Testing interval forecasts: a GMM-based approach [*]

Elena-Ivona Dumitrescu [†] Christophe Hurlin [‡] Jaouad Madkour [§]

August 2011

Abstract

This paper proposes a new evaluation framework for interval forecasts. Our model free test can be used to evaluate intervals forecasts and High Density Regions, potentially discontinuous and/or asymmetric. Using a simple J-statistic, based on the moments defined by the orthonormal polynomials associated with the Binomial distribution, this new approach presents many advantages. First, its implementation is extremely easy. Second, it allows for a separate test for unconditional coverage, independence and conditional coverage hypotheses. Third, Monte-Carlo simulations show that for realistic sample sizes, our GMM test has good small-sample properties. These results are corroborated by an empirical application on SP500 and Nikkei stock market indexes. It confirms that using this GMM test leads to major consequences for the ex-post evaluation of interval forecasts produced by linear versus nonlinear models.

Key words: Interval forecasts, High Density Region, GMM.

J.E.L Classification : C52, G28.

[*] The authors thank Gilbert Colletaz for comments on the paper as well as the participants of the 65th European Meeting of the Econometric Society (ESEM, Oslo, August 2011), the 43th Journées de Statistique (French Society of Statistics, Tunis, May 2011) and the 59th congress of the French Economic Association (AFSE, Nanterre, September 2010). Usual disclaimers apply.

[†] elena.dumitrescu@univ-orleans.fr, Maastricht University and LEO, University of Orléans, Laboratoire d'Economie d'Orléans (LEO), Rue de Blois - BP 6739. 45067 Orléans Cedex 2, France.

[‡] christophe.hurlin@univ-orleans.fr, University of Orléans, Laboratoire d'Economie d'Orléans (LEO), Rue de Blois, BP 6739, 45067, Orleans Cedex 2, France.

[§] jaouad.madkour@etu.univ-orleans.fr, University of Orléans, Laboratoire d'Economie d'Orléans (LEO), Rue de Blois, BP 6739, 45067, Orleans Cedex 2, France.

1 Introduction

In recent years, the contribution of nonlinear models to forecasting macroeconomic and financial series has been intensively debated (see Teräsvirta, 2006, Colletaz and Hurlin, 2005 for a survey). A suggested by Teräsvirta, there are relatively numerous studies in which the forecasting performance of nonlinear models is compared with that of linear models using actual series. In general, no dominant nonlinear (or linear) model has emerged. However, the use of nonlinear models has actually led to the renewal of the forecasting approach, especially through the emergence of concepts like High Density Regions (Hyndman, 1995, thereafter HDR) or density forecasts as opposed to point forecasts. Consequently, this debate on non-linearity and forecasting involves new forecast validation criteria. It is the case of density forecasts, for which many specific evaluation tests have been developed (Bao, Lee and Saltoglu, 2004, Corradi and Swanson 2006 etc.).

On the contrary, if there are numerous methods to calculate HDR and interval forecasts (Chatfield, 1993), only a few studies propose validation methods adapted to these kind of forecasts. This paradox is even more astounding if we take into consideration the fact that interval forecast is the most generally used method by applied economists to account for forecast uncertainty.

One of the main exceptions, is the seminal paper of Christoffersen (1998), that introduces general definitions of hypotheses allowing to assess the validity of an interval forecast obtained by using any type of model (linear or nonlinear). His model-free approach is based on the concept of violation: a violation is said to occur if the *ex-post* realization of the variable does not lie in the *ex-ante* forecast interval. Three validity hypothesis are then distinguished. The *unconditional coverage* hypothesis means that the expected frequency of violations is precisely equal to the coverage rate of the interval forecast. The *independence hypothesis* means that if the interval forecast is valid then violations must be distributed independently. In other words, there must not be any cluster in the violations sequence. Finally, under the *conditional coverage hypothesis* the violation process satisfies the assumptions of a martingale difference. Based on these definitions, Christoffersen proposes a Likelihood Ratio (hereafter LR) test for each of these hypotheses, by considering a binary first-order Markov chain representation under the alternative hypothesis.

More recently, Clements and Taylor (2002) applied a simple logistic regression with periodic dummies and modified the first-order Markov chain approach in order to detect dependence at a periodic lag. In 2003, Wallis recast Christoffersen (1998)'s tests in the framework of contingency tables increasing users' accessibility to these interval forecast evaluation methods. Owing to his innovative approach, it became possible to calculate exact p-values for the LR statistics in small sample cases.

Beyond their specificities, the main common characteristic of these tests is that assessing the validity of interval forecasts comes down to testing a distributional assumption for the violation process. If we define a binary indicator variable that takes the value one in case of violation, and zero otherwise, it is obvious that under the null of conditional coverage, the sum of the indicators associated to a sequences of interval forecasts follows a Binomial distribution.

On these grounds, we propose in this paper a GMM approach to test the interval forecasts and

HDR validity. To be more precise, we propose to test interval forecast using discrete polynomials. The series of violations, I_t, (a violation indicates whether the forecast belongs to the $1-\alpha$ confidence interval or not) is splitted into blocks of size N. The sum of I_t within each block follows a binomial distribution $B\left(N,\alpha\right)$. The test consists in testing that the series of sums is indeed a *i.i.d.* sequence of random variables which are binomially distributed. Relying on the GMM framework of Bontemps and Meddahi (2005), we propose simple J-statistics based on particular moments defined by the orthonormal polynomials associated with the Binomial distribution. A similar approach has been used by Candelon et al. (2011) in the context of the Value-at-Risk[1] backtesting. The authors test the VaR forecasts validity by testing the geometric distribution assumption for the durations observed between two consecutive VaR violations. Here, we propose a general approach for all kind of intervals and HDR forecasts, that directly exploits the properties of the violation process (and not the durations between violations). We adapt the GMM framework to the case of discrete distributions and more exactly to a binomial distribution.

Our approach has several advantages. First, we develop an unified framework in which the three hypotheses of unconditional coverage, independence and conditional coverage are tested independently. Second, this approach imposes no restrictions under the alternative hypothesis. Third, this GMM-based test is easy to implement and does not generate computational problems regardless of the sample size. Finally, some Monte-Carlo simulations indicate that for realistic sample sizes, our GMM test have good power properties.

The paper is structured as follows. Section 2 presents the general framework of interval forecast evaluation, while section 3 introduces our new GMM-based evaluation tests. In section 4 we scrutinize the finite-sample properties of the tests through Monte-Carlo simulations and in section 5 we propose an empirical application. Section 6 concludes.

