

Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Tests

Elena-Ivona Dumitrescu, Christophe Hurlin, Vinson Pham

▶ **To cite this version:**

Elena-Ivona Dumitrescu, Christophe Hurlin, Vinson Pham. Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Tests. 2012. halshs-00671658

HAL Id: halshs-00671658
https://shs.hal.science/halshs-00671658

Preprint submitted on 18 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Tests

Elena-Ivona Dumitrescu*, Christophe Hurlin†, and Vinson Pham‡

February 2012

Abstract

In this paper we propose a new tool for backtesting that examines the quality of Value-at-Risk (VaR) forecasts. To date, the most distinguished regression-based backtest, proposed by Engle and Manganelli (2004), relies on a linear model. However, in view of the dichotomic character of the series of violations, a non-linear model seems more appropriate. In this paper we thus propose a new tool for backtesting (denoted DB) based on a dynamic binary regression model. Our discrete-choice model, *e.g.* Probit, Logit, links the sequence of violations to a set of explanatory variables including the lagged VaR and the lagged violations in particular. It allows us to separately test the unconditional coverage, the independence and the conditional coverage hypotheses and it is easy to implement. Monte-Carlo experiments show that the DB test exhibits good small sample properties in realistic sample settings (5% coverage rate with estimation risk). An application on a portfolio composed of three assets included in the CAC40 market index is finally proposed.

- *Keywords*: Value-at-Risk; Risk Management; Dynamic Binary Choice Models
- *J.E.L Classification* : C22, C25, C52, G28

*Corresponding author: Maastricht University and University of Orléans (LEO, UMR CNRS 6221), Rue de Blois, BP 6739, 45067 Orléans Cedex 2, France. Email: elena.dumitrescu@univ-orleans.fr

†University of Orléans, (LEO, UMR CNRS 6221). Email: christophe.hurlin@univ-orleans.fr.

‡University of California at Santa Cruz (UCSA). Vinson Pham benefited from a grant from the European Program *Atlantis* AIME "Excellence in Mobility" for his visit at the University of Orléans.

1 Introduction

There is an intense academic debate on the validity of risk measures in general and on the validity of the Value-at-Risk (hereafter VaR) in particular. Indeed, this is a particular problem, since the VaR is not observable, and therefore we have to rely upon the analysis of the behaviour of the violations so as to test its validity. A violation is actually defined as a situation where the loss observed ex-post goes beyond the ex-ante value of the VaR in absolute value. A model is hence valid if the violation process satisfies the martingale difference hypothesis.

There are three main issues generally emphasized when one comes to evaluating VaR sequences. First, the power of the backtesting test, *the probability of rejecting a model that is not valid*, especially in small samples (250 to 500 observations, or, to put it differently, 1-2 years of VaR forecasts) plays a key role. It has been shown that generally these tests have low power, as the backtesting procedure is too optimistic in the sense that it does not reject the validity of a model as often as it should (see Hurlin and Tokpavi, 2008).

Second, the backtesting methodology has to be model-free. Indeed, the evaluation procedure must be implementable whatever the model used to generate the sequence of VaR, so as to reach a diagnostic regarding the validity of the VaR. Third, estimation risk must be taken into account. VaR series can be estimated using various models, some more, others less complicated, with a few or numerous parameters, according to the specific methodology of a certain financial institution. Testing procedures can thus successfully answer the question of VaR validity only by taking into account estimation error, as the risk of estimation error present in the estimates of the parameters pollutes VaR forecasts. Conditional on allowing for these errors, we should observe no particular orientation of the diagnostic of the backtest in the sense of under-rejecting or over-rejecting too often.

Various backtests have been proposed so as to satisfy these three requirements (high power, model-free, introduce estimation risk). They can be classified into four categories. First, in the pioneer works of Christoffersen (1998) the validity of VaR forecasts is tested through pa-

rameter restrictions on the transition probability matrix associated with a two-states Markov chain model (violation/no violation). To be more precise, two assumptions are derived from the martingale difference hypothesis, namely the unconditional coverage and the independence hypotheses. Second, tests relying on the duration between two consecutive violations are put forward by Christoffersen and Pelletier (2004), Haas (2005) and Candelon et al. (2008) in a likelihood-ratio framework. At the same time, the martingale difference assumption is tested directly by Berkowitz et al. (2011), Hurlin and Tokpavi (2007) or Perignon and Smith (2008). Last but not least, some tests are based on regression models (see Engle and Manganelli, 2004). The general idea is to project VaR violations onto a set of explanatory variables and subsequently test different restrictions on the parameters of the regression model, that correspond to the consequences of the martingale difference assumption. In such a context, both linear and non-linear regression models can be considered. For example, the recent paper of Gaglianone et al. (2011) proposes to evaluate the validity of the VaR by relying on quantile regression, which allows them to identify why and when a VaR model is misspecified.

Nevertheless, the most popular test of this category is Engle and Manganelli's Dynamic Quantile test (2004), hereafter DQ.[1] It consists in testing some linear restrictions in a linear model that links the violations to a set of explanatory variables. However, the dependent variable is by nature a binary one. It follows that linear regression models are not the most appropriate choice allowing to infer on the parameters and consequently on the hypothesis of validity of the VaR. The linear model has several shortcomings in this context. The innovations of the latent model are assumed to follow a discrete distribution. They are also heteroscedastic in a way that depends on the estimated parameters. At the same time, constraining the right part of the regression to the 0-1 interval implies negative variances and nonsense probabilities. Still, it is technically possible to test the significance of the slope parameters in the case of a binary dependent variable by relying on linear models (see Gourieroux, 2000).

In this paper we propose a new tool for backtesting VaR forecasts. Like Engle and Manganelli, we consider a regression model that links the violations to a set of explanatory variables.

[1]Note that the DQ backtest is not related to the quantile regression method used in the CAViaR method to forecast the VaR (Engle and Manganelli, 2004).

However, given the dichotomic character of the series of violations, we use a non- linear model and, more specifically, a Dynamic Binary (hereafter DB) regression model. The issue addressed in this paper is hence the improvement of the finite sample properties of the backtests, particularly the power of these tests, when using a link function that is more appropriate for the binary dimension of the regressand. Besides, these new tests are expected to be robust to estimation risk.

By proposing dynamic binary models, which rely on recent extensions advocated in the *Early Warning System* literature, the potential correlation between the violations (clusters) is taken into account in the estimation. Consequently, the tests used to assess the independence assumption for the violations and implicitly the ones testing the conditional coverage hypothesis are expected to exhibit higher power than the ones previously proposed in the literature. To be more precise, we propose seven different specifications, denoted by DB_1 to DB_7, inspired from the CAViaR specifications put forward by Engle and Manganelli (2004). The subspace of explanatory variables includes several lags of the violations series and of the VaR, to which the lagged index is added in view of the dynamic nature of the models. To test the accuracy of the VaR sequence, a two-step framework is thus implemented. First, the seven DB specifications are estimated by constraint maximum-likelihood (Kauppi and Saikonnen, 2008). Subsequently, likelihood-ratio statistics are used to assess the joint significance of the parameters and thus the validity of the VaR.

Note that this test has several advantages. First, it can be easily implemented. Second, it allows us to separately test the unconditional coverage, the independence and the conditional coverage hypotheses. Third, Monte-Carlo experiments show that by taking into account estimation risk, our conditional coverage test exhibits good finite sample properties in very small samples (250 observations) for a 5% coverage rate.

A main issue in VaR literature regards the consequences of the potential correlation amongst assets on the construction of risk measures. We thus propose to test the validity of the VaR obtained by estimating both multivariate models, *i.e.* models that take into account the correlation among assets and univariate models, *i.e.* models that do not care for the possible correlation

among assets. To achieve this aim, we consider a portfolio constituted from three assets included in the CAC40 market index for the period June 1, 2007 - June 1, 2009. Our backtest shows that the two approaches lead us to risk measures that are valid from the conditional coverage hypothesis viewpoint. These findings go along the lines of Berkowitz and O'Brien's diagnostic (2002).

The rest of this paper is organized as follows. Section 2 presents the testing framework. In section 3 the binary regression-based backtests are presented while in section 4 their small-sample properties are gauged. Section 5 reveals the main results of an empirical application on a three-asset illustrative portfolio.

2 Environment and testable hypotheses

Let us denote by r_t the return of an asset or of a portfolio at time t and by $VaR_{t|t-1}(\alpha)$ the *ex-ante* VaR for an $\alpha\%$ coverage rate forecast conditionally on an information set $\mathcal{F}_{t-1}$. Following the actuarial convention of a positive sign for a loss (see Gourieroux et al., 2000 and Scaillet, 2003, *inter alii*), the conditional VaR is actually defined as follows:

$$\Pr[r_t < -VaR_{t|t-1}(\alpha)] = \alpha, \ \ \forall t \in \mathbb{Z}. \tag{1}$$

Let $I_t(\alpha)$ be the binary variable associated with the *ex-post* observation of an $\alpha\%$ VaR violation at time t, *i.e.*:

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

As stressed by Christoffersen (1998), VaR forecasts are valid if and only if the violation process $I_t(\alpha)$ satisfies the following two hypotheses:

- The unconditional coverage (UC thereafter) hypothesis: the probability of an *ex-post* return exceeding the VaR forecast must be equal to the α coverage rate,

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

- The independence (IND thereafter) 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 $\alpha\%$ 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.

