

Backtesting Value-at-Risk Accuracy: A New Simple Test

Christophe Hurlin

LEO, Université d'Orléans. Email: christophe.hurlin@univ-orleans.fr.

Sessi Tokpavi

Corresponding author. LEO, Université d'Orléans. Rue de Blois. BP 6739. 45067 Orléans Cedex 2. France. Email: sessi.tokpavi@univ-orleans.fr.

Abstract

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.

Key words: Value-at-Risk; Risk Management; Backtesting.

JEL classification: C22, C52, G28.

1 Introduction

Nowadays, prudential rules defined within the framework of the Basle Accord, leave the financial institutions free to develop their own internal model of risk management and computation of Value at Risk (VaR). The method of standard assessment of these models then consists in estimating *ex-post* the precision of VaR forecasts by backtesting procedures (See for example Jorion (2001) for a description of backtesting procedures). However, no particular backtesting technique is today recommended by the regulatory authorities in spite of the very special attention given within the framework of the Second Pillar of the Basle Accord to the validation procedures of internal models of risk assessment. Now, it is obvious that the choice of validation technique is a key issue of the transparency policy and risk management of financial institutions. How can we guarantee the relevance and the accuracy of a risk measurement like VaR , which generally results from a relatively complex model and about which a bank can find it difficult or even undesirable to communicate?

This problem would not be so acute if the very various possible methods of computation of VaR led to similar assessments of risk for one and the same portfolio. Indeed in the literature a multitude of methods can be found, which make it possible to calculate VaR . The internal model of VaR ¹ computation can be either based on a non parametric approach (hybrid method, historical simulation, etc.) or, on the contrary, on a particular parametric approach (univariate or multivariate ARCH-GARCH models, RiskMetrics) or else 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. So, Beder (1995) by using eight fairly common VaR measurements based on the combination of three criteria (type of modeling used for portfolio returns, namely historical simulation or Monte Carlo; hypothesis concerning correlations between portfolio assets; hypothesis on the holding period) shows that there are wide gaps in forecast VaR values. From

¹ See Dowd (2005) for a review of recent literature on the methods of VaR computation.

then on, how can we evaluate the precision of a *VaR* computation and choose the "best" model?

To date, in the literature there are two main types of statistical tests which can be used within the framework of backtesting procedures to assess *VaR* validity. The first main approach consists in testing two fundamental hypotheses concerning the process of *VaR* violations for a given coverage rate: the hypothesis of unconditional coverage and the hypothesis of independence (Christoffersen, 1998). This approach is sometimes called 'Event Probability Forecast Approach' (Clements and Taylor, 2003). Let us remember that a violation corresponds to a situation in which *ex-post* portfolio returns are lower than *VaR* forecasts². The hypothesis of unconditional coverage simply means that *VaR* for a 5% coverage rate for example, is valid only on condition that the expected frequency of observed violations is also equal to 5%.

As for the independence hypothesis, it shows that if the model of *VaR* calculation is valid, violations must be distributed independently. In other words, there must not be any violation cluster: the occurrence of a loss exceeding *VaR* forecasts does not contain information useful to forecast future violations. Testing on either of these two hypotheses or on both jointly was proposed in the literature for a given coverage level. Among these, Christoffersen's test (1998) is based on the use of a Markov chain, the 'hit regression' test of Engle and Manganelli (2004) based on a linear auto-regressive model, and more recently the tests of Berkowitz et al. (2005) based on tests of martingale difference or weak white noise (absence of autocorrelation in the de-meanded process of violations or Hit function).

On the contrary the second approach is based on the assessment of the density of loss and profit probability which makes the *VaR* forecasts possible. The method used here is totally different because the model used to compute *VaR* is tested without computing *VaR* (Crnkovic and Drachman, 1997; Diebold et al., 1998; Berkowitz, 2001). In so doing, the validity of the whole density of the

² Or the opposite of *VaR* if the latter is defined as a loss in absolute value.

distribution of losses and profits modeled *ex-ante* is tested and so the implicit validity of *VaR* for all coverage rates between 0 and 1 is put to the test. The advantage of this approach is that it indirectly exploits the property of independence for violation linked with *VaR* for various coverage rates. Indeed, the process of violations associated to a 5% *VaR* must be independent from future or past violations associated to a 1% *VaR*. This assessment method, also called 'Density Forecast Approach', belongs to the more general type of assessment of the non-linear model forecasts using probabilities density (Clements, 2003; Clements and Smith, 2000). However, one of the major weaknesses of this approach lies in the fact that the validation of the mode of *VaR* model is of interest only for extreme events. Indeed, backtesting the validity of *VaR* for a 50% coverage rate is of little interest for risk management as envisaged by prudential regulations.

In this paper, we propose a new test of *VaR* validation which takes advantage of both approaches found in the literature. This test is based on the weak white noise property of the process of violations for a given coverage rate (Event Probability Forecast Approach). It also takes into account the joint validation of this hypothesis for a finite set of coverage rates (Density Forecast Approach). More precisely, we propose Multivariate Portmanteau test statistic applied to a finite group of *VaR* violation processes associated to several coverage rates considered as relevant for the analysis of extreme risks. The statistic used is a multivariate extension of the autocorrelation test of Box and Pierce (1970), proposed by Li and McLeod (1981). For example we can assess a *VaR* model particularly by testing the property of conditional efficiency for violation processes associated to *VaR* at 1%, 5% and 10%, and not only violations for a single 5% coverage rate. In so doing, we exploit a much larger dataset to verify the validity of the model, just like the tests based on Density Forecast Evaluation. However our approach stays clear of the stumbling block which consists in estimating *VaR* for coverage rates which are irrelevant for the management of extreme risks. So, we show that our multivariate test statistic has very good properties in small sample sizes. Using various exercises of power comparison, it turns out that our test notably has better properties than directly comparable tests of the Event Probability Forecast approach.

The rest of the paper is organized as follows. In section 2, we present the main *VaR* assessment tests by distinguishing the two approaches mentioned above. Section 3 presents our multivariate statistic. Section 4 deals with the study of our test properties at finite distance and with various Monte Carlo power comparison exercises with the other *VaR* validation methods. Finally, the last section concludes and submits various extensions to our test.

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 comparison of the various forecast models is thus generally made by using a criterion (or a statistical loss function) such as for example the Mean Squared Error (MSE) criterion or the standard information criteria (AIC, BIC etc.). However this method is possible only if the *ex-post* realization of the variable of interest is observable. When it is unobservable, the assessment exercise then requires the use of a proxy for the unobservable variable with good properties, notably in terms of bias. 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 intra-day returns (Andersen et al., 2003). However, it is relatively delicate to compute such a proxy variable in the *VaR* approach. That is why *VaR* assessment criteria are generally based on statistical tests of the two main hypotheses that the process associated to *VaR* forecast violations has to confirm, namely the unconditional coverage hypothesis and the independence hypothesis.