2 General Framework

Formally, let $x_t, t \in \{1, ..., T\}$ be a sample path of a time series x_t. Let us denote by $\left\{C_{t|t-1}(\alpha)\right\}_{t=1}^{T}$ the sequence of *out-of-sample* interval forecasts for the coverage probability α, so that

$$\Pr[x_t \in C_{t|t-1}(\alpha)] = \alpha. \tag{1}$$

Hyndman (1995) identifies three methods to construct a $100(1-\alpha)\%$ forecast region: (i) a symmetrical interval around the point forecast, (ii) an interval defined by the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the forecast distribution, (i) and a High Density Region (HDR). These three forecast regions are identical (symmetric and continuous) in the case of symmetric and unimodal distribution. By contrast, HDR_α is the smallest forecast region for asymmetric or multimodal distributions. When the interval forecast is continuous, $C_{t|t-1}(\alpha)$ can be defined as in Christoffersen (1998), by $C_{t|t-1}(\alpha) = [L_{t|t-1}(\alpha), U_{t|t-1}(\alpha)]$, where $L_{t|t-1}(\alpha)$ and $U_{t|t-1}(\alpha)$ are the limits of the *ex-ante* confidence interval for the coverage rate α.

[1] Note that the Value-at-Risk can be interpreted as a one-sided and open interval.

Whatever the form of the HDR or the interval forecasts (symmetric or asymmetric, continuous or discontinuous), we define an indicator variable $I_t(\alpha)$, also called violation, as a binary variable that takes a value one if the realization of x_t does not belong to this region:

$$I_t(\alpha) = \begin{cases} 1, & x_t \notin C_{t|t-1}(\alpha) \\ 0, & x_t \in C_{t|t-1}(\alpha) \end{cases}. \tag{2}$$

Based on the definition of the violations process, a general testing criterion for interval forecasts can be established. Indeed, as stressed by Christoffersen (1998), the interval forecasts are valid if and only if the conditional coverage (CC hereafter) hypothesis is fulfilled, implying that both the independence (IND hereafter) and unconditional coverage (UC hereafter) hypotheses are satisfied. Under the UC assumption, the probability to have a violation must be equal to the α coverage rate:

$$H_{0,UC} : \Pr[I_t(\alpha) = 1] = \mathbb{E}[I_t(\alpha)] = \alpha. \tag{3}$$

Under the IND hypothesis, violations observed at different moments in time for the same coverage rate (α%) must be independent. In other words, we do not observe any clusters of violations and past violations should not be informative about the present or future violations. The UC property places a restriction on how often violations may occur, whereas the IND assumption restricts the order in which these violations may appear.

Christoffersen (1998) pointed out that in the presence of higher-order dynamics it is important to go beyond the UC assumption and test the CC hypothesis. Under the CC assumption, the conditional (on a past information set Ω_{t-1}) probability to observe a violation must be equal to the α coverage rate, *i.e.* the I_t process satisfies the properties of a martingale difference:

$$H_{0,CC} : \mathbb{E}[I_t(\alpha) \mid \Omega_{t-1}] = \alpha. \tag{4}$$

Christoffersen considers an information set Ω_{t-1} that consists of past realizations of the indicator sequence $\Omega_{t-1} = \{I_{t-1}, I_{t-2}, .., I_1\}$. In this case, testing $\mathbb{E}[I_t(\alpha) \mid \Omega_{t-1}] = \alpha$ for all t is equivalent to testing that the sequence $\{I_t(\alpha)\}_{t=1}^{T}$ is identically and independently distributed Bernoulli with parameter α. So, a sequence of interval/HDR forecasts $\left\{C_{t|t-1}(\alpha)\right\}_{t=1}^{T}$ has correct conditional coverage, if:

$$I_t \overset{i.i.d}{\sim} Bernouilli(\alpha), \ \forall t. \tag{5}$$

This feature of the violation process is actually at the core of most of the interval forecast evaluation tests (Christoffersen, 1998, Clements and Taylor, 2002, etc.) and so it is for our GMM-based test.

3 A GMM-Based Test

In this paper we propose a unified GMM framework for evaluating interval forecasts and HDR by testing the Bernoulli distributional assumption of the violation series $I_t(\alpha)$. Our analysis is based on the recent GMM distributional testing framework developed by Bontemps and Meddahi (2005) and Bontemps (2006). We first present the environment of the test, then we define the moment conditions used to test the interval forecasts efficiency, and finally we propose simple J-statistics corresponding to the three hypotheses of UC, IND and CC.

3.1 Environment Testing

Given the result (5), it is obvious that if the interval forecast has a correct conditional coverage, the sum of violations follows a Binomial distribution

$$H_{0,CC} : \sum_{t=1}^{T} I_t\left(\alpha\right) \sim B(T, \alpha). \tag{6}$$

A natural way to test CC, consists in testing this distributional assumption. However this property cannot be directly used to develop an implementable testing procedure, since, for a given sequence $\left\{C_{t|t-1}(\alpha)\right\}_{t=1}^{T}$, we have only one observation for the sum of violations.

Therefore, we propose to divide the sample of violations into blocks. Since under the null hypothesis the violations $\{I_t(\alpha)\}_{t=1}^{T}$ are independent, it is possible to split the initial series of violations into H blocks of size N, where $H = [T/N]$ (see Figure 1).

[Insert Figure 1]

The sum of I_t within each block follows a binomial distribution $B\left(N, \alpha\right)$. More formally, for each block, we define y_h, $h \in \{1, ..., H\}$ as the sum of the corresponding N violations:

$$y_h = \sum_{t=(h-1)N+1}^{hN} I_t(\alpha). \tag{7}$$

As a result, under the null hypothesis, the constructed processes y_h are *i.i.d.* $B(N, \alpha)$, and thus the null of CC that the interval forecasts are well specified can simply be expressed as follows:

$$H_{0,CC} : y_h \sim B(N, \alpha),\ \forall h \in \{1, ..., H\}. \tag{8}$$

This approach can be compared to the sub-sampling methodology proposed by Politis, Romano and Wolf, (1999). However, the objective here is entirely different. In our case, we do not aim to obtain the finite sample distribution of a particular test statistic. We only divide the initial sample of T violations into H blocks of size N in order to compute our CC test (which is a simple distributional test). In other words, we choose the distributional assumption that we want to test. In order to test the CC assumption, we propose to test the $B\left(N, \alpha\right)$ distribution rather than the $B\left(T, \alpha\right)$ one, even if theoretically both approaches are possible. The advantages of this approach will be presented in the next sections, in the specific context of the test that we propose.