The UC hypothesis is a straightforward one. Indeed, if the frequency of violations observed over T periods is significantly lower (respectively higher) than the coverage rate α, then the VaR model overestimates (respectively underestimates) the true level of risk. However, the UC hypothesis shades no light on the possible dependence of VaR violations. Therefore, the independence property of violations is an essential one, 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. Consequently, there must be no dependence in the violations variable, whatever the coverage rate considered.

When the UC and IND hypotheses are simultaneously valid, VaR forecasts are said to have a correct conditional coverage (CC thereafter), and the VaR violation process is a martingale difference:

$$\mathbb{E}\left[I_t(\alpha) \mid \mathcal{F}_{t-1}\right] = \alpha. \tag{4}$$

This last property is at the core of most of the backtests for VaR models available in the literature (Christoffersen, 1998; Engle and Manganelli, 2004; Berkowitz et al., 2011; etc.).

Actually, this backtesting approach, based on the UC, IND and CC assumptions, can be considered as a special case of the event probability forecast evaluation and can be implemented through four main types of tests. To be more exact, the first backtesting tests (Kupiec, 1995; Christoffersen, 1998) were based on a Markov chain type model with two states (violation/no violation). In this framework, the UC, IND and CC assumptions are simply tested through parameter restrictions on the transition probability matrix associated with the Markov representation. Other tests (Christoffersen and Pelletier, 2004; Haas, 2005; Candelon et al., 2008) rely on the duration between two consecutive violations. Under the CC hypothesis, the dura-

tion variable follows a geometric distribution with parameter α. Exploiting this property, it is straightforward to develop a likelihood ratio (LR) test for the null hypothesis of conditional coverage. The general idea consists in specifying a lifetime distribution that nests the geometric, so that the memoryless property can be tested by means of LR tests. Third, some other tests check the martingale difference assumption directly, through its main correlation based implications. It is, for instance, the case of Berkowitz et al. (2011), Hurlin and Tokpavi (2007) or Perignon and Smith (2008). Finally, some tests are based on regression models. In this context, the most popular backtesting test, often called DQ (Dynamic Quantile) test, was proposed by Engle and Manganelli (2004). It is based on a linear regression model of the hits variable on a set of explanatory variables including a constant, the lagged values of the hit variable, and any function of the past information set suspected of being informative.

More formally, let us denote by $Hit_t(\alpha) = I_t(\alpha) - \alpha$ the demeaned process of violation, that takes the value $1-\alpha$ every times r_t is less than the *ex-ante* VaR and $-\alpha$ otherwise. From the definition of the VaR, the conditional expectation of $Hit_t(\alpha,)$ given the information known at $t-1$ must be zero. In particular, under the CC assumption, the variable $Hit_t(\alpha)$ must be uncorrelated with its own lagged values and with any other lagged variable (including past returns, past VaR, etc.), and its expected value must be equal to zero. To put it another way, the CC assumption can be tested in the following linear regression model:

$$\begin{aligned} Hit_t(\alpha) &= \delta + \sum_{k=1}^{K} \beta_k Hit_{t-k}(\alpha) \\ &\quad + \sum_{k=1}^{K} \gamma_k \; g\left[Hit_{t-k}(\alpha), Hit_{t-k-1}(\alpha), ..., z_{t-k}, z_{t-k-1}, ...\right] + \varepsilon_t, \end{aligned} \tag{5}$$

where ε_t is a discrete *i.i.d.* process and where $g(.)$ is a function of past violations and of the variables z_{t-k} belonging to the entire informational set available $\mathcal{F}_{t-1}$. To be more exact, we can consider variables like past returns r_{t-k}, the square of past returns r_{t-k}^2, the predicted Value-at-Risk $VaR_{t-k|t-k-1}(\alpha)$ or even implicit volatility data. Nevertheless, testing for the null hypothesis of conditional efficiency is equivalent to testing the joint nullity of the coefficients β_k, γ_k, $\forall k = 1, .., K$, and that of the intercept δ, independent of the VaR model specification

considered:

$$H_0 : \delta = \beta_1 = ... = \beta_k = \gamma_1 = ... = \gamma_k = 0, \qquad \forall k = 1, .., K \tag{6}$$

At the same time, present violations of the VaR are not correlated with past violations if $\beta_1 = ... = \beta_k = \gamma_1 = ... = \gamma_k = 0$ (as an implication of the independence hypothesis), while the unconditional coverage hypothesis is fulfilled if the constant δ is null. Indeed, under the null hypothesis, $\mathbb{E}\left[Hit_t\left(\alpha\right)\right] = \mathbb{E}\left(\varepsilon_t\right) = 0$, which means that by definition $\Pr\left[I_t\left(\alpha\right) = 1\right] = \mathbb{E}\left[I_t\left(\alpha\right)\right] = \alpha$. All in all, it results that the test of joint nullity of all coefficients corresponds to a conditional efficiency test.

Therefore, if we denote by $\Psi = (\delta,\ \beta_1,\ ..,\beta_K,\ \gamma_1,\ ..,\gamma_K)'$ the vector of the $2K+1$ parameters of the model and by Z the matrix of explanatory variables of model (5), the test statistic designated by DQ_{CC} and associated with the conditional coverage test[2] satisfies the following relation:

$$DQ_{CC} = \frac{\widehat{\Psi}' Z' Z \widehat{\Psi}}{\alpha\left(1-\alpha\right)} \xrightarrow[T\to\infty]{L} \chi^2\left(2K+1\right) \tag{7}$$

Similar to Christoffersen's test, the DQ test can be decomposed so that we can verify, for example, only the hypothesis of independence of violations. The test statistic DQ_{IND} associated with the independence hypothesis $H_0 : \beta_k = \gamma_k = 0$ satisfies:

$$DQ_{IND} = \frac{\widehat{\Psi}' R' \left[R\left(Z'Z\right)^{-1} R'\right]^{-1} R\widehat{\Psi}}{\alpha\left(1-\alpha\right)} \xrightarrow[T\to\infty]{L} \chi^2\left(2K\right) \tag{8}$$

where $R = [0 : I_{2K}]$ is a $(2K+1, 2K)$ matrix so that $R\Psi = \beta$, where $\beta = (\beta_1,\ ..,\beta_K,\ \gamma_1,\ ..,\gamma_K)'$.

3 A Dynamic Binary Response model

As aforementioned, Engle and Manganelli (2004) implement a test based on a simple linear regression which links the current hit (violation) to the lagged violations. Nevertheless, it is well known that this type of model is not suitable for dichotomous dependent variables (Gourieroux, 2000) since the residuals follow a discrete distribution and they are heteroscedastic by construc-

[2]Under the null hypothesis, the residuals ε_t correspond to the violation process $Hit_t\left(\alpha\right)$, which follows a Bernouilli distribution of parameter α and variance $\alpha\left(1-\alpha\right)$.

tion. Additionally, the scatter plot of the observations cannot be correctly adjusted with a single line as in the case of continuous variables.

A solution consists in using non-linear regression models suited for limited-dependent variables. Indeed, we propose a dichotomic model (logit, probit, cloglog,...) to establish a link between the current violations of VaR and the set of explanatory variables. We actually expect that the use of this appropriate non-linear methodology for binary dependent variables in the backtesting regression will improve the finite sample properties of the backtesting tests.

It should be noted that the backtesting VaR topic resembles at different levels the *Early Warning Systems* methodology which aims at forecasting financial crises or recessions. First, the object of the study has the same form, since in both cases the dependent variable is binary, *e.g.* the loss is greater then the VaR at time t or not; there is a crisis in a certain period or not. It follows that a special econometric methodology must be implemented so as to obtain consistent estimators. Furthermore, both events (loss or crisis) can be persistent as it is not always possible to shift from loss to profit or from crisis to calm in just one period. This dynamics should be taken into account in a VaR backtest regression models as it has been recently done in *Early Warning Systems* by Kauppi and Saikonnen (2008). Therefore, we explicitly implement a dynamic model whose main advantage is that it allows us to test both hypotheses of independence and conditional coverage.

In this part of the paper our new methodology based on a dynamic binary response model is introduced. First, the different model specifications considered are presented. Then, the estimation method is developed and finally the backtesting test statistics are introduced.

3.1 Model Specification

Let us consider the following dynamic binary response model, in which the conditional probability of violation at time t is given by:

$$\Pr\left[I_t(\alpha) = 1 \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[I_t(\alpha) \mid \mathcal{F}_{t-1}\right] = F\left(\pi_t\right). \tag{9}$$

where $F(.)$ denotes a c.d.f. (which does not necessarily correspond to the distribution of the returns) and $\mathcal{F}_{t-1}$ is a set of information available at $t-1$. We assume that the index π_t satisfies the following autoregressive representation:

$$\pi_t = c + \sum_{j=1}^{q_1} \beta_j \pi_{t-j} + \sum_{j=1}^{q_2} \delta_j I_{t-j}(\alpha) + \sum_{j=1}^{q_3} \psi_j l(x_{t-j}, \varphi) + \sum_{j=1}^{q_4} \gamma_j \, l(x_{t-j}, \varphi) \; I_{t-j}, \tag{10}$$

where $l(.)$ is a function of a finite number of lagged values of observables, and x_t is a vector of explicative variables. The role of $l(.)$, is to link the index π_t to the observable variables that belong to the information set. A natural choice for x_{t-j} is given by the lagged returns or the lagged VaR.