Formally, we denote r_t the return of an asset or a portfolio of assets at time t . The *ex-ante* *VaR* for a $\alpha\%$ coverage rate noted $VaR_{t|t-1}(\alpha)$, anticipated conditionally to an information set, Ω_{t-1} , available at time $t - 1$ is defined by the following relation:

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

Let $I_t(\alpha)$ be the indicator variable associated to the *ex-post* observation of a $\alpha\%$ *VaR* violation at time t .

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

Indeed Christoffersen (1998) shows that the problem of *VaR* validity corresponds to the problem of knowing whether the violation sequence $\{I_t\}_{t=1}^T$ confirms the following two hypotheses or not (Campbell, 2005):

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

$$\Pr [I_t(\alpha) = 1] = E [I_t(\alpha)] = \alpha \quad (3)$$

- **The independence hypothesis:** *VaR* violations observed at two different dates for the same coverage rate must be distributed independently. Formally, variable $I_t(\alpha)$ associated to the *VaR* violation at time t for a $\alpha\%$ coverage rate is independent from variable $I_{t-k}(\alpha)$, $\forall k \neq 0$. In other words, it means that past *VaR* violations do not hold information on current and future violations. This property is also valid for any variable belonging to the Ω_{t-1} information set available at time $t-1$. In particular, variable $I_t(\alpha)$ must be independent from past returns, past values of *VaR*, and also *VaR* violations associated to any other coverage rate $\beta \in]0, 1[$, *i.e.* $I_{t-k}(\beta)$, $\forall k \neq 0$.

The first hypothesis is fully intuitive: for a $\alpha\%$ coverage level, the occurrence of losses exceeding this *VaR* must correspond to $\alpha\%$ of the total number of periods during which *VaR* is used as a measure of extreme risk. For a 5% *VaR* used as a reference measure over 1000 periods, the expected number of violations must be equal to 50. If the violation number is significantly higher or lower than 50, the *VaR* model is not valid. Indeed, if the probability associated to event $I_t(\alpha) = 1$, noted $\pi_t = \Pr [I_t(\alpha) = 1]$, assessed by the frequency of violations observed over a time T , *i.e.* $(1/T) \sum_{t=1}^T I_t(\alpha)$, is significantly lower (resp. higher) than the α nominal coverage rate, it shows an overestimation (resp. underestimation) of *VaR* and thus too few (resp. too many) violations.

The tests of unconditional coverage were initially developed by Kupiec (1995). They were set up within the framework of backtesting procedures. Today, they are at the heart of the main assessment procedures for *VaR* models.

However, the unconditional coverage property does not give any information about the temporal independence of violations. The independence property of violations is nevertheless an essential property because any measure of risk must adapt automatically and immediately to any new information which entails a new evolution in the dynamics of asset returns. A model which does not take this aspect into account might create successive violation clustering. Extreme losses can then lead to extreme losses; this situation generally leads to bankruptcy. So, there must not be any form of dependence in the violation sequence, whatever the coverage rates considered. Indeed, the pioneering works of Berkowitz and O' Brien (2002) show that the *VaR* models used by six big American commercial banks tend not only to be very conservative as regards risk, - *i.e.* they tend to overestimate the actual fractiles of conditional P&L distribution -, but also to lead to violation clusters highlighting their inability to forecast changes in volatility.

" We evaluate the VaR forecasts in several ways. First, the null hypothesis of 99 percent coverage rate is tested. 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, page 1094)

It is important to note that these two *VaR* properties are independent one from the other. At this point, if a *VaR* measure does not satisfy either of these two hypotheses, it must be considered as not valid (Christoffersen, 1998). For example, satisfying the hypothesis of unconditional coverage does not compensate for the possible existence of violations clusters nor the non compliance with the independence hypothesis. On the contrary, there is conditional efficiency when the *VaR* measure confirms the two of unconditional coverage and

independence hypotheses.

Let us finally indicate that these two hypotheses (unconditional coverage and independence) are satisfied when the process associated to *VaR* violations is a martingale difference (Berkowitz et al., 2005), that is when:

$$E [I_t(\alpha) - \alpha \mid \Omega_{t-1}] = 0 \quad (4)$$

where information set Ω_{t-1} can include not only past *VaR* violations defined for the $\alpha\%$ reference rate, *i.e.* $\{I_{t-1}(\alpha), I_{t-2}(\alpha), \dots\}$, but also any variable Z_{t-k} known at time $t-1$, such as past *VaR* levels, returns, but also the violations associated to any other coverage rate β , *i.e.* $\{I_{t-1}(\beta), I_{t-2}(\beta), \dots\}$. Let us keep in mind that the martingale difference property indeed implies that $\forall Z_{t-k} \in \Omega_{t-1}$, $E [(I_t(\alpha) - \alpha) \otimes Z_{t-k}] = 0$ and in particular that if $I_{t-k}(\beta) \in \Omega_{t-1}$, then

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

Here we find the independence property again, whereas the unconditional coverage property stems from the property of iterated expectations, because the null conditional moment $E [I_t(\alpha) - \alpha \mid \Omega_{t-1}] = 0$ implies the null unconditional moment and so consequently the equality $E [I_t(\alpha)] = \alpha$.

To date, there are two major forms of conditional efficiency tests in the literature. The first category regroups all the tests set 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 , *i.e.* $I_{t-k}(\alpha)$. These tests, notably those based on the assessment of the occurrence of a particular event in time (here the occurrence of an $\alpha\%$ *VaR* violation) correspond to the approach called ‘Event Probability Forecast’. On the contrary, the second category regroups all the tests which jointly verify the conditional efficiency property for all possible coverage rates. These tests are no longer restricted to the study of *VaR* for a coverage rate arbitrarily set at 5% for instance. With this approach, the aim is thoroughly to assess the P&L distribution. Naturally, if this density is adequate, the validity of the computation of these quantiles, and so of *VaR*,

for any coverage rate included between $[0, 1]$ is guaranteed. These tests correspond to an approach of 'Density Forecast Assessment' type. Now we are going to present the main tests developed in line with these two approaches.

2.1 Event Probability Forecast Approach

As we have just mentioned, tests using the Event Probability Forecast approach evaluate a model of *VaR* calculation for a particular nominal coverage rate. What matters then is to jointly test the unconditional coverage and independence hypotheses. In this context, the major difficulty consists in specifying the form of the dependence of $I_t(\alpha)$ processes under the alternative hypothesis. Various tests of conditional efficiency are associated to the various suggested specifications.