3.2 Orthonormal Polynomials and Moment Conditions

There are many ways to test conditional coverage hypothesis through the distributional assumption (8). Following Bontemps and Meddahi (2005) and Bontemps (2006), we propose here to use a GMM-based framework. The general idea is that for many continuous and discrete distributions, it is possible to associate some particular orthonormal polynomials whose expectation is equal to zero. These orthonormal polynomials can be used as moment conditions in a GMM framework to test for a specific distributional assumption. For instance, the Hermite polynomials associated to the normal distribution can be employed to build a test for normality (Bontemps and Meddahi, 2005). Other particular polynomials are used by Candelon et al. to test for a geometric distribution hypothesis.

In the particular case of a Binomial distribution, the corresponding orthonormal polynomials are called Krawtchouk polynomials. These polynomials are defined as follows:

Definition 1. *Let us consider a discrete random variable y_h such that $y_h \sim B(N, \alpha)$. The corresponding orthonormal Krawtchouk polynomials are defined by the following recursive relationship*:

$$P_{j+1}^{(N,\alpha)}(y_h) = \frac{\alpha(N-j)+(1-\alpha)j-y_h}{\sqrt{\alpha(1-\alpha)(N-j)(j+1)}} P_j^{(N,\alpha)}(y_h) - \sqrt{\frac{j(N-j+1)}{(j+1)(N-j)}} P_{j-1}^{(N,\alpha)}(y_h),$$

where $j < N$ and $P_{-1}^{(N,\alpha)}(y_h) = 0$, $P_0^{(N,\alpha)}(y_h) = 1$ verify

$$\mathbb{E}\left[P_j^{(N,\alpha)}(y_h)\right] = 0, \quad \forall j < N. \tag{9}$$

Our test exploits these moment conditions. More precisely, let us define $\{y_1; ...; y_H\}$ a sequence of sums defined by (7) and computed from the sequence of violations $\{I_t(\alpha)\}_{t=1}^T$. Under the null of conditional coverage, variables y_h are *i.i.d.* and have a Binomial distribution $B(N, \alpha)$, where N is the block size. Hence, the null of CC can be expressed as follows:

$$H_{0,CC}: \ \mathbb{E}\left[P_j^{(N,\alpha)}(y_h)\right] = 0, \quad j = \{1, .., m\}, \tag{10}$$

with the number of moment conditions $m < N$. The expressions of the first two polynomials are the following:

$$P_1^{(N,\alpha)}(y_h) = \frac{\alpha N - y_h}{\sqrt{\alpha(1-\alpha)N}}, \tag{11}$$

$$P_2^{(N,\alpha)}(y_h) = \left(\frac{\alpha(N-1)+(1-\alpha)-y_h}{\sqrt{\alpha(1-\alpha)2(N-1)}}\right)\left(\frac{\alpha N - y_h}{\sqrt{\alpha(1-\alpha)N}}\right) - \sqrt{\frac{N}{2(N-1)}}. \tag{12}$$

An appealing property of the test is that it allows to test separately for the UC and IND hypotheses. Let us remind that under the UC assumption, the unconditional probability to have a violation is equal to the coverage rate α. Consequently, under UC, the expectation of the sum y_h is then equal to αN, since:

$$\mathbb{E}\left(y_h\right)=\sum_{t=(h-1)N+1}^{hN}\mathbb{E}\left[I_t\left(\alpha\right)\right]=\alpha N,\ \forall h\in\{1,...,H\}. \tag{13}$$

Given the properties of the Krawtchouk polynomials, the null UC hypothesis can be expressed as:

$$H_{0,UC}:\ \mathbb{E}\left[P_1^{(N,\alpha)}(y_h)\right]=0. \tag{14}$$

In this case, we need to use only the first moment condition defined by $P_1^{(N,\alpha)}(y_h)$, since the condition $\mathbb{E}\left[P_1^{(N,\alpha)}(y_h)\right]=0$ is equivalent to the UC condition $\mathbb{E}\left(y_h\right)=\alpha N$ or $\mathbb{E}\left(I_t\right)=\alpha$.

Under the IND hypothesis, the violations are independently and identically distributed, but their probability is not necessarily equal to the coverage rate α. Let us denote β the violation probability. If the violations are independent, the sum y_h follows a $B(N,\beta)$ distribution, where β may be different from α. Thus, the IND hypothesis can simply be expressed as:

$$H_{0,IND}:\ \mathbb{E}\left[P_j^{(N,\beta)}\left(y_h\right)\right]=0 \quad j=\left\{1,..,m\right\}, \tag{15}$$

with $m<N$.

3.3 Testing Procedure

Let $P^{(N,\alpha)}$ denote a $(m,1)$ vector whose components are the orthonormal polynomials $P_j^{(N,\alpha)}\left(y_h\right)$, for $j=1,..,m$, associated with the Binomial distribution $B\left(N,\alpha\right)$. Under the CC assumption and some regularity conditions (Hansen, 1982) it can be shown that:

$$\left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}P^{(N,\alpha)}(y_h)\right)'\Sigma^{-1}\left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}P^{(N,\alpha)}(y_h)\right)\underset{H\to\infty}{\overset{d}{\to}}\chi^2(m), \tag{16}$$

where Σ is the long-run variance-covariance matrix of $P^{(N,\alpha)}(y_h)$. By the definition of orthonormal polynomials, this long-run variance-covariance matrix corresponds to the identity matrix.[2] Therefore, the corresponding J-statistic is very easy to implement. Let us denote by $J_{CC}(m)$ the CC test-statistic associated with the $(m,1)$ vector of orthonormal polynomials $P^{(N,\alpha)}(y_h)$.

Definition 2. *Under the null hypothesis of conditional coverage, the CC test statistic verifies*:

$$J_{CC}(m)=\frac{1}{H}\sum_{j=1}^{m}\left(\sum_{h=1}^{H}P_j^{(N,\alpha)}(y_h)\right)^2\underset{H\to\infty}{\overset{d}{\to}}\chi^2(m), \tag{17}$$

where $P_j^{(N,\alpha)}(y_h)$ denotes the Krawtchouk polynomial corresponding to a Binomial distribution $B(N,\alpha)$ of order j, for $j\leq m$.