Consequently, we propose seven specifications, denoted by DB_1 to DB_7, (where DB stands for *Dynamic Binary Model*):

$$DB1 : \pi_t = c + \beta_1 \pi_{t-1}, \tag{11}$$

$$DB2 : \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha), \tag{12}$$

$$DB3 : \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \delta_2 I_{t-2}(\alpha), \tag{13}$$

$$DB4 : \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \delta_2 I_{t-2} + \delta_3 I_{t-3}(\alpha), \tag{14}$$

$$DB5 : \pi_t = c + \beta_1 \pi_{t-1} + + \psi_1 VaR_{t-1}, \tag{15}$$

$$DB6 : \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha) + \psi_1 VaR_{t-1}, \tag{16}$$

$$DB7 : \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha) + \psi_1 VaR_{t-1} + \gamma_1 VaR_{t-1} \; I_{t-1}. \tag{17}$$

The first four specifications correspond to a dynamic binary response model including the lagged index as an explanatory variable and some additional information through the observed past values of the violation process. The fifth and sixth model are derived from the autoregressive quantile specifications CAViaR used by Engle and Mangenelli (2004), and more specifically from their symmetric absolute value specification. To be more precise, the index is assumed to respond symmetrically to lagged VaR values. The last specification DB_7 introduces an asymmetry in the response of the index to past VaR according to the appearance of a VaR violation.

3.2 Estimation Technique

There are several ways to estimate a dynamic binary model. A maximum smoothly simulated likelihood estimation framework may be employed, as in Falcetti and Tudela (2006). Nonetheless, an easier method has been recently proposed by Kauppi and Saikkonen (2008). They propose a model that includes both the lagged value of the index and the one of the binary variable itself. Their model is easy to estimate in a maximum likelihood framework, under appropriate regularity conditions (stationarity of explanatory variables and normality of the random variables).

Let us recall the general form of the model: $\Pr_{t-1}(I_t(\alpha) = 1) = F(\pi_t)$, where π_t corresponds to one of the seven specifications considered in equations 11 to 17. The log-likelihood function may be expressed as:

$$\ln L(\theta; I(\alpha), Z) = \sum_{t=1}^{T} [I_t(\alpha) \ln F(\pi_t(\theta, Z_t)) + (1 - I_t(\alpha)) \ln(1 - F(\pi_t(\theta, Z_t)))], \tag{18}$$

where θ is the vector of parameters of the binary choice model except for the intercept, *i.e.* $\theta = [\beta' \ , \gamma', \ \psi', \ \delta']$ and Z_t is the vector of explanatory variables at time t corresponding to a certain specification from the seven models considered in the paper (eq. 11 to 17).

Given the maximum-likelihood framework, this dynamic time-series model is easy to implement and the maximum-likelihood estimators have the desired large-sample properties. Nevertheless, one should not lose sight of the fact that since β is an autoregressive parameter, it has to satisfy the usual stationarity condition that the roots of the corresponding polynomial lie outside the unit circle. This means that we have to proceed to a constrained maximum likelihood estimation (for more details, see Appendix 1).

3.3 Backtesting test

Testing the null hypothesis of conditional coverage by using our dynamic binary model is similar to the case of the *DQ* test of Engle and Manganelli (2004), since it is based on a test of nullity of the coefficients of the regression.

Given the model in equation (10), the CC assumption $\mathbb{E}\left[I_t(\alpha) \mid \mathcal{F}_{t-1}\right] = \alpha$ can be simply

defined as:

$$H_0 : \beta = 0,\ \delta = 0,\ \psi = 0,\ \gamma = 0 \text{ and } c = F^{-1}(\alpha). \tag{19}$$

Under the null hypothesis, the hit variable is orthogonal to any explanatory variable that belongs to the information set $\mathcal{F}_{t-1}$ and the conditional probability of a violation is equal to α:

$$\Pr(I_t = 1 \mid \mathcal{F}_{t-1}) = F(F^{-1}(\alpha)) = \alpha. \tag{20}$$

Three tests can be used in such a context, *i.e.* LR, $Wald$ and LM. However, we present only the results of the LR test, as it has better small sample properties, particularly in terms of power. The dynamic binary LR test, denoted by $DB_{LR_{CC}}$, which tests the efficiency assumption of the VaR model takes the following form:

$$\begin{aligned} DB_{LR_{CC}} &= -2\left\{\ln L(0, F^{-1}(\alpha); I_t(\alpha), Z_t) - \ln L(\hat{\theta}, \hat{c}; I_t(\alpha), Z_t)\right\} \\ &\xrightarrow[T\to\infty]{d} \chi^2\left(dim(Z_t)\right) \end{aligned} \tag{21}$$

where $\widehat{\theta}$ is the vector of estimated parameters of the binary-choice model (under the alternative hypothesis, by maximum-likelihood) and $\hat{c}$ is the estimated intercept of the model.

Similarly, the test of the independence assumption is based on H_0: $\beta = 0,\ \delta = 0,\ \psi = 0,\ \gamma = 0$, *i.e.* the nullity of the coefficients, but this time in a model where the constant is no longer a function of α, *i.e.* c is a free parameter. Under the null of IND, there is no temporal dependence of violations, but the VaR violation probability could be different from the coverage rate α:

$$\Pr(I_t = 1 \mid \mathcal{F}_{t-1}) = F(c). \tag{22}$$

The corresponding LR test statistic is similar to the one associated with the conditional coverage hypothesis, except that the degrees of freedom have to be adjusted, *i.e.* $dim(Z_t) - 1$.

In a similar way, it is also possible to define a test of the null of UC as follows:

$$\mathbb{E}\left[I_t(\alpha)\right] = \lim_{j\to\infty} \mathbb{E}\left[I_t(\alpha) \mid \mathcal{F}_{t-j}\right] = \alpha. \tag{23}$$

This test rejects the null hypothesis of an accurate VaR if the true fraction of VaR violations in the sample is statistically different from α. Actually, particular constraints, different from one model specification to another, are imposed to the parameters of each model. For instance, if we consider our second model specification (eq. 12), it is possible to show that $\mathbb{E}(\pi_t|\mathcal{F}_{t-j}) = c*(1-\beta^{j+1})/(1-\beta)+\beta^j\mathbb{E}(\pi_{t-j}|\mathcal{F}_{t-j})+\delta\mathbb{E}(I_t(\alpha)|\mathcal{F}_{t-j})(1-\beta^j)/(1-\beta)$. Thus, the null hypothesis corresponding to the UC assumption can be expressed as follows: $c = (\beta-1)F^{-1}(\alpha)-\alpha\delta$. A LR test statistic associated with the UC hypothesis can be computed and the degrees of freedom adjusted so as to account for the additional constraint(s) corresponding to a specific model.

4 Small Sample Properties

In this section, we use Monte Carlo simulations to illustrate the finite sample properties (empirical size and power) of our conditional coverage test statistic. However, it is worth noting that one of the main issues in the literature on VaR assessment is the relative scarcity of violations. Indeed, even with one year of daily returns, the number of observed violations may often be dramatically small (in particular for a 1% coverage rate), and this situation can induce poor small sample properties. For this reason, the size of the test has to be controlled using, for example, the Monte Carlo (MC) testing approach of Dufour (2006), as done in Christoffersen and Pelletier (2004) and Berkowitz et al. (2011) *inter alii*.

4.1 Empirical Size Analysis

To illustrate the size performance of our non-linear regression-based test in finite samples, two cases are considered. First, we generate a sequence of violations by taking independent draws from a Bernoulli distribution. At the same time, assessing the robustness of risk models is a very important topic nowadays (see Escanciano and Olmo, 2010). Indeed, VaR models are specific to each bank, which means that estimation error can be an important issue in the computation of violation series. We hence propose a second experiment, based on the introduction of estimation risk via a t-Garch model, so as to analyze the performance of the test in an environment closer to reality. In this context, we are interested in the deformation scale of the tests engendered by

estimation-error.

In both cases a dynamic logit model has been chosen for the backtesting test, as it is usually done in the Early Warning Systems literature. We set the link function $F(.)$ to the c.d.f of the logistic distribution, $F(.) = \Lambda(.) = e^{\omega}/(1+e^{\omega})$, where ω is the vector of parameters of the binary regression model, $i.e.$, $\omega = [c\ \theta]$.

First, we take the case of VaR violations drawn directly from a Bernoulli distribution. This means that the VaR series are not available and, as a result, only the first four specifications can actually be implemented, $i.e.$, eq. 11 to 14. Several sample sizes ranging from 250 (which roughly corresponds to one year of trading days) to 1500 were considered. Furthermore, the reported empirical sizes correspond to the rejection rates calculated over $10,000$ simulations.

[insert Table 1 here]

Table 1 reports the empirical size of the DB test statistic for various specifications of the dynamic binary model based on the binomial experiment, first considering a coverage rate of 5% then one of 1%. For comparison reasons, we also display the results of the DQ test (considering three different specifications DQ_1, DQ_2 and DQ_3, which include from one to three lags of the binary variable) and the LR test proposed by Christoffersen (1998), denoted LR_{CC}, which is actually based on a Markov chain assumption. Our choice relies on the fact that these tests are currently some of the most used in empirical applications. On the one hand, it seems that for a 5% coverage rate all the tests are oversized in small samples. On the other hand, in large samples (with at least 500 observations), our tests are oversized, the LR_{CC} one is undersized (except in very large samples where it is slightly oversized), while the DQ one is quite close to the nominal level.

For a 1% coverage rate, the results vary severely from one specification to another. More exactly, the LR_{CC} test and our first specification (DB_1) are roughly undersized, especially in small samples. All three other DB specifications are quite well sized in small samples, whereas only DB_2 remains well sized in large samples, the other two being slightly undersized. By contrast, the DQ tests are somewhat oversized in large samples.