LR Test of Christoffersen (1998)

Christoffersen (1998) supposes that, under the alternative hypothesis of *VaR* inefficiency, the process of $I_t(\alpha)$ violations is modeled with a Markov chain whose matrix of transition probabilities is defined by:

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

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

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

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 the α , coverage rate, *i.e.* $\pi_t = \Pr [I_t(\alpha) = 1] = \alpha$. Furthermore, the probability of a violation at time t is independent from the state in $t-1$. A simple likelihood ratio statistic, denoted LR_{CC} , then allows to test the null hypothesis of conditional efficiency. Under H_0 , Christoffersen shows that:

$$LR_{CC} = -2 \left\{ \ln L [\Pi_\alpha, I_1(\alpha), \dots, I_T(\alpha)] - \ln L [\hat{\Pi}, I_1(\alpha), \dots, I_T(\alpha)] \right\} \xrightarrow[T \rightarrow \infty]{L} \chi^2(2)$$

where $\hat{\Pi}$ is the maximum likelihood estimator of the transition matrix under the alternative hypothesis and where $\ln L [\Pi, I_1(\alpha), \dots, I_T(\alpha)]$ is the log-likelihood of violations $I_t(\alpha)$ associated to a transition matrix Π , such as:

$$L [\Pi, I_1(\alpha), \dots, I_T(\alpha)] = (1 - \pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1 - \pi_{11})^{n_{10}} \pi_{11}^{n_{11}} \quad (8)$$

with n_{ij} the number of times we have $I_t(\alpha) = j$ and $I_{t-1}(\alpha) = i$.

While this test is easy to use, it seems rather limitative for two main reasons. Firstly, independence is tested against a very particular form which 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 that of any other exogenous variable. The tests recently proposed by Engle and Manganelli (2004) and Patton (2002) overcome these two drawbacks.

DQ Test of Engle and Manganelli (2004)

Engle and Manganelli (2004) suggest using a linear regression model linking current violations to past violations so as to test the conditional efficiency hypothesis. Let $Hit(\alpha) = I_t(\alpha) - \alpha$, be the de-meanded process on α associated to $I_t(\alpha)$:

$$Hit_t(\alpha) = \begin{cases} 1 - \alpha & \text{if } r_t < VaR_{t|t-1}(\alpha) \\ -\alpha & \text{else} \end{cases} \quad (9)$$

Let us consider the following linear regression model:

$$Hit_t(\alpha) = \delta + \sum_{k=1}^K \beta_k Hit_{t-k}(\alpha) + \sum_{k=1}^K \gamma_k g[Hit_{t-k}(\alpha), Hit_{t-k-1}(\alpha), \dots, z_{t-k}, z_{t-k-1}, \dots] + \varepsilon_t \quad (10)$$

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} . For example we can consider past returns r_{t-k} , the square of past returns r_{t-k}^2 , the values of *VaR* forecast, $VaR_{t-k|t-k-1}(\alpha)$ or also implicit volatility data. But, whatever the chosen specification, the null hypothesis test of conditional efficiency corresponds to testing the joint nullity of coefficients β_k and γ_k and of constant δ :

$$H_0 : \delta = \beta_k = \gamma_k = 0, \quad \forall k = 1, \dots, K \quad (11)$$

The current *VaR* violations are uncorrelated to past violations since $\beta_k = \gamma_k = 0$ (consequence of the independence hypothesis), whereas the unconditional coverage hypothesis is verified when δ is null. Indeed, under the null hypothesis, $E[Hit_t(\alpha)] = E(\varepsilon_t) = 0$, which implies by definition that $\Pr[I_t(\alpha) = 1] = E[I_t(\alpha)] = \alpha$. The joint nullity test of all coefficients, including the constant, thus corresponds to a conditional efficiency test. A LR statistic or a Wald statistic can easily be used to test the simultaneous nullity of these coefficients. So, if we denote $\Psi = (\delta \ \beta_1 \ \dots \ \beta_K \ \gamma_1 \ \dots \ \gamma_K)'$ the vector of the $2K + 1$ parameters in this model and Z the matrix of explanatory variables of model (10), the Wald statistic, noted DQ_{CC} , in association with the test of conditional efficiency hypothesis then verifies³:

$$DQ_{CC} = \frac{\widehat{\Psi}' Z' Z \widehat{\Psi}}{\alpha(1-\alpha)} \xrightarrow[T \rightarrow \infty]{L} \chi^2(2K+1) \quad (12)$$

A natural extension to the test of Engle and Manganelli (2004) simply consists in considering a (probit or logit) binary model linking current violations to past ones. Indeed, it is accepted that the linear regression model is not the most adapted model when, as is case here, the dependent variable is a binary

³ Under the null hypothesis, ε_t corresponds to the violation process $Hit_t(\alpha)$ which follows a Bernouilli distribution of parameters α and $\alpha(1-\alpha)$.

one (Gourieroux, 2000). So, Patton (2002) proposes a LR test based on a logit model linking the violation probability at time t to a set of explanatory variables Z_t (possibly including past VaR violations).

Martingale difference test by Berkowitz et al. (2005)

The objective of Berkowitz et al. (2005) is to propose a unified approach of VaR assessment. They start from the fact that the unconditional coverage and independence hypotheses are nothing but consequences of the martingale difference hypothesis of the $Hit(\alpha) = I_t(\alpha) - \alpha$ process. In this perspective, several tests of the martingale difference hypothesis can be used to test the validity of VaR models for a given coverage rate α . On the basis of various tests listed by Durlauf (1991), the authors particularly focus on tests based on the examination of the spectral density of $Hit(\alpha)$, but also on the univariate Ljung-Box test which makes it possible to test the absence of autocorrelation in the $Hit(\alpha)$ sequence. For the Ljung-Box test, the statistic associated to the nullity test of the first K auto correlations of the violation process verifies:

$$LB(K) = T(T+2) \sum_{i=1}^K \frac{\hat{r}_i^2}{T-i} \xrightarrow[T \rightarrow \infty]{L} \chi^2(K) \quad (13)$$

where \hat{r}_i is the empirical autocorrelation of order i of the $Hit(\alpha)$ process. The Monte Carlo simulations made by the authors show that this test has good properties at finite distance if $K > 1$ ($K = 5$ in their simulations). This conclusion highlights the restrictive character of the Christoffersen (1998) test which takes into account only the auto correlation of order 1 in the Hit sequence function.

Other tests belonging to the Event Probability Forecast Approach can finally be mentioned. In particular, the independence tests based on the modeling of duration between two violations, such as that of Christoffersen and Pelletier (2004), allow us to consider wider dependences than those chosen under the Markov chain hypothesis or within the frame of the linear probability model. However, the test logic remains unchanged and consists in putting the conditional coverage hypothesis to the test, for a given coverage level.