Proof : see appendix A. Since the $J_{UC}(m)$ statistic corresponding to the UC hypothesis is a

[2] If we neglect this property, it is also possible to use the Kernel estimate of the long-run variance covariance matrix as it is usually done in the GMM literature.

special case of the $J_{CC}(m)$ test statistic, it can be immediately computed by taking into account only the first moment condition $\mathbb{E}\left[P_1^{(N,\alpha)}(y_h)\right] = 0$, and can be expressed as follows:

$$J_{UC} = J_{CC}(1) = \frac{1}{H}\left(\sum_{h=1}^{H} P_1^{(N,\alpha)}(y_h)\right)^2 \underset{H\to\infty}{\overset{d}{\to}} \chi^2(1), \tag{18}$$

Finally, the independence hypothesis statistic, denoted $J_{IND}(m)$ takes the form of:

$$J_{IND}(m) = \frac{1}{H}\sum_{j=1}^{m}\left(\sum_{h=1}^{H} P_j^{(N,\beta)}(y_h)\right)^2 \underset{H\to\infty}{\overset{d}{\to}} \chi^2(m), \tag{19}$$

where $P_j^{(N,\beta)}(y_h)$ is the orthonormal polynomial of order $j \leq m$ associated with a Binomial distribution $B(N,\beta)$, where β can be different from α. The coverage rate β is generally unknown, and thus it has to be estimated. When using a consistent estimator $\hat{\beta} = (1/T)\sum_t^T I_t(\alpha)$ instead of β, the degree of freedom of the GMM-statistic $J_{IND}(m)$ has to be adjusted accordingly:

$$J_{IND}(m) = \frac{1}{H}\sum_{j=1}^{m}\left(\sum_{h=1}^{H} P_j^{(N,\widehat{\beta})}(y_h)\right)^2 \underset{H\to\infty}{\overset{d}{\to}} \chi^2(m-1), \tag{20}$$

where $P_j^{(N,\widehat{\beta})}(y_h)$ is the orthonormal polynomial of order j associated to a Binomial distribution $B\left(N,\widehat{\beta}\right)$ and $\widehat{\beta}$ is the estimated coverage rate.

Our block-based approach has many advantages for testing the validity of interval forecasts, especially in finite samples with relatively small size (as it will be shown in the Monte Carlo simulation section). First, let us consider without loss of generality the case of two moment conditions. The test statistic $J_{CC}(2)$ based on $P_1^{(N,\alpha)}(y_h)$ and $P_2^{(N,\alpha)}(y_h)$, can be viewed as a function of both y_h and y_h^2 which, once expanded, involves the cross product $I_t(\alpha)\,I_s(\alpha)$ for two periods t and s within a given block. When the block size N is equal to 2, $J_{CC}(2)$ is close to the joint test of Christoffersen. When $N = 3$, the test statistic involves the product (i.e. correlation) between $I_{t-2}(\alpha)$. $I_{t-1}(\alpha)$ and $I_t(\alpha)$ and more generally, for any N, it includes the correlations between $I_{t-h}(\alpha)$ for $h = 1, .., N$ and $I_t(\alpha)$. Consequently we expect that our approach will reveal some dependencies that cannot be identified by Christoffersen's approach.

Second, when the block size N is small, H is relatively important, and many observations of the sums y_h are available. The finite sample distribution of the J-statistic is then close to its asymptotic chi-squared distribution. On the contrary, when N is large compared to T, the binomial distribution $B(N,\alpha)$ can be approximated by a normal distribution. Then, each sum y_h has also a normal distribution and their sum of squares has a chi-squared distribution. Consequently, as we will show in the next section, the J-statistics have a finite chi-squared distribution.

4 Monte-Carlo Experiments

In this section we gauge the finite sample properties of our GMM-based test using Monte-Carlo experiments. We first analyze the size performance of the test and we then investigate its empirical power in the same framework as in Berkowitz et al. (2010). A comparison with Christoffersen (1998)'s LR tests is provided for both analyses. In order to control for size distortions, we use Dufour (2006)'s Monte-Carlo method.

4.1 Empirical Size Analysis

To illustrate the size performance of our UC and CC tests in finite sample, we generate a sequence of T violations by taking independent draws from a Bernoulli distribution, considering successively a coverage rate $\alpha = 1\%$ and $\alpha = 5\%$. Several sample sizes T ranging from 250 (which roughly corresponds to one year of daily forecasts) to $1,500$ are considered. The size of the blocks (used to compute the H sums y_h) is fixed to $N = 100$ or $N = 25$ observations. Additionally, we consider several moment conditions m from 1 (for the UC test statistic J_{UC}) to 5. Based on a sequence $\{y_h\}_{h=1}^{H}$, with $H = [T/N]$, we compute both statistics J_{UC} and $J_{CC}\left(m\right)$. The reported empirical sizes correspond to the rejection rates calculated over $10,000$ simulations for a nominal size equal to 5%.

[Insert Table 1]

In table 1, the rejection frequencies for the $J_{CC}(m)$ statistic and a block size N equal to 100 are presented. For comparison reasons, the rejection frequencies for the Christoffersen (1998)'s LR_{UC} and LR_{CC} test statistics are also reported. For a 5% coverage rate and whatever the choice of m, the empirical size of the J_{CC} test is close to the nominal size, even for small sample sizes. For a 1% VaR, the J_{CC} test is also well sized, whereas the LR_{CC} test seems to be undersized in small samples (especially for $\alpha = 1\%$), although it size converges to the nominal one as T increases.[3] On the contrary, the performance of our J_{UC} statistic and the LR_{UC} are quite comparable (especially for $T \geq 500$). It can be proved that J_{UC} is a local expansion of the unconditional test of Christoffersen. Indeed, our statistic can be expressed as a simple function of the sample size and the total number of hits $\sum_{h=1}^{H} y_h$, since :

$$J_{UC} = \frac{1}{H}\left(\sum_{h=1}^{H}\frac{y_h - N\alpha}{\sqrt{\alpha\left(1-\alpha\right)N}}\right)^2 = \frac{T}{\alpha\left(1-\alpha\right)}\left(\alpha - \frac{1}{T}\sum_{h=1}^{H} y_h\right)^2 . \tag{21}$$

[3] Berkowitz et al. (2010) and Candelon et al. (2011) found that the LR_{CC} is oversized in small sample. The difference comes from the computation of the LR statistic. Under H_1, the computation of the LR_{CC} statistic requires calculating the sum of joint violations $I_t\left(\alpha\right)$ and $I_{t-1}\left(\alpha\right)$. Consequently, the size of the available sample is equal to $T-1$. On the contrary, under H_0, the likelihood depends on the sample size and the coverage rate α. Contrary to previous studies, we compute the likelihood under H_0 by adjusting the sample size to $T-1$. This slight difference explains the differences in the results. By considering a sample size T under H_0, we get exactly the same empirical size as in Berkowitz and al. (2010) or Candelon et al. (2011).