In a second experiment we propose to assess the impact of estimation risk on the finite sample

properties of our backtesting test. Not accounting for estimation error in the simulations would pollute the estimates of the coefficients of the binary response model and thus the properties of our backtests. Therefore, we propose a new Bootstrap procedure based on Christoffersen and Gonçalves (2005) in order to compute corrected critical values:

Step A1 - Generate a series of returns $\{r_t\}_{t=1}^T$ *according to a standard t-Garch(1,1) DGP given a parameterization of the coefficients (page 11)*:

$$r_t = c + \sigma_t \sqrt{\left(\frac{v-2}{v}\right)} z_t, \tag{24}$$

where $\epsilon_t = z_t \sigma_t \sqrt{(v-2)/v}$, z_t *is an i.i.d. series from Student's t-distribution with* ν *degrees of freedom, and where the conditional variance* σ_t^2 *is given by the following recursive relation*:

$$\sigma_t^2 = \beta_0 + \beta_1 \epsilon_{t-1}^2 + \beta_2 \sigma_{t-1}^2. \tag{25}$$

The coefficients of the model are parametrized as follows: $c = 0.000999$, $\beta_0 = 7.9778e^{-7}$, $\beta_1 = 0.0896$, $\beta_2 = 0.9098$ *and* $\nu = 1/0.1633$. *These parameters correspond to those estimated by a t-Garch(1,1) for the SP500 index spanning the period from January 2, 1970 to May 5, 2006.*

Step A2 - Given the simulated series of returns, a t-Garch (1,1) model is estimated in-sample, i.e. on the first R observations so as to introduce estimation risk in the analysis and the series of VaR forecasts is obtained using the out of sample part of the dataset (T-R observations) according to the following equation:

$$\widehat{VaR}_{t|t-1}(\alpha) = \hat{c} + \sqrt{\left(\frac{\widehat{\nu}-2}{\widehat{\nu}}\right)} \widehat{\sigma}_t G^{-1}(\alpha, \widehat{\nu}), \tag{26}$$

where the conditional variance is given by $\widehat{\sigma}_t^2 = \hat{\beta}_0 + \hat{\beta}_1 (r_{t-1} - \hat{c})^2 + \hat{\beta}_2 \widehat{\sigma}_{t-1}^2$ *and* $G^{-1}(\alpha, \nu)$ *stands for the inverse of the Student distribution for a nominal coverage rate equal to* α *and* ν *degrees of freedom. Besides, we do not use rolling estimation, but suppose that the estimated parameters are stable during the whole period considered. Next, the series of violations is computed, the backtesting tests are implemented and the test-statistics are compared with the corresponding*

empirical critical value obtained by bootstrapping the standardized residuals (see Steps B1 to B6). Note that if, by contrast, the test-statistics are compared with the corresponding asymptotic critical value, the size of the test hence obtained is 'polluted' by estimation error.

Step A3 - Repeat Steps A1 and A2 a large number of times and count the number of rejections for each backtesting statistic.

To gauge this estimation error issue, bootstrapped critical values are used instead of asymptotic critical values. Indeed, to compute the empirical critical values, at each repetition of steps A1 to A2 we depart from Step A2 and implement the following steps:

Step B1 - Relying on the in-sample estimates of the parameters, $\hat{\theta} = \{\hat{c}, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{v}\}$ *and estimated conditional variance* $\hat{\sigma}_t^2$*, we compute the in-sample i.i.d. Student standardized residuals* $\{\hat{u}_t\}_{t=1}^R$*, given by* $\hat{u}_t = (r_t - \hat{c})\sqrt{(\hat{v}/(\hat{v}-2))}/\hat{\sigma}_t)$.

Step B2 - Let $\{u_t^b\}_{t=1}^R$ *be a resample drawn with replacement from the standardized residuals series. We can hence compute the resampled conditional variance series by considering the following recursive relation* $\sigma_t^{2b} = \hat{\beta}_0 + \hat{\beta}_1 \sigma_{t-1}^{2b}(u_{t-1}^b)^2(\hat{v}-2)/\hat{v} + \hat{\beta}_2 \sigma_{t-1}^{2b}$*, as well as the resampled in-sample returns series* $\{\hat{r}_t^b\}_{t=1}^R$ *by using* σ_t^{2b} *and* u_t^b *in eq. 24.*

Step B3 - Next, we estimate a t-Garch(1,1) model on this new in-sample data ($\{\hat{r}_t^b\}_{t=1}^R$*). Denote by* $\tilde{\theta} = \{\tilde{c}, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{v}\}$ *the corresponding vector of parameters. Out-of-sample bootstrap predictions of the conditional variance can be hence obtained according to*: $\tilde{\sigma_t^2}^b = \tilde{\beta}_0 + \tilde{\beta}_1(r_{t-1} - c)^2 + \tilde{\beta}_2 \tilde{\sigma_{t-1}^2}^b$*, where* $t = R+1, ..., T$*, given the historical returns at time* $t-1$ *and the initial value of the conditional variance:*

$$\tilde{\sigma_R^2}^b = \frac{\tilde{\beta}_0}{1-\tilde{\beta}_1-\tilde{\beta}_2} + \tilde{\beta}_1 \sum_{j=0}^{R-2} \tilde{\beta}_2^j \left((r_{R-j-1} - c)^2 - \frac{\tilde{\beta}_0}{1-\tilde{\beta}_1-\tilde{\beta}_2} \right).$$

The associated out-of-sample VaR is next computed as $VaR_t = \tilde{\beta}_0 + \sqrt{(\tilde{v}-1/\tilde{v})}\tilde{\sigma_t}^b G^{-1}(\alpha, \tilde{v})$*, where* $t = \{R+1, ..., T\}$ *and* $G^{-1}(\alpha, \tilde{v})$ *is the inverse of the Student distribution for an* α *nominal coverage and* $\tilde{v}$ *degrees of freedom, so that the series of violations* $\{I_t\}_{R+1}^T$*, taking the value of 1 in the case of a* VaR *exception, is finally obtained.*

Step B4 - At this step, all the backtesting tests can be implemented, and the corresponding test-statistics are regrouped in a row vector $\{DB_1^b, DB_2^b, DB_3^b, DB_4^b, DB_5^b, DB_6^b, DB_7^b, DQ_1^b,$

$DQ_2^b, DQ_3^b, LR^b\}$.

Step B5 - Repeat Steps B1 to B4 a large number of times and save the row vector of test-statistics obtained at each iteration (i) as the i^{th} row of the matrix of test statistics. The empirical critical value for each of the tests is given by the 95% quantile of the corresponding series of test-statistics (the appropriate column of the matrix of test-statistics).

These nested simulations are very time-consuming. This is why we use 1000 simulations in steps A1-A2, whereas we consider 100 repetitions for the steps B1 to B4. The coverage rate is fixed to 5%, and several in-sample/out-of-sample sizes are considered. We thus illustrate the impact of the in-sample size on the estimation error and that of the out-of-sample size on the small-sample properties of the backtests. Table 2 shows the influence of the estimation risk on the size of the backtesting tests. It reports the uncorrected size and the corrected one. To be more precise, the former is computed by relying on asymptotic critical values, whereas the latter, *i.e.* the corrected size, is obtained by implementing the procedure presented above (steps A1-A4 and B1-B5) and it is reported between parentheses.

[insert Table 2 here]

The most striking finding is that in the presence of estimation error all tests are oversized whatever the in-sample/out-of-sample sizes. To be more exact, the larger the in-sample size, the closer the size of the tests is to the nominal 5% level. As expected, the estimated parameters asymptotically converge to the true parameter values and the impact of estimation error asymptotically vanishes. Besides, for the same in-sample size (R=250), an increase in the out-of-sample window (from 250 to 500) leads to a loss in the size properties of the tests, *i.e.* the tests become more oversized. This finding corresponds to our a-priori expectation that the larger the number of forecasts based upon parameters affected by estimation error, the more the backtest applied to these forecasts will be likely to 'wrongly' reject the null hypothesis of valid VaR forecasts.

We can emphasize, however, the fact that when correcting for estimation error, the size of the tests is close to the nominal one. To be more precise, in small samples most of the tests are undersized, and DB_1 seems to works better than the others. For larger sample sizes this test becomes oversized, which is also the case for DQ_1 and LR_{CC}, whereas DQ_2 and DQ_3 remain

undersized. The other DB tests are in fact the closest to the nominal size in large samples (DB_3 in particular).

Tables 3 and 4 present the feasibility ratios of the size simulations based on the Bernoulli distribution and the t-Garch(1,1) returns respectively.

[insert Tables 3 and 4 here]

It can be noticed that the variables specific to dynamic discrete choice models, namely the lagged probability and the lagged binary variables, are strongly correlated and thus sometimes they completely separate the observations into their response groups (violation/no violation). This situation is called complete separation of data points. Its main consequence is that the likelihood maximization algorithm does not have a finite solution (or in other words, the maximum likelihood estimate does not exist). More specifically, this problem arises in the models containing both the lagged probability and lagged binary variables, and its negative impact rebounds on the feasibility of these tests. Consequently, it is only the DB_1 specification that can be always computed, followed by DB_5 and DB_6. It is also worth noting that the LR_{CC} and DQ tests cannot be computed when there is no violation in the sample, which reduces their feasibility particularly for a 1% nominal coverage, and in small samples. Similar results are obtained for the t-Garch experiment. In consequence, we present the power results only for the DB_1, the DB_5 and the DB_6 specifications, as well as the DQ and the LR_{CC} tests.

