6834
$T = 750$	0.5239	0.6473	0.7233	0.7775	0.7963
$T = 1{,}000$	0.5942	0.7192	0.7894	0.8394	0.8625

For each simulation, the profit/loss distribution is generated under an EGARCH(1, 1) distribution with normal disturbances. At each date, the delta-normal VAR is simply computed using formula (33). It does not satisfy the independence assumption and/or the nominal coverage. The empirical power is computed using the Dufour methodology. K denotes the lag order of the portmanteau test, T denotes the length of the VAR sample, and m corresponds to the number of different coverage rates included in the multivariate Portmanteau test. Nominal size is 10%.

5 Conclusions

In this work we have proposed a new VAR validation test that is halfway between the "interval forecast" and "density forecast" approaches to backtesting. Our test looks at violations over multiple coverage rates (or confidence levels). It is based on the white noise property of the process of violations for any given coverage rate, but it also looks at multiple coverage rates at the same time.

The proposed test is easy to implement, and we showed in a set of Monte Carlo replications that it is more powerful in small sample sizes than the more traditional interval forecasts VAR validation tests, which deal with only one level coverage rate. We have no doubt that the method proposed here can be expanded further, and an obvious example would be to expand the Engle–Manganelli (2004) test using a multivariate binary model.

TABLE 5 Empirical power of the LR_{CC}, DQ_{CC} and $Q_m(K)$ tests. Case 1: historical simulation.

	LR_{CC}			DQ_{CC}			$Q_2(K)$	$Q_3(K)$
	1%	5%	10%	1%	5%	10%	—	—
T = 250	0.2128	0.2953	0.2365	0.3892	0.4817	0.4860	0.4813	0.5025
T = 500	0.2314	0.2687	0.2540	0.4302	0.5886	0.5795	0.6258	0.6869
T = 750	0.2112	0.2855	0.2843	0.4904	0.6889	0.6802	0.7254	0.7889
T = 1,000	0.1812	0.3058	0.3526	0.5604	0.7677	0.7685	0.8106	0.8485

For each simulation, the profit/loss distribution is generated under an EGARCH(1, 1) distribution with normal disturbances. At each date, the historical simulation VAR is simply the unconditional quantile of the past 250 daily observations. It does not satisfy the independence assumption and/or the nominal coverage. The empirical power is computed using the Dufour methodology. The lag order K of the multivariate portmanteau test is kept at 5. T denotes the length of the VAR sample, and m corresponds to the number of different coverage rates included in the multivariate portmanteau test. Nominal size is 10%.

TABLE 6 Empirical power of the LR_{CC}, DQ_{CC} and $Q_m(K)$ tests. Case 2: delta-normal.

	LR_{CC}			DQ_{CC}			$Q_2(K)$	$Q_3(K)$
	1%	5%	10%	1%	5%	10%	—	—
T = 250	0.3191	0.3066	0.2590	0.3838	0.4148	0.4141	0.4842	0.5160
T = 500	0.3713	0.3139	0.3038	0.4849	0.4904	0.4821	0.6270	0.6834
T = 750	0.4259	0.3316	0.3463	0.5560	0.5655	0.5536	0.7365	0.7963
T = 1,000	0.4426	0.3409	0.4347	0.5959	0.6209	0.6446	0.8165	0.8625

For each simulation, the profit/loss distribution is generated under an EGARCH(1, 1) distribution with normal disturbances. At each date, the delta-normal VAR is simply computed using formula (33). It does not satisfy the independence assumption and/or the nominal coverage. The empirical power is computed using the Dufour methodology. The lag order K of the multivariate portmanteau test is kept at 5. T denotes the length of the VAR sample, and m corresponds to the number of different coverage rates included in the multivariate portmanteau test. Nominal size is 10%.

REFERENCES

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625.

Beder, T. (1995). VaR: seductive but dangerous. *Financial Analysts Journal* **51**(5), 12–24.

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

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

Berkowitz, J., and O'Brien J. (2002). How accurate are the value-at-risk models at commercial banks. *Journal of Finance* **57**, 1093–111.

Box, G. E. P., and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series. *Journal of the American Statistical Association* **65**, 1509–26.

Campbell, S. D. (2006). A review of backtesting and backtesting procedures. *Journal of Risk* **9**(2), current issue.

Chitturi, R. V. (1974). Distributions of residual autocorrelations in multiple autoregressive schemes. *Journal of the American Statistical Association* **69**, 928–34.

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

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

Clements, M. P., and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: application to output growth and unemployment. *Journal of Forecasting* **19**, 255–76.

Clements, M. P., and Taylor, N. (2003). Evaluating prediction intervals for high-frequency data. *Journal of Applied Econometrics* **18**, 445–56.

Crnkovic, C., and Drachman, J. (1996) "Quality control." *Risk* **9** (September), 139–43.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts. *International Economic Review* **39**, 863–83.

Dowd, K. (2005). *Measuring market risk.* John Wiley & Sons.

Dufour, J. M. (2004). Monte Carlo tests with nuisance parameters: a general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, forthcoming

Durlauf, S. N. (1991). Spectral based testing of the martingale hypothesis. *Journal of Econometrics* **50**, 355–76.

Engle, R. F., and Manganelli, S. (2001). Value at risk models in finance. Working paper series 75, European Central Bank.

Engle, R. F., and Manganelli, S. (2004). CAViaR: conditional autoregressive value-at-risk by regression quantiles. *Journal of Business and Economic Statistics* **22**, 367–81.

Gouriéroux, C. (2000). *Econometrics of qualitative dependent variables.* Cambridge University Press.

Hosking, J. R. M. (1980). The multivariate portmanteau statistic. *Journal of the American Statistical Association* **75**, 602–8.

Jorion P. (2001). *Value-at-risk: the new benchmark for managing financial risk.* McGraw-Hill.

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

Li, W. K., and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society, Series B,* **43**, 231–9.

Patton, A. J. (2002). Application of copula theory in financial econometrics. Ph.D. dissertation, Department of Economics, UCSD.

We'll deal with pagination.