2.2 Density Forecast Approach

The tests mentioned above deal with conditional efficiency for a given nominal coverage rate α . Now, the property of conditional efficiency must be valid for any coverage rate. If a model of *VaR* calculation leads to independent violations for a 1% coverage rate, but leads to violation clusters for a 5% coverage rate, it can not be considered as valid. This reasoning, pushed to the limit, then leads to testing conditional efficiency for all possible coverage rates between zero and one. This is the basic principle exploited by tests based on the Density Forecast Approach (Crnkovic and Drachman, 1997; Diebold et al., 1998; Berkowitz, 2001). Assessing the P&L distribution forecast thus means evaluating the probability of violation occurrence, $\Pr [I_t(\alpha) = 1]$, obtained for all values $\alpha \in]0, 1[$.

These tests of assessment of forecast density use in particular the transformation⁴ of Rosenblatt (1952) known as *P.I.T.* (*Probability Integral Transformation*). Suppose r_t is the observed return of an asset or of an asset portfolio, and suppose $F_{t-1}(\cdot)$ is the expected distribution function for this return. Let us keep in mind that this distribution function notably determines the forecast *VaR* value of the portfolio such as $VaR_{t|t-1}(\alpha) = F_{t-1}^{-1}(\alpha)$. Under the null hypothesis, this distribution function corresponds to the *ex-post* return distribution (which implies that the *VaR* calculation model is a "good" model), the Rosenblatt transformation then implies that:

$$X_t = F_{t-1}(r_t) \sim i.i.d. U_{[0,1]} \quad (14)$$

Testing the validity of the *VaR* model corresponds to testing this hypothesis. As Berkowitz (2001) underlines:

"Therefore, if banks are required to regularly report forecast distributions

⁴ 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 $[0, 1]$.

$\hat{F}(\cdot)$, 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 $\hat{F}(\cdot)$ changes over time”, (Berkowitz, 2001, page 7).

Given this general principle, different techniques can be used, to test independence and/or uniformity. So Crnkovic and Drachman (1997) suggest using Kuiper statistics to test uniformity. As for Diebold, Gunther and Tay (1998), they suggest the use of non parametric tests (Kolmogorov-Smirnoff, Cramer-Von Mises) to evaluate the significance of the distance between the transformed serie and theoretical distribution $U_{[0,1]}$. The major problem with these non parametric tests is that they require many observations in order to reach a reasonable power level. That is why Berkowitz (2001) proposes a parametric test which is based on another transformation of the variable X_t of equation (14). Indeed, under the null hypothesis, that reserved model is adequate, variable X_t is *i.i.d.* Consequently if we denote $\Phi^{-1}(\cdot)$ the inverse of the cumulative distribution function of the standard normal distribution, we have:

$$Z_t = \Phi^{-1}(X_t) = \Phi^{-1}[F_{t-1}(r_t)] \sim i.i.d. N(0, 1) \quad (15)$$

Berkowitz then proposes a variety of likelihood ratio tests which make it possible to test the independence and/or uniformity of the distribution X_t , by way of transformed variable Z_t . One possible approach consists in testing $H_0 : Z_t \sim i.i.d N(0, 1)$ against an alternative of this type:

$$H_1 : Z_t = \mu + \rho_1 Z_{t-1} + \dots + \rho_n Z_{t-n} + \gamma_1 Z_{t-1}^2 + \dots + \gamma_m Z_{t-m}^2 + \mu_t \quad (16)$$

So the null hypothesis implies $n + m + 2$ constraints, *i.e.* $\mu = \rho_1 = \dots = \rho_n = \gamma_1 = \dots = \gamma_m = 0$ and $\sigma_{Z_t} = 1$. From various Monte Carlo simulation exercises, Berkowitz shows that the LR test associated to H_0 is a powerful test even with sample sizes as small as 100.

However, even though this approach and the contribution of Berkowitz (2001) seem attractive, its transposition in the field of risk management seems a lit-

tle delicate. Indeed, what matters most in portfolio management is the precise measure of extreme risks associated to wide negative variations of the portfolio's market value. So it is the tail of P&L distribution which matters most. On the contrary, the density forecast approach similarly takes into account an error on the tail of P&L distribution as an error near the average of returns. Thus an internal model of VaR computation may be wrongly rejected because it does not seize with accuracy the density of the generative process around the average of returns while at the level of the tails of distribution this model gives a good fit of extreme risks. Berkowitz (2001), aware of this limit, proposed a test of likelihood ratio named LR_{tail} adapted to the assessment of the models of *Expected Shortfall* type. Crnkovic and Drachman (1997) also proposed a modification of the Kuiper statistics. But from this viewpoint it soon becomes very difficult and relatively inappropriate to adapt this approach to the assessment of the return distribution only on the tail of distribution associated to extreme losses. That is why we propose a different method in this paper. Rather than adapting a test with the '*Density Forecast Evaluation*' approach to the assessment of the tail of return distribution only, we suggest adapting a test of the '*Event Probability Forecast Evaluation*' approach to assess the validity of VaR for a relevant set of coverage rates.

3 A Multivariate Portmanteau Statistic

We propose here to expand on the multivariate case the test of absence of autocorrelation of violations proposed by Berkowitz et al. (2005). Our test, based on a multivariate Portmanteau statistic enables us to jointly test the property of conditional coverage for various relevant rates of coverage. So it exploits a larger information set than those generally used in tests of the '*Event Probability Forecast Evaluation*' category, yet without the drawbacks of tests based on the assessment of return density.

Formally, we denote $VaR_{t|t-1}(\alpha)$, VaR at time t , for a coverage rate of $\alpha\%$, anticipated conditionally in a set of information Ω_{t-1} . Suppose $Hit_t(\alpha)$ is the

indicator value $1 - \alpha$ in case of violations and $-\alpha$ otherwise:

$$Hit_t(\alpha) = \begin{cases} 1 - \alpha & \text{if } r_t < VaR_{t|t-1}(\alpha) \\ -\alpha & \text{else} \end{cases} \quad (17)$$

The martingale difference hypothesis as formulated by Berkowitz et al. (2005), *i.e.* $E[Hit_t(\alpha) | \Omega_{t-1}] = 0$, implies in particular (property of iterated expectancy) that for coverage rate α :

$$E[Hit_t(\alpha) Hit_{t-k}(\alpha)] = 0 \quad \forall k \in \mathbb{N}^* \quad (18)$$

As we said above, this hypothesis also implies the independence of the violation processes associated to two different coverage rates, α and β :

$$E[Hit_t(\alpha) Hit_{t-k}(\beta)] = 0 \quad \forall k \in \mathbb{N}^*, \quad \forall(\alpha, \beta) \quad (19)$$

This latter property of *VaR* violations is the basis of the multivariate extension of the Portmanteau test we propose. Our idea simply consists in testing the validity of the *VaR* determination model for a finite sample by various coverage rates considered as relevant *a priori* for risk management. To do this, we suggest building a multivariate test statistic which will test the absence of autocorrelation of the violation processes associated to different coverage rates. This statistic takes into account both violation autocorrelations for a finite sample of a given coverage rate, and also 'combined' autocorrelations between violations obtained for different coverage rates. Furthermore, the use of a multivariate test rather than of a group of univariate tests allows us to control the size of the test accurately.