4.2 Power analysis

We now investigate the power of the test for two different alternative hypotheses. First, an *i.i.d.* Bernoulli violations series with a false coverage rate is considered under the alternative hypothesis. As usual in this literature, the Monte Carlo technique proposed by Dufour (2006) is implemented in this case. Second, a t-Garch(1,1) model is considered. Another major implication of the estimation error concerns the power results of this model. Thus, the empirical power has to be computed from the rejection rates obtained with the bootstrapped critical values.

Indeed, we implement a two-step procedure specially designed to size-adjust the power results, as follows. *First, we generate a returns series* $\{r_t\}_{t=1}^{T}$ *according to a standard t-Garch(1,1) DGP*

following the different specifications considered (page 11). Then, we compute the sequence of VaR forecasts under the alternative hypothesis by relying on the Historical Simulations method. Next, we implement the backtests and obtain the test-statistics under the alternative. 1000 repetitions are considered, and for each of them we compute bootstrapped critical values based on 100 simulations under the null hypothesis, performed by following the same procedure as for the size simulations (see steps A2 and B1-B5). Note that a t-Garch(1,1) model is used for the critical values and not the Historical Simulation method, since these simulations are performed under the null hypothesis.

The rejection frequency for the conditional coverage tests based on i.i.d. Bernoulli sequences of violation are reported in Table 5 (both at 5% and 1% coverage rate).

[insert Tables 5 and 6 here]

Under the alternative hypothesis a 3% coverage rate has been considered. It seems that the LR_{CC} and DB tests outperform the DQ specifications, particularly in small samples and for a 5% nominal coverage rate.

In the t-Garch experiment (see Table 6), a VaR-HS is estimated under the alternative hypothesis, while under the null hypothesis the VaR computation is based on the t-Garch(1,1) model. Recall that to size adjust the power results, the backtest-statistics computed under the alternative hypothesis are compared with the corresponding bootstrapped critical values. The coverage rate is fixed to 5% and the same pairs of in-sample/out-of-sample size are used as in the size simulations.

As the main result, the DB, DQ_1 and LR_{CC} tests always exhibit low power. Furthermore, the results in Table 1 show that, for a backtest which relies on dynamic regression models, the introduction of the VaR apart from that of the lagged index (which gives the dynamics of the system) improves the power results. In particular, it is the case of the DB_5 specification, which seems to outperform all other specifications, regardless of the pair of in-sample/out-of-sample size we choose.

Besides, the power of the tests slowly increases with the in-sample size while it decreases when the out-of-sample window grows but the in-sample one stays the same.

Similar to the case of size simulations, the DB_1 conditional coverage test can always be computed, whereas the other tests are not always feasible, especially when estimation error is taken into account (see Tables 7 and 8).

[insert Tables 7 and 8 here]

To sum up, all tests have ups and downs, especially for the size simulations when estimation error is taken into account, as they over-reject the validity of the VaR. However, our test seems to work slightly better than the others in this framework, as it rejects less frequently the validity of the VaR sequence. Besides, in all sets of simulations our test exhibits good power properties, better than the other two tests, particularly in small samples (T=250), even in the presence of estimation risk. Since it appears that there is no "best backtest", as none of them outperforms the others in all cases, our findings support the use of several validation tests whenever possible, and the implementation of a non-linear regression-based test in small samples.

5 Empirical application

Berkowitz and O'Brien (2002) compare VaR forecasts obtained from the structural models of 6 large U.S. banks to those issued from a standard GARCH model applied to the P&L volatility for each bank and show that, against expectations, the VaR models used by those banks are not better than simple models of volatility. In a similar vein, we gauge the impact of correlations between the assets composing a portfolio on the computation and implicitly on the validation of the VaR for that portfolio.

We thus illustrate our new backtest by considering a portfolio of three assets from the CAC40 market index, *i.e.* Alcatel-Lucent, Carrefour and Essilor International, for the period June 1, 2007 - June 1, 2009. The weights used to construct the portfolio have been chosen as the fraction of the market capitalization of the portfolio corresponding to each asset. Besides, the composition of the portfolio is assumed to be constant over time.

Two cases are considered. First, we care for the correlation between assets by using a multivariate framework, *i.e.* the Dynamic Conditional Correlation Multivariate GARCH model

(Engle and Sheppard, 2001, Engle, 2002), hereafter DCC. Indeed, the VaR is computed by using the conditional variance of the portfolio issued from the second step of the DCC model. Second, the VaR of the portfolio is issued from a standard Garch(1,1) model estimated for the portfolio returns, *i.e.* that neglects the correlations between the assets.

Figure 1 presents the three conditional correlation coefficients obtained from the DCC model.

[insert Figure 1 here]

It seems that there is a strong correlation between the three assets (Alcatel-Lucent, Carrefour and Essilor International) during this period. We proceed hence to the validation of the two VaR sequences obtained from a model taking into account the correlation between assets and one that does not consider this correlation.

The statistics and p-values of the DB tests are reported both at a 5% and at a 1% coverage rate. For comparison purposes, the findings of the DQ tests and that of Christoffersen's LR_{CC} test are also displayed. The results of the backtests for the first sequence of VaR (based on the DCC model) are displayed in the left part of Table 9 while those corresponding to the second VaR sequence (relying on a Garch(1,1) model) are presented in the right part of this table.

[insert Table 9 here]

The tests conclude to the validity of both risk measures for a 5% as well as for a 1% coverage rate. To be more precise, the two VaR sequences (obtained from the multivariate DCC model and from the univariate Garch model respectively) are too similar for the backtests to discriminate between them. In other words, the correlations amongst assets do not have any impact on the computation of the VaR and implicitly on its validity. Our findings thus go along the lines of Berkowitz and O'Brien's diagnostic (2002).

Besides, Figure 2 depicts the series of violation probabilities issued from the seven Dynamic Binary backtesting specifications we propose. More exactly, it corresponds to the in-sample VaR(5%) resulting from the Garch(1,1) model on the returns of the aggregated portfolio.[3]

[insert Figure 2 here]

[3]The plots corresponding to a 1% coverage rate and/or the DCC model are available upon request.

For each Dynamic Binary backtest (DB1 to DB6), Figure 2 depicts the estimated hit probability from June 1, 2007 to June 1, 2009. Most importantly, this graphical result confirms the fact that the univariate VaR-Garch validity cannot be rejected, whatever the specification of the Dynamic Binary test. Recall that under the null, the hit probability is precisely equal to the coverage rate, *i.e.* 5% in our case. We can observe that for all specifications, the estimated probabilities level close to the coverage rate all over the period. We do not report a confidence interval on the estimated probabilities in Figure 2, since this information is redundant with that given by the formal backtests. Obviously, the richer the dynamic specification of the binary model is, (including lagged values of VaR, lagged values of the hit variable, etc.), the more volatile the estimated probabilities are and the lower the power. Besides, the introduction of additional regressors reduces the feasibility of the tests (see the size and power feasibility tables). This is why we recommend the first and the fifth specification, DB_1 and DB_5, that include the lagged index and the lagged VaR and the lagged index, respectively.

6 Conclusion

In this paper we propose a new backtest based on non-linear regression models. The baseline idea is that the validity of the VaR can be tested through joint restrictions on the parameters of a regression in which the dependent variable is the series of violations. In view of the discrete choice character of the series of violations, the use of a non-linear model seems to be more appropriate than the estimation of a linear one. We expect to ameliorate the finite-sample properties of the tests by gauging the impact of modelling the binary choice indicator in a proper non-linear framework on the properties of the backtest.

Departing from the *Early Warning System* literature (Kauppi and Saikonnen, 2008), Dynamic Binary logit models including the lagged index, *i.e.* DB, are estimated and classic likelihood-ratio tests are implemented. Several specifications are considered for the DB test, including the lagged violations and the lagged VaR. For comparison reasons the DQ and LR_{CC} tests are also presented.

In a Monte Carlo experiment, we show that the use of a framework which is appropriate for

the binary character of the dependent variable slightly improves the finite sample properties of the backtesting tests. This is particularly true in small samples, when estimation risk is taken into account, case that is notably interesting from the practitioner's point of view.

An empirical application on an illustrative portfolio constituted from three assets included in the CAC40 market index for the period June 1, 2007 - June 1, 2009 is finally proposed. It reveals that the correlations amongst assets do not have a significant impact on the computation of the VaR and implicitly on the validity of the VaR. To be more exact, the VaR sequence issued from a multivariate DCC model, as well as the one based on a univariate Garch model, are both valid at a 5% and 1% coverage rate.

References

Berkowitz, J. and J. O'Brien, "How Accurate are Value-at-Risk Models at Commercial Banks?," *Journal of Finance*, vol. 57, 2002, 1093-1111.

Berkowitz, J., P. F. Christoffersen and D. Pelletier, "Evaluating Value-at-Risk Models with Desk-Level Data," *Management Science*, vol. 57, no. 12, December 2011, 2213-2227.

Campbell, S. D., "A Review of Backtesting and Backtesting Procedures," *Journal of Risk*, vol. 9, no. 2, 2007, 1-18.

Candelon, B., G. Colletaz, C. Hurlin and S. Tokpavi, "Backtesting Value-at-Risk: A GMM Duration-Based Test," forthcoming in Journal of Financial Econometrics, 2011.

Christoffersen, P. F., "Evaluating Interval Forecasts," *International Economic Review*, vol. 39, 1998, 841-862.

Christoffersen, P. F. and D. Pelletier, "Backtesting Value-at-Risk: A Duration-Based Approach," *Journal of Financial Econometrics*, vol. 2, no. 1, 2004, 84-108.