The performance of our test is quite remarkable, since under the null, in a sample with $T = 250$ and a coverage rate equal to 1%, the expected number of violations lies between two and three. It is worth noting that even if our asymptotic result requires that the number of blocks H tends to infinity, our testing procedure works even with very small H values. Indeed, when the block size is substantial there is also an asymptotic normality that explains these results. For instance, let us consider the UC statistic J_{UC}, defined by the first orthonormal polynomial $P_1^{(N,\alpha)}$. For $N = 100$ and $\alpha = 0.05$, the binomial distribution can be approximated by a normal distribution (since $N\alpha \geq 5$, $N(1-\alpha) \geq 5$ and $N > 30$), so that under UC:

$$P_1^{(N,\alpha)}(y_h) = \frac{y_h - N\alpha}{\sqrt{\alpha(1-\alpha)N}} \sim N(0,1), \ \forall h \in \{1, ..., H\}. \tag{22}$$

Consequently, for $H = 1$ ($N = T$), it is obvious that our J_{UC} statistic (equation 18) has a chi-squared distribution:

$$J_{UC} = \left[P_1^{(N,\alpha)}(y_1)\right]^2 \sim \chi^2(1). \tag{23}$$

For values of $H > 1$, we have the same result. For instance let us consider the case where $H = 2$, *i.e.* where the block size N is equal to $T/2$. Then, the J_{UC} statistic is defined as follows (equation 18):

$$J_{UC} = \frac{1}{2}\left[P_1^{(N,\alpha)}(y_1) + P_1^{(N,\alpha)}(y_2)\right]^2, \tag{24}$$

or equivalently by

$$J_{UC} = \left(\frac{P_1^{(N,\alpha)}(y_1) + P_1^{(N,\alpha)}(y_2)}{\sqrt{2}}\right)^2, \tag{25}$$

where $P_1^{(N,\alpha)}(y_1) + P_1^{(N,\alpha)}(y_2)$ is the sum of two independent standard normal variables provided that the blocks are independent. So, under the UC assumption, we have:

$$\frac{P_1^{(N,\alpha)}(y_1) + P_1^{(N,\alpha)}(y_2)}{\sqrt{2}} \sim N(0,1), \tag{26}$$

and consequently $J_{UC} \sim \chi^2(1)$.

The same type of results can be observed when the block size N is decreased. The rejection frequencies of the Monte-Carlo experiments for the $J_{CC}(m)$ GMM-based test statistic, both for a coverage rate of 5% and of 1% and for a block of size 25 are reported in table 2. In that case, the normal approximation of the binomial distribution is not valid (since $N\alpha = 1.25$ or 0.25 given the values of α) and cannot be invoked to explain the quality of the results of our test. However, the number of observations H increases for a given size T (relatively to the previous case $N = 100$), so the J statistics converge more quickly to a chi-squared distribution.

[Insert Table 2]

It is important to note that these rejection frequencies are only calculated for the simulations providing a LR test statistic. Indeed, for realistic sample size (for instance $T = 250$) and a coverage rate of 1%, some simulations do not deliver a LR statistic. The LR_{CC} test statistic is computable only if there is at least one violation in the sample. Thus, at a 1% coverage rate for which the scarcity of violations is more obvious, a large sample size is required in order to compute this test statistic. The fraction of samples for which a test is feasible is reported for each sample size, both for the size and power experiments (at 5% and 1% coverage rate), are reported in table 3. By contrast, our GMM-based test can always be computed as long as the number of moment conditions m is inferior or equal to the block size N. It is one of the advantages of our approach.

[Insert Table 3]

4.2 Empirical Power Analysis

We now investigate the empirical power of our GMM test, especially in the context of risk management. As previously mentioned, Value-at-Risk (VaR) forecasts can be interpreted as one-sided and open forecast intervals. More formally, let us consider an interval $CI_{t|t-1}(\alpha) = [-\infty, VaR_{t|t-1}(\alpha)]$, where $VaR_{t|t-1}(\alpha)$ denotes the conditional VaR obtained for a coverage (or risk) equal to α%. As usual in the backtesting literature, our power experiment is based on a particular DGP for financial returns and a method to compute VaR out-of-sample forecasts. This method has to be chosen to produce invalid VaR forecasts according to Christoffersen's hypotheses.

Following Berkowitz et al. (2010), we assume that returns r_t are issued from a simple t-GARCH model with an asymmetric leverage effect:

$$r_t = \sigma_t z_t \sqrt{\frac{\nu - 2}{\nu}}, \tag{27}$$

where 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:

$$\sigma_t^2 = \omega + \gamma \sigma_{t-1}^2 \left(\sqrt{\frac{\nu - 2}{\nu}} z_{t-1} - \theta \right)^2 + \beta \sigma_{t-1}^2. \tag{28}$$

Once the returns series has been generated, a method of VaR out-of-sample forecasting must be selected.[4] Obviously, this choice has deep implications in terms of power performance for the interval forecast evaluation tests. We consider the same method as in Berkowitz et al. (2010), *i.e.* the historical simulation (HS), with a rolling window size T_e equal to 250. This unconditional forecasting method generally produces clusters of violations (violation of the independence assumption), and some slight deviations from the unconditional coverage assumption when we consider out-of-sample forecasts (these deviations depend on the size of the rolling window). Formally, we define the HS-VaR as following:

$$VaR_{t|t-1}(\alpha) = Percentile\left(\{r_i\}_{i=t-T_e}^{t-1}, 100\alpha \right). \tag{29}$$

[4] The coefficients of the model are parametrized as in Berkowitz et al. (2010) : $\gamma = 0.1$, $\theta = 0.5$, $\beta = 0.85$, $\omega = 3.9683e^{-6}$ and $d = 8$. At the same time, ω has been chosen so as to be consistent with a 0.2 annual standard deviation. Additionally, the global parametrization corresponds to a daily volatility persistence of 0.975.

For each simulation, a violation sequences $\{I\}_{t=1}^{T}$ is then constructed, by comparing the *ex-ante* $VaR_{t|t-1}(\alpha)$ forecasts to the *ex post* returns r_t. Next, the sequence $\{y_h\}_{h=1}^{H}$ is computed for a given block size N by summing the corresponding I_t observations (see section 3.1). Based on this sequence, the J_{CC} test statistics are then implemented for different number of moment conditions and sample sizes T ranging from 250 to 1500. For comparison, both LR_{UC} and LR_{CC} statistics are also computed for each simulation. The rejection frequencies, at a 5% nominal size, are based on 10,000 simulations. In order to control for size distortions between LR and J_{CC} tests and to get a fair power comparison, we use Dufour (2006)'s Monte-Carlo method (see appendix B).