Let $\Theta = \{\theta_1, \dots, \theta_m\}$ be a discrete set of m different coverage rates, strictly between 0 and 1 and considered as relevant for risk analysis. Let $Hit_t = [Hit_t(\theta_1) : Hit_t(\theta_2) : \dots : Hit_t(\theta_m)]'$ be the vector of dimension $(m, 1)$ regrouping the violation sequences associated to these m coverage rates, at time $t, \theta_1, \dots, \theta_m$. The conditional efficiency hypothesis for the vectorial process im-

plies that:

$$\text{cov}(Hit_t, Hit_{t-k}) = E[Hit_t Hit'_{t-k}] = V * \delta_k \quad (20)$$

where V is a (m, m) symmetric non zero matrix and where δ_k is a scalar⁵:

$$\delta_k = \begin{cases} 1 & \text{si } k = 0 \\ 0 & \text{else} \end{cases} \quad (21)$$

The practical application of our conditional efficiency test consists, for a given order $K \geq 1$, in testing the null hypothesis corresponding to the joint nullity by the autocorrelations of order 1 in K for the vectorial process Hit_t :

$$H_0 : \text{cov}(Hit_t, Hit_{t-k}) = V * \delta_k \quad \forall k = 1, \dots, K \quad (22)$$

Or in an equivalent way:

$$H_0 : E[Hit_t(\theta_i) Hit'_{t-k}(\theta_j)] = 0 \quad \forall k = 1, \dots, K, \forall (\theta_i, \theta_j) \in \Theta \quad (23)$$

This test is nothing but a multivariate extension of common Portmanteau tests (Box-Pierce or Ljung-Box). To implement this test, several multivariate Portmanteau statistics can be used (Chitturi, 1974; Hosking, 1980; Li and McLeod, 1981). Let us call \hat{C}_k the matrix of empirical covariance associated to vector Hit_t :

$$\hat{C}_k = (\hat{c}_{ijk}) = \sum_{t=k+1}^T Hit_t Hit'_{t-k} \quad \forall k \in N^* \quad (24)$$

We assert $\hat{R}_k = D\hat{C}_k D$ where D is the diagonal matrix having as its constituents, standard deviations associated to univariate processes $Hit_t(\theta_i)$ defined by $\sqrt{\hat{c}_{ii0}}$, for $i = 1, \dots, m$. We denote $Q_m(K)$ the multivariate Portman-

⁵ It is also possible to test the independence hypothesis by defining covariance in the following way: $H_0 : \text{cov}(Hit_t, Hit_{t-k}) = E\{[Hit_t - E(Hit_t)][Hit_{t-k} - E(Hit_{t-k})]\} = V * \delta_k$.

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

teau statistic proposed by Li and McLeod (1981)⁶:

$$Q_m(K) = T \sum_{k=1}^K (\text{vec} \hat{R}_k)' \left(\hat{R}_0^{-1} \otimes \hat{R}_0^{-1} \right) (\text{vec} \hat{R}_k) \quad (25)$$

Under the null hypothesis of absence of autocorrelation in the vector Hit_t , Li and McLeod (1981) prove that:

$$Q_m(K) \xrightarrow[T \rightarrow \infty]{L} \chi^2(Km^2) \quad (26)$$

Two points should be underlined concerning the choice of K and m respectively. First of all, as regards number m of coverage rates, it is clear that our method implies the use of at least two coverage rates ($m \geq 2$). Furthermore, coverage rates θ_j , $j = 1, \dots, m$ must be relatively small to test *VaR* validity only in the low tail of the P&L distribution. However, using a high value for m has two drawbacks. The first one is purely mathematical and concerns the computation of the $Q_m(K)$ statistic. Indeed, $Q_m(K)$ cannot be calculated if matrix \hat{R}_0 is singular. Now the probability of this matrix being singular increases when we use very close coverage rates, *i.e.* when we increase m while limiting rates θ_j to a given interval. As an example, let us suppose that Θ is such as $\theta_{min} = 1\%$ and $\theta_{max} = 5\%$. If we take nine coverage rates uniformly distributed in this interval, that is $\Theta = \{1\%, 1.5\%, 2\%, 2.5\%, 3\%, 3.5\%, 4\%, 4.5\%, 5\%\}$, 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), and so it will imply a singularity of \hat{R}_0 . In other words, the higher number m of coverage rates, the closer we get to a Density forecast approach when using a test statistic which is not adapted to this type of problem. That is why we have to choose m very carefully: as we shall see afterwards, taking more coverage rates into account improves the power of our test, but beyond a certain level a higher m makes our approach based on an Event Probability Forecast Approach statistic irrelevant. That is why we suggest using at most three coverage rates m . For $m = 2$, we suggest considering the set $\Theta = \{1\%, 5\%\}$

⁶ Let us note that if $m = 1$ the Li and McLeod statistic (1981) is equal to that of Box and Pierce (1970).

and for $m = 3$ the set $\Theta = \{1\%, 5\%, 10\%\}$. These relatively distant nominal coverage rates, generally avoid the singularity of matrix \hat{R}_0 and correspond to common coverage rates used in risk management.

As regards the choice of lag order, Hosking (1980) shows that K has to be $O(T^{1/2})$ to insure the asymptotic convergence of the statistic. However, Bender and Grouven (1993), through simulations, highlight the dependence of K not only on T , but also on m . That is why, to determine the choice of the acceptable values of lag order K , we use the method of Bender and Grouven (1993). We set sample size T at 250, 500, 750 and 1000. For every couple left (T, m) , with $m = 2, 3$, we simulate a process of multivariate white noise of dimension (T, m) . The statistic of Li and McLeod (1981) is then calculated for various values of K . We consider 14 different values for K from 1 to 50. We repeat the simulation 1000 times, which finally gives 14 series of length 1000 for every couple (T, m) . These series are in theory asymptotically distributed according to $\chi^2(Km^2)$. For each series, the Kolmogorov-Smirnov (K-S) test is run to assess the equivalence with the theoretical distribution. P-values of the K-S test are shown for every triplet (T, m, K) in Table 1 of the Appendix. 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. Indeed we observe a strong rejection of the null hypothesis of test K-S. Thus, for these sample sizes, we suggest the following values for K , $K \in \{1, 2, 3, 4, 5\}$.

4 Short Sample Properties

In this section, we first study the finite sample properties of our test to characterize the influence of the choice of both lag order K and number m of coverage rates. Then, in a second time, we will propose a comparison of sizes and powers of our test with the main conditional coverage tests mentioned in the second section.