Christoffersen, P. F. and S. Gonçalves, "Estimation Risk in Financial Risk Management," *Journal of Risk*, vol. 7, 2005, 1-28.

Dufour, J.-M., "Monte Carlo Tests with Nuisance Parameters: a General Approach to Finite Sample Inference and Nonstandard Asymptotics," *Journal of Econometrics*, vol. 127, no. 2,

2006, 443-477.

Engle, R. F. and K. Sheppard, "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH," NBER Working Papers 8554, 2001.

Engle, R. F., Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models, *Journal of Business and Economic Statistics*, vol. 20, no. 3, 2002, 339-50.

Engle, R. F. and S. Manganelli, "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles," *Journal of Business & Economic Statistics*, vol. 22, 2004, 367-381.

Escanciano, J. C. and J. Olmo, "Backtesting Parametric Value-at-Risk with Estimation Risk," *Journal of Business & Economic Statistics*, vol 28, no. 1, 2010, 36-51.

Falcetti, E. and M. Tudela, "Modelling Currency Crises in Emerging Markets: A Dynamic Probit Model With Unobserved Heterogeneity and Autocorrelated Errors," *Oxford Bulletin of Economics and Statistics*, vol. 68, no. 4, 2006, 445-471.

Gaglianone, W.P., L.R. Lima, O. Linton and D.R. Smith, "Evaluating Value-at-Risk Models via Quantile Regression," *Journal of Business & Economic Statistics*, vol. 29, no.1, 2011, 150-160.

Gourieroux, C., *Econometrics of Qualitative Dependent Variables*, 2000, Cambridge University Press.

Gourieroux, C., J.P. Laurent and O. Scaillet, "Sensitivity Analysis of Value at Risk," *Journal of Empirical Finance*, vol. 7, 2000, 225-245.

Haas, M., "Improved Duration-Based Backtesting of Value-at-Risk," *Journal of Risk*, vol. 8, no. 2, 2005, 17-36.

Hurlin, C. et S. Tokpavi, "Un Test de Validité de la Value-at-Risk," *Revue Economique*, vol. 58, no. 3, 2007, 599-608.

Hurlin, C. and S. Tokpavi, "Une Evaluation des Procédures de Backtesting: Tout va pour le Mieux dans le Meilleur des Mondes," *Finance*, vol. 29, no. 1, 2008, 53-80.

Jorion, P., *Value-at-Risk, Third edition: The New Benchmark for Managing Financial Risk*, 2007, McGraw-Hill.

Kauppi, H. and P. Saikkonen, "Predicting U.S. Recessions with Dynamic Binary Response Models," *The Review of Economics and Statistics*, vol. 90, no. 4, 2008, 777-791.

Kupiec, P., "Techniques for Verifying the Accuracy of Risk Measurement Models," *The Journal of Derivatives*, vol. 3, no. 2, 1995, 73-84.

Pérignon C. and D.L.R. Smith, "A New Approach to Comparing VaR Estimation Methods," *The Journal of Derivatives*, vol. 16, no. 2, 2008, 54-66.

Scaillet, O., "The Origin and Development of VaR," *Modern Risk Management: A History, 15th Anniversary of Risk*, 2003, 151-158.

Appendix 1: Constrained Maximum Likelihood Estimation (Kauppi and Saikkonen, 2008)

Let us recall the general form of the dynamic binary response model: $P_{t-1}(I_t(\alpha)=1)=\Lambda(\beta\pi_{t-1}+\delta I_{t-1}(\alpha)+x_{t-1}\xi)$, where $\xi=[\psi,\gamma]$. The initial value π_0 is set to $(\bar{x}\xi)/(1-\beta)$, $\bar{x}$ being the sample mean of the exogenous variables and ξ being the vector of parameters $\xi=[c,\ \psi,\ \gamma]$. The initial condition for the ξ vector of parameters is given by an OLS estimation, while the initial β is set to 0.

Since β is an autoregressive parameter, a constrained maximum likelihood estimation must be implemented. Nevertheless, the same results can be reached in a faster and easier way by using a transformation of the β parameter in the classical maximum likelihood process. Thus, in order to solve this problem, we denote by η the new maximization parameter identified so that β is equal to $\eta/(1+|\eta|)$, *i.e.*, β takes values in the interval [0,1].

We must emphasize the fact that, in view of the parameter transformation from β to η, the maximization variance-covariance matrix corresponds to the parameters $[\xi\ \eta]$, and not to the initial parameters $[\xi\ \beta]$. Thus, we must proceed to a modification of the variance-covariance matrix from the first space to the second one. To this purpose, the approximation of the transformation function around the point η_0 is computed based on Taylor's theorem. To be more

exact, since the estimated parameter $\hat{\beta} = f(\hat{\eta})$, where $f(\hat{\eta}) = \hat{\eta}/(1 + |\hat{\eta}|)$, the approximation becomes:

$$\hat{\beta} = f(\hat{\eta}) \simeq f(\eta_0) + \frac{\partial f(\hat{\eta})'}{\partial \eta}|_{\eta_0}(\hat{\eta} - \eta_0). \tag{27}$$

Since we actually aim at finding the variance of β, we use the formula $Var(a'X) = a'Var(X)a$:

$$Var(\hat{\beta}) \simeq 0 + \frac{\partial f(\hat{\eta})'}{\partial \eta}|_{\eta_0} Var(\hat{\eta}) \frac{\partial f(\hat{\eta})}{\partial \eta}|_{\eta_0} \tag{28}$$

Since $\hat{\eta} \xrightarrow{p} \eta_0$, we can replace η_0 with the estimator $\hat{\eta}$ in eq. 28:

$$Var(\hat{\beta}) \simeq 0 + \frac{\partial f(\hat{\eta})'}{\partial \eta}|_{\hat{\eta}} Var(\hat{\eta}) \frac{\partial f(\hat{\eta})}{\partial \eta}|_{\hat{\eta}} \tag{29}$$

Last but not least, the first derivative of the transformation function $f(\hat{\eta})$ with respect to $(\hat{\eta})$ can be computed through finite differences. Consequently, the standard errors obtained as the square root of the elements laying on the first diagonal of the variance-covariance matrix are consistent with the $[\beta\ \eta]$ vector of parameters.