[Insert Tables 4 and 5]

Tables 4 and 5 report the corrected power of the J_{UC}, $J_{CC}(m)$, LR_{UC} and LR_{CC} tests for different sample sizes T, in the case of a 5% and a 1% coverage rate, both for a block size $N = 100$ and $N = 25$. We can observe that the two GMM-based tests (J_{UC} and J_{CC}) have good small sample power properties, whatever the sample size T and the block size N considered. Additionally, our test is proven to be quite robust to the choice of the number of moment conditions m. Nevertheless, in our simulations it appears that the optimal power of our GMM-based test is reached when considering two or three moment conditions. For a 5% coverage rate, a sample size $T = 250$ and a block size $N = 25$, the power of our $J_{CC}(2)$ test statistic is approximately two times the power of the corresponding LR test for that experiment. For a coverage rate $\alpha = 1\%$, the power of our $J_{CC}(2)$ test remains by 30% higher than the one of the LR test. On the contrary, our unconditional coverage J_{UC} test does not outperform the LR test. This result is logical, since both exploit approximately the same information, *i.e.* the frequency of violations. Note that for UC tests (J and LR tests), the empirical power is decreasing, contrary to the CC tests. This result is specific to that experiment and comes from the use of the historical simulation to produce out-of-sample VaR forecasts. For large T sample, the deviation from the CC mainly comes from the clusters of violations. The empirical frequencies of hits is then close to the nominal coverage rate α.

The choice of the block size N has two opposite effects on the empirical power. A decrease in the block size N leads to an increase in the length of the sequence $\{y_h\}_{h=1}^{H}$ used to compute the J-statistic, and then leads to an increase in its empirical power. On the contrary, when the block size N increases, the normal approximation of the binomial distribution is more accurate. Thus, the finite sample distribution of our J statistics is close to the chi-squared distribution. This result is not due to the number of observations H, but to the normal approximation of the binomial. Figure 2 displays the Dufour's corrected empirical power of the $J_{CC}\,(2)$ statistic as a function of the sample size T, for three values $(2, 25$ and $100)$ of the block size N. We note that, whatever the sample size, the power for a block size $N = 100$ is always lesser than that obtained for a block size equal to 25. In the same time, the power with $N = 100$ is always larger than that with $N = 2$. In order to get a more precise idea of the link between the power and the block size N, the Figure 3 displays the Dufour's corrected empirical power of the $J_{CC}\,(2)$ statistic as a function of the block size N, for three values $(250, 750$ and $1500)$ of the sample size T. The highest corrected power corresponds to block sizes between 20 and 40, that is why we recommend a value of $N = 25$ for applications. Other simulations based on Bernoulli trials with a false coverage rate (available upon request) confirm this choice.

[Insert Figures 2 and 3]

Thus, our new GMM-based interval forecasts evaluation tests seems to perform better both in terms of size and power than the traditional LR ones.

5 An Empirical Application

Now, we propose an empirical application based on two series of daily returns, namely the SP500 (from 05 January 1953 to 19 December 1957) and the Nikkei (from 27 January 1987 to 21 February 1992). The baseline idea is to select some periods and assets for which the linearity assumption is strongly rejected by standard specification tests. Then, we use (at wrong) a linear model to produce a sequence of invalid interval forecasts. The issue is then to check if our evaluation tests are able to reject the nulls of UC, IND and/or CC.

Here we use the nonlinearity test recently proposed by Harvey and Laybourne (2007). This takes into account both an ESTAR or LSTAR alternative hypothesis, and has very good small sample properties. For the considered periods, the conclusion of the test are clear: the linearity assumption is strongly rejected for both assets. For the SP500 (respectively Nikkei), the statistic is equal to 24.509 (respectively 89.496) with a p-value less than 0.001. As previously mentioned, we use simple autoregressive linear models AR(1) to produce forecasts and interval forecasts at an horizon $h = 1, 5$ or 10 days. More precisely, each model is estimated on the first 1,000 in sample observations, while continuous and symmetrical confidence intervals are computed for each sequence of 250 out-of-sample observations both at a 5% and 1% coverage rate.

[Insert Tables 6 and 7]

Tables 6 and 7 report the main results of the interval forecast tests, based on a block size N equal to 25. It appears that for the SP500 index (see Table 6) our GMM-based test always rejects the CC hypothesis and thus, the validity of the forecasts. In this case, the LR_{CC} test does not reject this hypothesis for a 5% coverage rate. When considering a 1% coverage rate, both CC tests succeed in rejecting the null hypothesis. Still, further clarifications are required. Both the UC and IND hypothesis are rejected when using GMM-based tests, whereas the only assumption rejected by the LR tests is the UC one. Similar results are obtained for the Nikkei series (see Table 7). Thus, the two series of interval forecasts are characterized by clusters of violations detected only by our GMM-based test. On the contrary, the LR_{IND} test appears not to be powerful enough to reject the independence assumption. This analysis proves that our evaluation tests for interval forecasts have interesting properties for applied econometricians, especially when they have to evaluate the validity of interval forecasts on short samples.

6 Conclusion

This paper proposes a new evaluation framework of interval and HDR forecasts based on simple J-statistics. Our test is model free and can be applied to intervals and/or HDR forecasts, potentially discontinuous and/or asymmetric. The underlying idea is that if the interval forecast is correctly specified, then the sum of the violations should be distributed according to Binomial distribution with a success probability equal to the coverage rate. So, we adapt the GMM framework proposed by Bontemps (2006) in order to test for this distributional assumption that corresponds to the null of interval forecast validity.

More precisely, we propose an original approach that transforms the violation series into a series of sums of violations defined for H blocks of size N. Under the null of validity, these sums are distributed according to a Binomial distribution.

Our approach has several advantages. First, all three hypotheses of unconditional coverage, independence and conditional coverage can be tested independently. Second, these tests are easy to implement. Third, Monte-Carlo simulations show that all our GMM-based tests have good properties in terms of power, especially in small samples and for a 5% coverage rate (95% interval forecasts), which are the most interesting cases from a practical viewpoint.

Assessing the impact of the estimation risk for the parameters of the model that generated the HDR or the interval forecasts (and not for the distributional parameters) on the distribution of the GMM test-statistic by using a subsampling approach or a parametric correction is left for future research.

Bibliography

Références

[1] Berkowitz, J., Christoffersen, P., Pelletier, D., (2010). Evaluating Value-at-Risk Models with Desk-Level Data, forthcoming in *Management Science*.

[2] Bontemps, C., (2006). Moment-based tests for discrete distributions, *Working Paper TSE.*

[3] Bontemps, C., and Meddahi, N., (2005). Testing normality: a GMM approach, *Journal of Econometrics,* 124, 149-186.

[4] Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S., (2011). Backtesting Value-at-Risk : a GMM duration-based test, forthcoming in *Journal of Financial Econometrics*.

[5] Chatfield, C., (1993). Calculating interval forecasts, *Journal of Business and Economic Statistics*, 11, issue 2, 121-135.

[6] Christoffersen, F.P., (1998). Evaluating interval forecasts, *International Economic Review*, 39, 841-862.