Empirical Size of the $Q_m(K)$ test

In the literature on *VaR* assessment, we generally observe (Cf. Berkowitz, 2001 and Berkowitz et al., 2005) an important deformation of the empirical size of conditional coverage tests, in particular for 1% coverage rate (the value recommended by supervision authorities). That is why we suggest assessing first the empirical size of our new test for various values of parameters m and K . To do this, we simulate the P&L distribution from an EGARCH process. Here we consider the calibration proposed by Campbell (2005).

$$r_{t,t-1} = v_{t,t-1} \quad (27)$$

$$v_{t,t-1} \sim N(0, \sigma_t^2) \quad (28)$$

$$\ln(\sigma_t^2) = 0.02 + 0.94 \ln(\sigma_{t-1}^2) + 0.22 \left| \frac{v_{t,t-1}}{\sigma_{t-1}} \right| - 0.05 \frac{v_{t,t-1}}{\sigma_{t-1}} \quad (29)$$

Campbell shows that such a model duplicates perfectly the dynamics of monthly returns for several American indexes over a relatively long period (1927-1998).

The empirical size of our test is then assessed by using a method of *VaR* computation in which the true dynamics of asset profitability is known. Thus, *VaR* is computed for a given coverage rate from the conditional variance determined by equation (29)⁷. The series of *out-of-sample VaR* are generated for sample sizes $T = 250, 500, 750, 1000$. From the *VaR* violation sequences observed *ex-post* (or Hit functions), test statistic $Q_m(K)$ is computed for a relatively high number of simulations. Here we consider 10 000 simulations. Empirical size then corresponds to the frequency of rejection of the conditional coverage hypothesis observed in these simulations. If the asymptotic distribution of our test is adequate, this refusal frequency must be close to the nominal size set at 10% in our experiments.

Table 2 presents the empirical size of conditional coverage test $Q_m(K)$ for different sample sizes T , different lag orders K and different dimensions m of our multivariate statistic. The first part of Table 2 shows the empirical size

⁷ The VaR is then equal to:

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

of the Ljung-Box test for 1% VaR . The following parts show the empirical size of the Multivariate Portmanteau test respectively for $m = 2$ coverage rates ($m = 3$ coverage rates). The next two parts of Table 2 indicate that the shift to multivariate frame (taken into account by 2 or 3 different coverage rates) leads to a slight increase in our test's empirical size. Whatever the sample size and the value of K , we note a stabilization of the empirical size around 14% on average for a 10% asymptotic size. Thus the test proposed in this paper is slightly oversized. So, considering the fact that for small-sized samples around 250, the univariate Portmanteau test is slightly undersized whereas our multivariate test appears slightly oversized, it is not appropriate to directly compare⁸ the powers of these two tests to show the contribution of our multivariate extension. That is why we will choose the Dufour (2004) method which authorizes the calculation of a test's empirical power at a finite distance, while respecting nominal size independently from the number of simulations. This method can be found in Christoffersen and Pelletier (2004) and also in Berkowitz et al. (2005).

Finite Sample Power of the $Q_m(K)$ test

We propose two different exercises to assess the power of our conditional coverage test. The first exercise, similar to that of Berkowitz and al is based on the use of Historic Simulation (HS) for VaR calculation. By definition, VaR calculated at time $t - 1$ and summarizing potential maximum loss for the following period (with probability $1 - \alpha$) is the empirical α -quantile of past returns observed on the last Te periods. Here we set Te at 250. Formally, VaR is defined by the following relation:

$$VaR_{t|t-1}(\alpha) = percentile\left(\{R_j\}_{j=t-Te}^{t-1}, 100\alpha\right) \quad (30)$$

⁸ This direct comparison of test powers would then be to the detriment of the test with the lowest empirical size, that is to the detriment of the test of Berkowitz et al. (2005). The results of these exercises of power comparison are available on request.

The second exercise is based on the use of the delta normal method. In that case, VaR is defined by the following equality:

$$VaR_{t|t-1}(\alpha) = \Phi^{-1}(\alpha) \left[\text{Variance} \left(\{R_j\}_{j=t-T_e}^{t-1} \right) \right]^{\frac{1}{2}} \quad (31)$$

The choice of these two VaR calculation methods to quantify the test power is not neutral. Indeed, it is necessary to choose VaR calculation methods which are not adapted to the P&L distribution and therefore violate efficiency, *i.e.* the nominal coverage \ independence hypothesis. Here Historical Simulation and the Delta Normal methods seem to be sensible choices as they generate VaR violations clusters as illustrated in Figures 1 and 2. These graphs compare observed returns (equation 29) with VaR obtained via the two computation methods for an EGARCH simulation. We notice at once that violation clusters are quite graphic, whether for 1% VaR or for 5% VaR .

We use the Dufour (2004) methodology for a comparison of power independent of size. Let us have S a statistic of the test of continuous survival function⁹ $G(\cdot)$ such as $Prob[S_i = S_j] = 0$. Theoretical P-value $G(\cdot)$ can be approximated by its empirical counterpart: $\hat{G}_M(x) = 1/M \sum_{i=1}^M \mathbb{I}(S_i \geq x)$ where $\mathbb{I}(\cdot)$ is the indicator function. S_i is the test statistic for a sample simulated under the null hypothesis. Dufour (2004) demonstrates that if M is big enough, whatever the value of S_0 , theoretical critical region $G(S_0) < a$, with a , the asymptotic nominal size, is equivalent to the critical region $\hat{p}_M(S_0) \leq a_1$, with

$$\hat{p}_M(S_0) = \frac{M \hat{G}_M(S_0) + 1}{M + 1} \quad (32)$$

and this $\forall a_1$. When $Prob[S_i = S_j] \neq 0$, or when it is possible for a given simulation of the test statistic (under H_0) to find the same value of S for two or more times, the function of empirical survival can be written as follows:

$$\tilde{G}_M(S_0) = 1 - \frac{1}{M} \sum_{i=1}^M \mathbb{I}(S_i \leq S_0) + \frac{1}{M} \sum_{i=1}^M \mathbb{I}(S_i = S_0) \times \mathbb{I}(U_i \geq U_0) \quad (33)$$

where U_i , $i = 0, 1, \dots, M$ correspond to realizations of a uniform $[0, 1]$ variable. So, to calculate the empirical power of test statistic $Q_m(K)$, we just need to

⁹ This is also called Tail area or p-value function.

simulate under H_0 , M independent realizations of our test statistic $Q_m^1(K)$, $Q_m^2(K)$, ..., $Q_m^M(K)$. If we note $Q_m^0(K)$ the test statistic obtained under H_1 (by using Historical Simulation or Delta Normal method), we reject H_0 if:

$$\hat{p}_M [Q_m^0(K)] = \frac{M \tilde{G}_M [Q_m^0(K)] + 1}{M + 1} \leq a \quad (34)$$

Statistic $Q_m^0(K)$ simulated N times under H_1 gives the test power, equal to the number of times when $\hat{p}_M [Q_m^0(K)]$ is inferior or equal to a , divided by N simulations (the test's nominal size is thus respected at finite distance). Here we set M at 10 001. This method makes it possible to compare the empirical power of several tests independently from the deformation of the empirical size. Naturally, such a correction increases (resp. decreases) the power of the undersized (resp. oversized) test (compared to non corrected power). So, in this case, the power of our multivariate test is at a disadvantage in comparison with that of the univariate test, since we have showed that it was slightly oversized for the different sample sizes considered.