Appendix 2: Figures and Tables

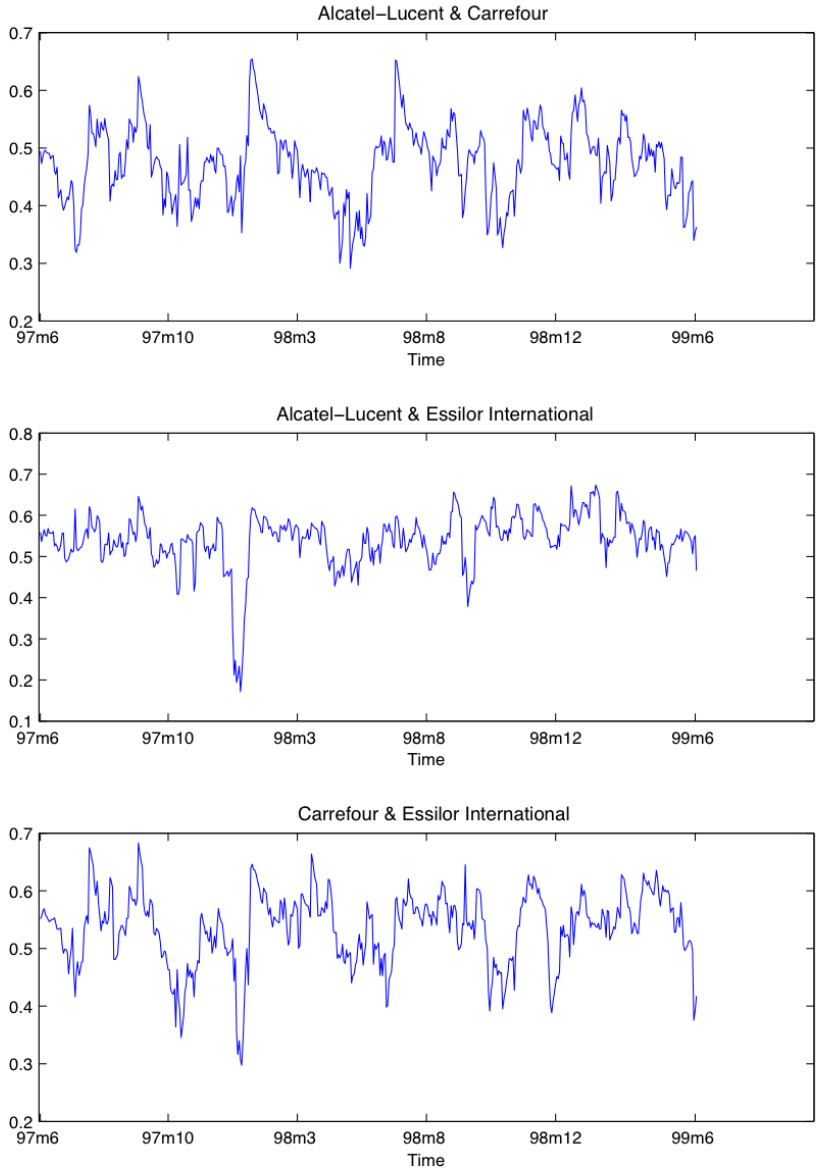


Figure 1: Conditional correlations

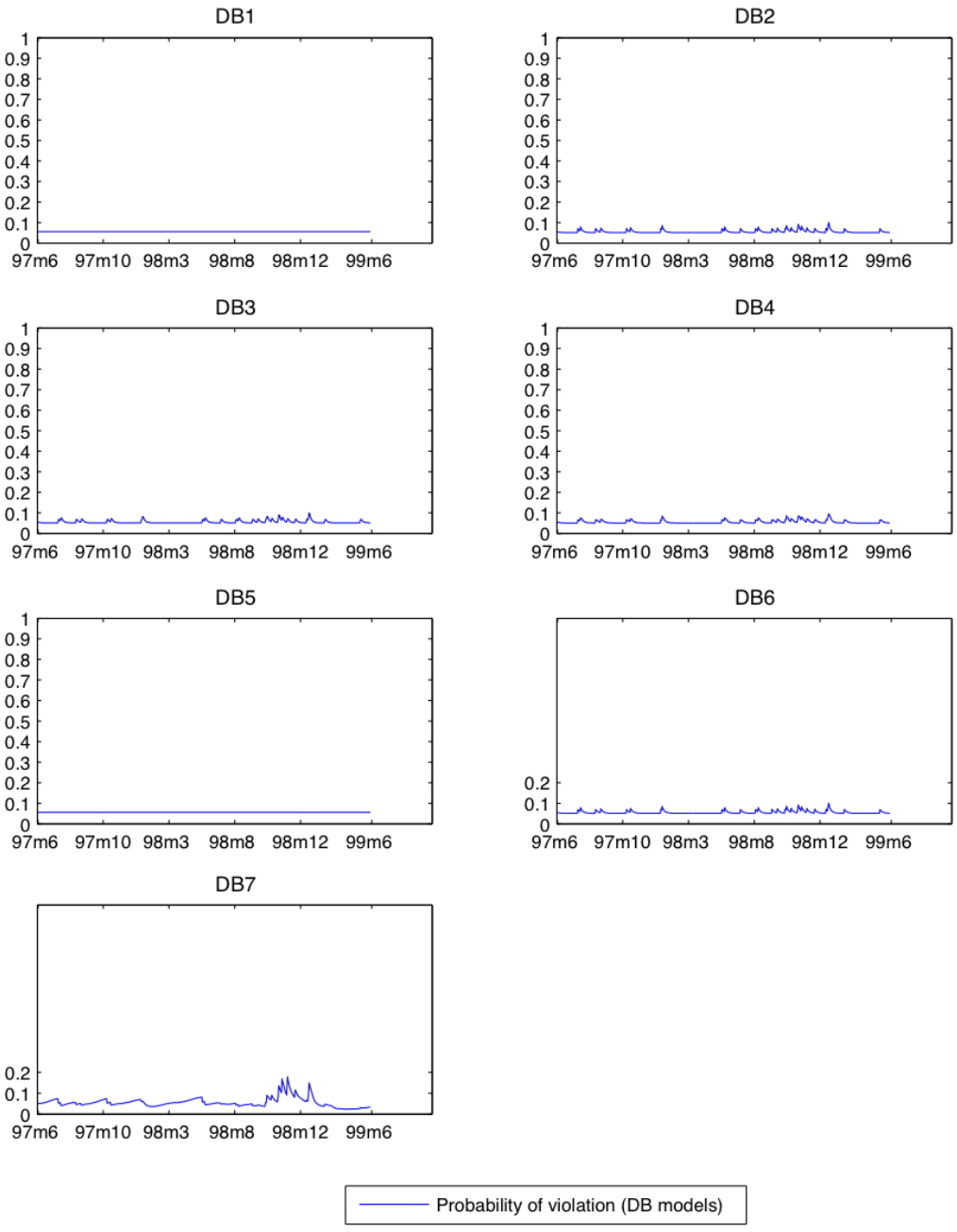


Figure 2: Violation Probabilities - DB specifications (VaR-Garch - 5% coverage rate)

Table 1. Empirical size - CC tests

Coverage rate 5%								
Sample size	DB_1	DB_2	DB_3	DB_4	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	0.0940	0.1293	0.1209	0.1075	0.1029	0.0890	0.0903	0.0632
$T = 500$	0.0664	0.0876	0.0801	0.0788	0.0529	0.0630	0.0648	0.0376
$T = 750$	0.0624	0.0759	0.0728	0.0724	0.0481	0.0495	0.0495	0.0373
$T = 1000$	0.0690	0.0834	0.0848	0.0821	0.0518	0.0492	0.0521	0.0486
$T = 1500$	0.0625	0.0846	0.0785	0.0780	0.0472	0.0494	0.0504	0.0574
Coverage rate 1%								
Sample size	DB_1	DB_2	DB_3	DB_4	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	0.0113	0.0612	0.0580	0.0490	0.0290	0.0416	0.0594	0.0065
$T = 500$	0.0109	0.0425	0.0286	0.0209	0.0485	0.0771	0.1112	0.0053
$T = 750$	0.0152	0.0479	0.0313	0.0243	0.0835	0.1256	0.1375	0.0169
$T = 1000$	0.0244	0.0474	0.0339	0.0237	0.1117	0.0859	0.1064	0.0267
$T = 1500$	0.0227	0.0474	0.0300	0.0252	0.0621	0.0612	0.0690	0.0325

Note: Under the null, the series of violations is i.i.d. from a Bernoulli distribution. For each sample, we provide the percentage of rejection of the coverage tests both at a 5% and 1% nominal level based on 10000 replications. DB_j denotes the Dynamic Binary test for the j^{th} specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004).

Table 2. Empirical size - CC tests with estimation error

Coverage rate 5%											
R	T	DB_1	DB_2	DB_3	DB_4	DB_5	DB_6	DB_7	DQ_1	DQ_2	DQ_3	LR_{CC}
250	500	0.2400	0.2835	0.2675	0.2450	0.2490	0.3030	0.2640	0.2115	0.1990	0.1835	0.2315
		(0.0420)	(0.0270)	(0.0220)	(0.0210)	(0.0310)	(0.0300)	(0.0290)	(0.0680)	(0.0140)	(0.0100)	(0.0380)
250	750	0.3467	0.4006	0.3840	0.3720	0.3557	0.4259	0.3862	0.3636	0.3418	0.3263	0.3263
		(0.0650)	(0.0410)	(0.0360)	(0.0410)	(0.0380)	(0.0400)	(0.0360)	(0.0810)	(0.0240)	(0.0170)	(0.0630)
500	750	0.1510	0.1945	0.1835	0.1750	0.1660	0.2230	0.1885	0.1325	0.1270	0.1275	0.1360
		(0.0630)	(0.0440)	(0.0390)	(0.0340)	(0.0320)	(0.0330)	(0.0320)	(0.0740)	(0.0170)	(0.0100)	(0.0540)
750	1000	0.1070	0.1535	0.1385	0.1310	0.1265	0.1995	0.1645	0.0975	0.0920	0.0900	0.0910
		(0.0690)	(0.0560)	(0.0480)	(0.0420)	(0.0420)	(0.0480)	(0.0440)	(0.0750)	(0.0170)	(0.0110)	(0.0670)
1000	1250	0.0900	0.1395	0.1400	0.1165	0.1240	0.1935	0.1690	0.0895	0.0960	0.1035	0.0760
		(0.0700)	(0.0570)	(0.0510)	(0.0510)	(0.0560)	(0.0400)	(0.0380)	(0.0880)	(0.0230)	(0.0180)	(0.0630)

Note: We generate returns using a Garch(1,1) model with standardized Student-t innovations and compute the out-of-sample VaR sequence using a t-Garch(1,1) model estimated on the first R observations. We provide the results 'polluted' by estimation error as well as those corrected for estimation error by using bootstrapped instead of asymptotic critical values (the corrected size is between parentheses). For each in-sample/out-of-sample pair (R, $T-R$), we show the percentage of rejection of the coverage tests at the 5% nominal level based on 1000 replications. At each replication, the empirical critical values are calculated by performing 100

resamplings of the standardized residuals from the t-Garch model. DB_j denotes the Dynamic Binary test for the $'j'^{th}$ specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004).

Table 3. Feasibility ratios - size simulations

	5% Coverage rate							
Sample size	DB_1	DB_2	DB_3	DB_4	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	1.0000	0.7952	0.7090	0.6446	1.0000	1.0000	1.0000	1.0000
$T = 500$	1.0000	0.9228	0.8648	0.8252	1.0000	1.0000	1.0000	1.0000
$T = 750$	1.0000	0.9653	0.9393	0.9141	1.0000	1.0000	1.0000	1.0000
$T = 1000$	1.0000	0.9834	0.9677	0.9574	1.0000	1.0000	1.0000	1.0000
$T = 1500$	1.0000	0.9946	0.9913	0.9881	1.0000	1.0000	1.0000	1.0000
	1% Coverage rate							
Sample size	DB_1	DB_2	DB_3	DB_4	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	1.0000	0.9516	0.9351	0.9181	0.9203	0.9203	0.9203	0.9194
$T = 500$	1.0000	0.9715	0.9653	0.9397	0.9944	0.9944	0.9944	0.9943
$T = 750$	1.0000	0.9775	0.9725	0.9452	0.9995	0.9995	0.9995	0.9995
$T = 1000$	1.0000	0.9766	0.9773	0.9458	0.9999	0.9999	0.9999	0.9999
$T = 1500$	1.0000	0.9835	0.9824	0.9654	1.0000	1.0000	1.0000	1.0000

Note: The results correspond to the fraction of the sample for which a certain test can be computed. DB_j denotes the Dynamic Binary test for the j'^{th} specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004). Also note that the series of violations is i.i.d. Bernoulli.