[7] Clements, M.P., Taylor, N. (2002). Evaluating interval forecasts of high-frequency financial data, *Journal of Applied Econometrics*, 18, Issue 4, 445 - 456.

[8] Colletaz G. and Hurlin C. (2005), Modèles non linéaires et prévision, *Rapport Institut CDC pour la recherche*, 106 pages.

[9] Dufour, J.-M., (2006). Monte Carlo tests with nuisance parameters: a general approach to finite sample inference and nonstandard asymptotics, *Journal of Econometrics*, 127, issue 2, 443-477.

[10] Hansen, L.P., (1982). Large sample properties of Generalized Method of Moments estimators, *Econometrica*, 50, 1029-1054.

[11] Harvey, D. I., and Leybourne, S. J., (2007), Testing for time series linearity, *Econometrics Journal,* 10, 149-165.

[12] Hyndman, R.J., (1995). Highest-density forecast regions for non-linear and non-normal time-series models, *Journal of Forecasting,* 14, 431-441.

[13] Politis, D.N., Romano, J.P., and Wolf, M., (1999), Subsampling, Springer-Verlag, New-York.

[14] Teräsvirta, T., (2006), Forecasting economic variables with non linear models, in Handbook of Economic Forecasting, G. Elliott, C.W.J. Granger and A. Timmermann editors, Elsevier, volume 1, Chapter 8, 413-457.

[15] Wallis, K.F., (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting*, 19, 165-175.

Appendix A: J statistics

Let us denote by $P^{(N,\alpha)} = \left(P_1^{(N,\alpha)}, .., P_m^{(N,\alpha)}\right)$ a $(m,1)$ vector whose components are the orthonormal polynomials $P_j^{(N,\beta)}(y_h)$ associated with the Binomial distribution $B(N,\alpha)$. Under the CC assumptions, the J statistic is simply defined by

$$J_{CC}(m) = \left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\alpha)}(y_h)\right)' \Sigma^{-1} \left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\alpha)}(y_h)\right), \tag{30}$$

where Σ denotes the long-run variance-covariance matrix of $P^{(N,\alpha)}(y_h)$. Since Σ is by definition equal to the identity matrix, we have

$$J_{CC}(m) = \frac{1}{H}\sum_{j=1}^{m}\left(\sum_{h=1}^{H} P_j^{(N,\alpha)}(y_h)\right)^2. \tag{31}$$

Similarly, the independence hypothesis statistic, denoted $J_{IND}(m)$ takes the form of:

$$\begin{aligned} J_{IND}(m) &= \left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\beta)}(y_h)\right)' \left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\beta)}(y_h)\right) \\ &= \frac{1}{H}\sum_{j=1}^{m}\left(\sum_{h=1}^{H} P_j^{(N,\beta)}(y_h)\right)^2, \end{aligned} \tag{32}$$

where $P^{(N,\beta)}(y_h)$ is the $(m,1)$ vector of orthonormal polynomials $P_j^{(N,\beta)}(y_h)$ defined for a coverage rate β that can be different from α.

Appendix B: Dufour (2006) Monte-Carlo Corrected Method

To implement MC tests, first generate M independent realizations of the test statistic, say S_i, $i = 1, \ldots, M$, under the null hypothesis. Denote by S_0 the value of the test statistic obtained for the original sample. As shown by Dufour (2006) in a general case, the MC critical region is obtained as $\hat{p}_M(S_0) \leq \eta$ with $1 - \eta$ the confidence level and $\hat{p}_M(S_0)$ defined as

$$\hat{p}_M(S_0) = \frac{M\ \hat{G}_M(S_0) + 1}{M + 1}, \tag{33}$$

where

$$\widehat{G}_M(S_0) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(S_i \geq S_0), \tag{34}$$

when the ties have zero probability, *i.e.* $\Pr(S_i = S_j) \neq 0$, and otherwise,

$$\widehat{G}_M(S_0) = 1 - \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left(S_i \leq S_0\right) + \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left(S_i = S_0\right) \times \mathbb{I}\left(U_i \geq U_0\right). \tag{35}$$

Variables U_0 and U_i are uniform draws from the interval $[0,1]$ and $\mathbb{I}(.)$ is the indicator function. As an example, for MC tests procedure applied to the test statistic $S_0 = J_{CC}(m)$, we just need to simulate under H_0, M independent realizations of the test statistic (*i.e.*, using durations constructed from independent Bernoulli hit sequences with parameter α) and then apply formulas (33 to 35) to make inference at the confidence level $1 - \eta$. Throughout the paper, we set M at $9,999$.

Table 1. Empirical size (block size $N = 100$, nominal size 5%)

Coverage rate 5%

T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.0316	0.0643	0.0499	0.0442	0.0587	0.0404
500	5	0.0521	0.0556	0.0615	0.0662	0.0544	0.0443
750	7	0.0409	0.0513	0.0595	0.0734	0.0503	0.0462
1000	10	0.0487	0.0535	0.0614	0.0655	0.0503	0.0565
1250	12	0.0522	0.0490	0.0543	0.0577	0.0417	0.0781
1500	15	0.0489	0.0479	0.0596	0.0577	0.0489	0.0656

Coverage rate 1%

T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.0516	0.0397	0.0397	0.0397	0.0132	0.0112
500	5	0.0314	0.0397	0.0360	0.0383	0.0640	0.0113
750	7	0.0330	0.0543	0.0456	0.0425	0.0384	0.0220
1000	10	0.0361	0.0482	0.0548	0.0473	0.0572	0.0251
1250	12	0.0575	0.0517	0.0592	0.0575	0.0627	0.0286
1500	15	0.0487	0.0518	0.0489	0.0414	0.0541	0.0312

Note: Under the null hypothesis, the violations are i.i.d. and follows a Bernoulli distribution. The results are based on 10,000 replications. For each sample, we provide the percentage of rejection at a 5% level. J_{CC}(m) denotes the GMM based conditional coverage test with m moment conditions. J_{UC} denotes the unconditional coverage test obtained for m=1. LR_{CC} (resp. LRuc) denotes the Christoffersen's conditional (resp. unconditional) coverage test. T denotes the sample size of the sequence of interval forecats violations I$_t$, while H=[T/N] denotes the number of block (size N=100) used to define the sums (y$_h$) of violations.