Tables 3 and 4 report the power of tests $LB(K)$ and $Q_m(K)$ for Historical Simulation and for Delta Normal. As for the comparison of empirical sizes, the power of conditional coverage test $Q_m(K)$ is reported for different T sample sizes, different lag orders K and different dimensions m of our multivariate statistic. In the first part of the table we find the power of univariate test $LB(K)$. Recall that it only tests the *VaR* validity for one coverage rate ($m = 1$, $\alpha = 1\%$). The interest of our multivariate approach which consists in taking into account several coverage rates to test *VaR* validity is highlighted in the following parts of Tables 3 and 4. These sections show the empirical powers of the Multivariate Portmanteau test respectively for $m = 2$ coverage rates (1% and 5%) and $m = 3$ coverage rates (1%, 5% and 10%). Indeed we show that whatever lag order K , whatever sample size T , the shift from a univariate dimension to a multivariate dimension improves the power of the conditional coverage test very significantly. So for example, for a sample size of 250 points (that is the equivalent of a year of quotations) and for a lag order equal to 5, our test power is about 50 % when we consider two or three coverage rates, while it is only 32% in the univariate case.

Finite Sample Properties of Conditional Coverage tests

Now we compare the size and power properties of our test with those of the various conditional coverage tests surveyed in the first section. Let us consider two tests¹⁰: Christoffersen's (1998) LR_{CC} test and Engle and Manganelli's (2004) DQ_{CC} test. Table 5 shows the empirical size of the different conditional coverage tests for a 10% nominal size. Since the LR_{CC} and DQ_{CC} tests assess VaR validity for a given coverage rate, we report the results for three different coverage rates: 1%, 5% and 10%. As for our test statistic, it is given for two coverage rates (1% and 5%) then for three coverage rates (1%, 5% and 10%) with lag order $K = 5$ in each case. We can observe that the empirical size of Engle and Manganelli's test is always relatively close to nominal size and this, even for small samples and whatever the considered VaR . On the other hand, we note that for small samples, the Christoffersen test shows a significant size distortion: this test tends to be undersized for a 1% VaR and on the contrary oversized for VaR with a higher coverage rate (10% in particular).

Tables 6 and 7 show the empirical powers of all three tests. Powers are calculated using the Dufour methodology. We note that our test gives much better results than Christoffersen's LR_{CC} test (1998). The same is true when comparing with the DQ_{CC} test (especially for a 1% coverage rate, the value recommended by market supervision authorities), with a power improvement ranging from 12% to 29%. In addition to the power gain, it should be noticed that the null hypothesis in our new test is more general than that of LR_{CC} or DQ_{CC} tests.

5 Conclusion

The assessment of VaR models is a key issue in a situation where this measure is today a standard for risk management. In this work we propose a validation

¹⁰The powers comparison between our multivariate test and the Ljung-Box test of Berkowitz et al. (2005) has already been studied (see Tables 3 and 4).

test halfway between the 'Event Probability Forecast', and the 'Density Forecast Assessment' approaches. This test is easy to implement and more powerful than the major existing conditional coverage tests found in the literature. Besides it has the advantage of testing the validity of a *VaR* calculation model for a set of coverage rates considered as relevant for the analysis of financial risks. Our test can be expanded on several ways. The Engle and Manganelli test (2004) can be expanded to the multivariate frame by using a *VAR* (Vectorial AutoRegression) model. A multivariate Logit model also is a natural extension of our test.

A 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, 465-474.
- BERKOWITZ, J., AND O'BRIEN J. , 2002. How Accurate are the Value-at-Risk Models at Commercial Banks. *Journal of Finance*, 57, 1093-1111.
- 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, 1509-1526.
- CAMPBELL, S. D., 2005. A Review of Backtesting and Backtesting Procedures. Working Paper, 2005-21, *Finance and Economics Discussion Series*.
- CHITTURI, R. V., 1974. Distributions of Residual Autocorrelations in Multiple Autoregressive Schemes. *Journal of the American Statistical Association*, 69, 928-934.
- CHRISTOFFERSEN, P. F., 1998. Evaluating Interval Forecasts. *International Economic Review*, 39, 841-862.

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., 2003. Evaluating the Survey of Professional Forecasters Probability Distributions of Expected Inflation Based on Derived Event Probability Forecasts. *mimeo, Department of Economics, University of Warwick*.

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, 255-276.

CLEMENTS, M. P., AND TAYLOR, N., 2003. Evaluating Prediction Intervals for High-Frequency Data. *Journal of Applied Econometrics*, 18, 445-456.

CRNKOVIC, C., AND DRACHMAN, J. (1997), *Quality Control in VaR: Understanding and Applying Value-at-Risk*, London, Risk Publications.

DIEBOLD, F. X., GUNTHER, T. A. AND TAY, A. S., 1998. Evaluating Density Forecasts. *International Economic Review*, 39, 863-883.

DOWD, K., 2005. *Measuring market risk*, John Wiley & Sons Ltd.

DUFOUR, J. M., 2004. Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics. Forthcoming in the *Journal of Econometrics*.

DURLAUF, S. N., 1991. Spectral Based Testing of the Martingale Hypothesis *Journal of Econometrics*, 50, 355-376.

ENGLE, R. F., AND MANGANELLI, S., 2004. CAViaR: Conditional Autoregressive Value-at-Risk by Regression Quantiles. *Journal of Business and Economic Statistics*, 22, 367-381.

GOURIEROUX, 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, 602-608.

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*, serie B, 43, 231-239.

PATTON, A. J., 2002. Application of Copula Theory in Financial Econometrics. *Ph.D. Dissertation*, submitted in Partial Satisfaction.