Table 4. Feasibility ratios - size simulations with estimation risk

		Coverage rate 5%										
R	T	DB_1	DB_2	DB_3	DB_4	DB_5	DB_6	DB_7	DQ_1	DQ_2	DQ_3	LR_{CC}
250	500	1.0000	0.8300	0.7542	0.7071	0.9787	0.9143	0.9063	0.9911	0.9911	0.9911	0.9911
		(0.9961)	(0.8319)	(0.7347)	(0.7308)	(0.9815)	(0.9077)	(0.8980)	(0.9864)	(0.9864)	(0.9864)	(0.9864)
250	750	1.0000	0.9333	0.8921	0.8567	0.9942	0.9493	0.9419	0.9887	0.9887	0.9887	0.9887
		(0.9970)	(0.9453)	(0.8946)	(0.8678)	(0.9891)	(0.9374)	(0.9314)	(0.9940)	(0.9940)	(0.9940)	(0.9940)
500	750	1.0000	0.8109	0.7260	0.6826	0.9741	0.9132	0.9107	0.9980	0.9980	0.9980	0.9980
		(0.9953)	(0.7978)	(0.7575)	(0.6807)	(0.9813)	(0.8886)	(0.8867)	(0.9916)	(0.9916)	(0.9916)	(0.9916)
750	1000	1.0000	0.7996	0.7181	0.6577	0.9825	0.9095	0.9035	0.9995	0.9995	0.9995	0.9995
		(0.9973)	(0.7815)	(0.7157)	(0.6654)	(0.9671)	(0.9031)	(0.8958)	(0.9963)	(0.9963)	(0.9963)	(0.9963)
1000	1250	1.0000	0.8001	0.7091	0.6667	0.9780	0.9080	0.8936	0.9995	0.9995	0.9995	0.9995
		(0.9991)	(0.8070)	(0.7160)	(0.6601)	(0.9729)	(0.9116)	(0.8999)	(0.9991)	(0.9991)	(0.9991)	(0.9991)

Note: The results correspond to the fraction of the sample for which a certain test can be computed. We provide the results for the size 'polluted' by estimation error as well as that corrected for estimation error by using bootstrapped instead of asymptotic critical values (the corrected size is between parentheses). DB_j denotes the Dynamic Binary test for the $'j'^{th}$ specification. Similarly,

LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004). Also note that a t-Garch(1,1) model has been considered both for the generation of the returns and the estimation of the VaR.

Table 5. Empirical power - CC tests

			Coverage rate 5%		
Sample size	DB_1	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	0.2394	0.1472	0.1171	0.0667	0.3119
$T = 500$	0.3649	0.3876	0.2961	0.2241	0.5072
$T = 750$	0.5592	0.5716	0.4390	0.3130	0.6884
$T = 1000$	0.7698	0.7284	0.6310	0.5258	0.8271
$T = 1500$	0.9394	0.9332	0.8812	0.8272	0.9417
			Coverage rate 1%		
Sample size	DB_1	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	0.6563	0.6190	0.4911	0.3960	0.6292
$T = 500$	0.8737	0.8035	0.7685	0.7448	0.8655
$T = 750$	0.9785	0.9583	0.9464	0.9226	0.9801
$T = 1000$	0.9954	0.9934	0.9905	0.9872	0.9943
$T = 1500$	0.9999	0.9999	0.9998	0.9992	0.9997

Note: Under the null hypothesis, the series of violations is i.i.d. Bernoulli. For each sample, we provide the percentage of rejection of the coverage tests both at a 5% and 1% nominal level based on 10000 replications. A 3% false coverage rate is considered under the alternative hypothesis. DB_1 denotes the Dynamic Binary test for the first specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on a Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004).

Table 6. Empirical power
CC tests with estimation error

					Coverage rate 5%			
R	T	DB_1	DB_5	DB_6	DQ_1	DQ_2	DQ_3	LR_{CC}
250	500	0.0820	0.1740	0.1520	0.1510	0.1000	0.0950	0.0890
250	750	0.0220	0.1210	0.1120	0.1170	0.0650	0.0630	0.0360
500	750	0.1820	0.2640	0.2380	0.2170	0.1370	0.1250	0.1770
750	1000	0.2680	0.3550	0.3220	0.3080	0.2470	0.2190	0.2840
1000	1250	0.3290	0.4300	0.3960	0.3790	0.2880	0.2690	0.3340

Note: We generate returns using a t-Garch(1,1) model with standardized Student-t innovations and compute the out-of-sample VaR sequence using the Historical Simulation method. For each sample, we provide the percentage of rejection of the coverage tests at a 5% nominal level based on 1000 replications and bootstrapped critical values so as to account for estimation error. The empirical critical values are obtained under the null by performing 100 resamplings of the standardized residuals from the t-Garch model. DB_j denotes the Dynamic Binary test for the j^{th} specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998)

based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004). R stands for the in-sample size whereas T represents the full-sample size.

Table 7. Feasibility ratios - power simulations

	5% Coverage rate				
Sample size	DB_1	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	1.0000	0.9995	0.9995	0.9995	0.9995
$T = 500$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 750$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1500$	1.0000	1.0000	1.0000	1.0000	1.0000
	1% Coverage rate				
Sample size	DB_1	DQ_1	DQ_2	DQ_3	LR_{CC}
$T = 250$	1.0000	0.9995	0.9995	0.9995	0.9995
$T = 500$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 750$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1500$	1.0000	1.0000	1.0000	1.0000	1.0000

Note: The results correspond to the fraction of the sample for which a certain test can be computed. DB_1 denotes the Dynamic Binary test for the first specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004). Also note that the series of violations is i.i.d. Bernoulli with a false 3% coverage rate under the alternative hypothesis.

Table 8. Feasibility ratios - power simulations with estimation error

		5% Coverage rate						
R	T	DB_1	DB_5	DB_6	DQ_1	DQ_2	DQ_3	LR_{CC}
250	500	1.0000	0.9388	0.7570	0.9815	0.9815	0.9815	0.9815
250	750	1.0000	0.9374	0.7624	1.0000	1.0000	1.0000	1.0000
500	750	1.0000	0.9101	0.7837	0.9410	0.9410	0.9410	0.9410
750	1000	1.0000	0.9013	0.7788	0.9168	0.9168	0.9168	0.9168
1000	1250	1.0000	0.9134	0.8115	0.9026	0.9026	0.9026	0.9026

Note: The results correspond to the fraction of the sample for which a certain test can be computed when analyzing the power of the tests corrected for estimation error (by relying on bootstrapped critical values). DB_j denotes the Dynamic Binary test for the j^{th} specification. Similarly, LR_{CC} represents Christoffersen's LR test (1998) based on Markov chain approach, while DQ denotes the Dynamic Quantile test proposed by Engle and Manganelli (2004). Also note that a t-Garch(1,1) model has been considered for the generation of the returns and the Historical Simulations method has been used under the alternative, to compute the VaR forecasts. R stands for the in-sample size whereas T represents the full-sample size.

Table 9. Conditional coverage Backtesting tests:

	DCC		Garch	
	VaR 5%	VaR 1%	VaR 5%	VaR 1%
DB_1	0.4076	0.0206	0.6632	0.6990
	(0.8156)	(0.9898)	(0.7178)	(0.7050)
DB_2	0.9146	2.7400	0.6838	7.3746
	(0.8219)	(0.4335)	(0.8770)	(0.0609)
DB_3	0.9397	2.7424	1.1083	7.3514
	(0.9188)	(0.6018)	(0.8929)	(0.1184)
DB_4	0.9481	2.7411	2.2938	7.3273
	(0.9666)	(0.7398)	(0.8072)	(0.1974)
DB_5	0.4074	1.3250	0.6642	2.8436
	(0.9387)	(0.7232)	(0.8816)	(0.4164)
DB_6	0.9187	9.2295	0.6849	9.4812
	(0.9219)	(0.0556)	(0.9532)	(0.0501)
DB_7	4.0038	9.4527	0.8590	9.4802
	(0.5489)	(0.0923)	(0.9731)	(0.0914)
DQ_1	0.4649	0.0505	0.6575	0.8885
	(0.7926)	(0.9751)	(0.7198)	(0.6413)
DQ_2	0.6211	0.1013	0.9994	1.0296
	(0.8916)	(0.9917)	(0.8014)	(0.7941)
DQ_3	0.8749	0.1532	1.3205	1.1748
	(0.9281)	(0.9972)	(0.8579)	(0.8822)
$DQVaR_1$	0.4654	1.1330	0.6611	3.3977
	(0.9264)	(0.7691)	(0.8823)	(0.3343)
$DQVaR_2$	0.7352	1.2037	1.1721	3.5630
	(0.9810)	(0.9445)	(0.9475)	(0.6139)
$DQVaR_3$	1.8576	1.3115	2.1057	3.6923
	(0.9673)	(0.9881)	(0.9538)	(0.8145)
LR	0.4388	0.1005	0.6323	0.8757
	(0.8030)	(0.9510)	(0.7290)	(0.6454)

Note: A portfolio composed of three assets is considered, the returns being observed for the period between June 1, 2007 and June 1, 2009. In a first case, the series of VaR for the portfolio is issued from a DCC model while in a second case from a t-Garch model. 13 conditional coverage tests are used, namely 7 dynamic binary specifications DB, 6 DQ tests (Engle and Manganelli, 2004) including several lags of the violations variable and VaR, as well as Christoffersen's LR_{CC} test. These tests are applied both at a 5% and 1% coverage rate respectively. The corresponding p-values are presented in the parentheses.