Figure 1: P&L Simulated Distribution and HS VaR

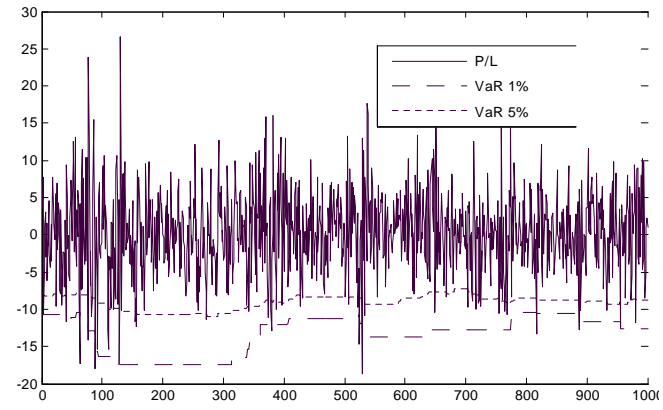


Figure 2: P&L Simulated Distribution and Delta Normal VaR

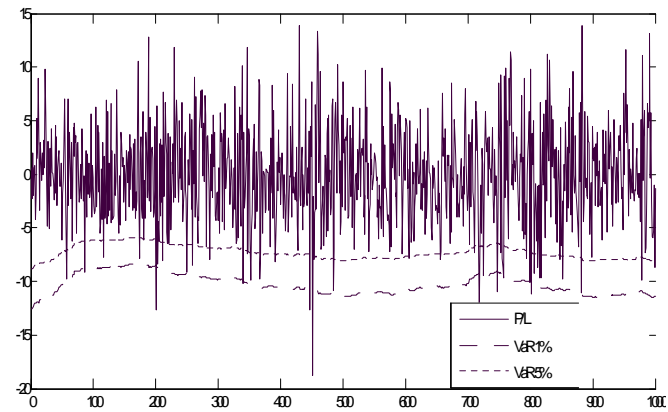


Table 1: P-values of Kolmogorov-Smirnov Tests¹⁴

$m = 2$							
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 10$	$K = 15$
$T = 250$	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
$T = 1000$	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
$T = 1000$	0.2043	0.5653	0.7909	0.1506	0.3675	0.0473	0.0013
$m = 2$							
	$K = 20$	$K = 25$	$K = 30$	$K = 35$	$K = 40$	$K = 45$	$K = 50$
$T = 250$	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
$T = 1000$	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
$T = 1000$	0.0006	0.0000	0.0187	0.0000	0.0000	0.0000	0.0000

Notes: K denotes the lag order of the Portmanteau test, T denotes the length of the VaR sample, and m corresponds to the number of VaRs included in the multivariate Portmanteau test. For each configuration the p -value of the KS adequation with the chi-squared distribution is reported. P-values are computed with 1000 simulations.

Table 2: Actual Sizes of $LB(K)$ and $Q_m(K)$ Tests¹⁵

$LB(K), \Theta = \{1\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.0801	0.1192	0.1456	0.1607	0.1407
$Q_2(K), \Theta = \{1\%, 5\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.1246	0.1368	0.1430	0.1490	0.1457
$Q_3(K), \Theta = \{1\%, 5\%, 10\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.1286	0.1428	0.1323	0.1377	0.1419

Notes: For each simulation, the profit and loss (P&L) distribution is generated under 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 size of the Ljung Box Test and the Multivariate Portmanteau test corresponds to the frequency of rejection of the null 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 VaRs included in the multivariate Portmanteau test. Nominal size is 10%

Table 3: Empirical Power of $LB(K)$ and $Q_m(K)$ Tests
Case 1: Historical Simulation ¹⁶

$LB(K), \Theta = \{1\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.3910	0.4525	0.4883	0.4948	0.5123
$Q_2(K), \Theta = \{1\%, 5\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.5484	0.6604	0.7393	0.7896	0.8106
$Q_3(K), \Theta = \{1\%, 5\%, 10\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.5695	0.7121	0.7879	0.8321	0.8485

Notes: For each simulation, the profit and loss distribution (P&L) is generated under 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 of the P&L. It does not satisfy the independence assumption and /or the nominal coverage. The empirical power is computed using Dufour methodology as described above. K denotes the lag order of the Portmanteau test, T denotes the length of the VaR sample, and m corresponds to the number of VaRs included in the multivariate Portmanteau test. Nominal size is 10%

Table 4: Empirical Power of $LB(K)$ and $Q_m(K)$ Tests
Case 2: Delta Normal¹⁷

$LB(K), \Theta = \{1\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.5379	0.4766	0.5362	0.5447	0.5640
$Q_2(K), \Theta = \{1\%, 5\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.5436	0.6826	0.7467	0.7879	0.8165
$Q_3(K), \Theta = \{1\%, 5\%, 10\%\}$					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$T = 250$	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
$T = 1000$	0.5942	0.7192	0.7894	0.8394	0.8625

Notes: For each simulation, the profit and loss distribution (P&L) is generated under EGARCH(1,1) distribution with normal disturbances. At each date, the Delta Normal VaR is simply computed using formula (31). It does not satisfy the independence assumption and /or the nominal coverage. The empirical power is computed using Dufour methodology as described above. K denotes the lag order of the Portmanteau test, T denotes the length of the VaR sample, and m corresponds to the number of VaRs included in the multivariate Portmanteau test. Nominal size is 10%

Table 5: Empirical sizes of LR_{CC} and DQ_{CC} Tests ¹⁸

$\alpha =$	LR_{CC}			DQ_{CC}		
	1%	5%	10%	1%	5%	10%
$T = 250$	0.0381	0.0826	0.1543	0.0757	0.0936	0.0947
$T = 500$	0.0637	0.0943	0.1083	0.0950	0.0921	0.0981
$T = 750$	0.0555	0.1090	0.1082	0.1113	0.0904	0.1017
$T = 1000$	0.0640	0.1623	0.0970	0.1293	0.0985	0.0977

Notes: For each simulation, the profit and loss distribution (P&L) is generated under 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 size of the tests corresponds to the frequency of rejection of the null obtained with 10 000 simulations. T denotes the length of the VaR sample. Nominal size is 10%

Table 6: Empirical Power of LR_{CC} , DQ_{CC} and $Q_m(K)$ Tests

Case 1: Historical Simulation ¹⁹

α	LR_{CC}			DQ_{CC}			$Q_2(5)$	$Q_3(5)$
	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	—	—
$T = 250$	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
$T = 1000$	0.1812	0.3058	0.3526	0.5604	0.7677	0.7685	0.8106	0.8485

Notes: For each simulation, the profit and loss distribution (P&L) is generated under 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 of the P&L. It does not satisfy the independence assumption and /or the nominal coverage. The empirical power is computed using Dufour methodology as described above. 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 VaRs included in the multivariate Portmanteau test. Nominal size is 10%

Table 7: Empirical Power of LR_{CC} , DQ_{CC} and $Q_m(K)$ Tests
Case 2: Delta Normal ²⁰

	LR_{CC}			DQ_{CC}			$Q_2(5)$	$Q_3(5)$
	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	—	—
$T = 250$	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
$T = 1000$	0.4426	0.3409	0.4347	0.5959	0.6209	0.6446	0.8165	0.8625

Notes: For each simulation, the profit and loss distribution (P&L) is generated under EGARCH(1,1) distribution with normal disturbances. At each date, the Delta Normal VaR is simply computed using formula (31). It does not satisfy the independence assumption and /or the nominal coverage. The empirical power is computed using Dufour methodology as described above. 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 VaRs included in the multivariate Portmanteau test. Nominal size is 10%