Table 2. Empirical size (block size $N = 25$, nominal size 5%)

Coverage rate 5%

T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.0386	0.0481	0.0417	0.0345	0.0558	0.0417
500	20	0.0547	0.0546	0.0550	0.0469	0.0573	0.0425
750	30	0.0461	0.0520	0.0583	0.0533	0.0572	0.0496
1000	40	0.0545	0.0567	0.0607	0.0510	0.0573	0.0592
1250	50	0.0489	0.0472	0.0555	0.0476	0.0423	0.0745
1500	60	0.0503	0.0515	0.0546	0.0472	0.0532	0.0685

Coverage rate 1%

T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.0456	0.0551	0.0551	0.0462	0.0157	0.0128
500	20	0.0309	0.0673	0.0632	0.0537	0.0651	0.0114
750	30	0.0592	0.0588	0.0645	0.0624	0.0390	0.0196
1000	40	0.0345	0.0498	0.0508	0.0849	0.0534	0.0231
1250	50	0.0423	0.0546	0.0448	0.0438	0.0582	0.0244
1500	60	0.0461	0.0540	0.0449	0.0289	0.0513	0.0286

Note: Under the null hypothesis, the violations are i.i.d. and follows a Bernoulli distribution. The results are based on 10,000 replications. For each sample, we provide the percentage of rejection at a 5% level. J_{CC}(m) denotes the GMM based conditional coverage test with m moment conditions. J_{UC} denotes the unconditional coverage test obtained for m=1. LR_{CC} (resp. LRuc) denotes the Christoffersen's conditional (resp. unconditional) coverage test. T denotes the sample size of the sequence of interval forecats violations I$_t$, while H=[T/N] denotes the number of block (size N=25) used to define the sums (y$_h$) of violations.

Table 3. Feasibility ratios (coverage rate $\alpha = 1\%$)

Size simulations				
	$T = 250$	$T = 500$	$T = 750$	$T = 1000$
LR_{UC}	0.9185	0.9939	0.9991	0.9999
LR_{CC}	0.9179	0.9936	0.9991	0.9999
Power simulations				
	$T = 250$	$T = 500$	$T = 750$	$T = 1000$
LR_{UC}	0.9023	0.9966	1.0000	1.0000
LR_{CC}	0.9010	0.9966	1.0000	1.0000

Note: the fraction of samples for which a test is feasible is reported for each sample size, both for the size and power tests for a coverage rate equal to 1%. LR_{UC} and LR_{CC} are Christoffersen (1998)'s unconditional and conditional coverage LR tests. Note that for J_{CC} the feasibility ratios are independent of the number of moment conditions m and are equal to 1. All results are based on 10,000 simulations. Note also that at 5% the LR tests can always be computed.

Table 4. Empirical Power (block size $N = 100$)

Coverage rate 5%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.2776	0.3991	0.4274	0.4203	0.2268	0.3333
500	5	0.1586	0.6151	0.6379	0.6221	0.1464	0.3298
750	7	0.1457	0.7197	0.7280	0.7099	0.1209	0.3632
1000	10	0.1302	0.8164	0.8209	0.8116	0.1152	0.4212
1250	12	0.1266	0.8703	0.8774	0.8639	0.1179	0.4874
1500	15	0.1367	0.9122	0.9118	0.9079	0.1322	0.5207
Coverage rate 1%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.1828	0.2709	0.2709	0.2820	0.1662	0.2730
500	5	0.2348	0.4525	0.4601	0.4403	0.1498	0.2361
750	7	0.2604	0.5410	0.5458	0.5516	0.2175	0.3073
1000	10	0.2980	0.6495	0.6596	0.6518	0.2116	0.3786
1250	12	0.3422	0.7051	0.6999	0.7058	0.2771	0.4407
1500	15	0.3663	0.7795	0.7738	0.7686	0.3330	0.4899

Note: Power simulation results are provided for different sample sizes T and number of blocks H, both at a 5% and 1% coverage rate. $J_{CC}(m)$ denotes the conditional coverage test with m moment conditions, J_{UC} represents the unconditional coverage test for the particular case when $m = 1$, and LR_{UC} and LR_{CC} are the unconditional and respectively conditional coverage tests of Christoffersen (1998). The results are obtained after 10,000 simulations by using Dufour (2005)'s Monte-Carlo procedure with ns=9999. The rejection frequencies are based on a 5% nominal size.

Table 5. Empirical Power (block size $N = 25$)

				Coverage rate 5%			
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.2656	0.5229	0.5314	0.4864	0.2285	0.3355
500	20	0.1842	0.7116	0.7022	0.6815	0.1482	0.3334
750	30	0.1509	0.8333	0.8277	0.8098	0.1155	0.3605
1000	40	0.1441	0.9091	0.9073	0.8919	0.1154	0.4374
1250	50	0.1444	0.9492	0.9439	0.9358	0.1218	0.4881
1500	60	0.1529	0.9717	0.9674	0.9637	0.1287	0.4981
				Coverage rate 1%			
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.2447	0.3697	0.3825	0.3866	0.1835	0.3355
500	20	0.2423	0.5163	0.5368	0.5410	0.1455	0.3334
750	30	0.2721	0.6436	0.6569	0.6232	0.2112	0.3605
1000	40	0.3253	0.7176	0.7428	0.7226	0.2044	0.4374
1250	50	0.3753	0.7926	0.7911	0.7896	0.2741	0.4881
1500	60	0.4373	0.8499	0.8456	0.8352	0.3368	0.4981

Note : Power simulation results are provided for different sample sizes T and number of blocks H, both at a 5% and 1% coverage rate. $J_{CC}(m)$ denotes the conditional coverage test with m moment conditions, J_{UC} represents the unconditional coverage test for the particular case when $m = 1$, and LR_{UC} and LR_{CC} are the unconditional and respectively conditional coverage tests of Christoffersen (1998). The results are obtained after 10,000 simulations by using Dufour (2005)'s Monte-Carlo procedure with ns=9999. The rejection frequencies are based on a 5% nominal size.

Table 6. Interval Forecast Evaluation, (SP500)

	Coverage rate 5%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	J_{IND} (2)	J_{CC} (2)	LR_{UC}	LR_{IND}	LR_{CC}
1	2.5263 (0.1120)	11.612 (0.0006)	29.493 (<0.0001)	2.4217 (0.1197)	3.3138 (0.0687)	5.8816 (0.0528)
5	4.4912 (0.0341)	10.615 (0.0011)	37.604 (<0.0001)	4.0607 (0.0439)	7.5661 (0.0059)	11.787 (0.0028)
10	2.5263 (0.1120)	19.605 (<0.0001)	46.040 (<0.0001)	2.4217 (0.1197)	3.3138 (0.0687)	5.8816 (0.0528)
	Coverage rate 1%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	J_{IND} (2)	J_{CC} (2)	LR_{UC}	LR_{IND}	LR_{CC}
1	109.09 (<0.0001)	11.612 (0.0006)	2072.4 (<0.0001)	49.234 (<0.0001)	3.3138 (0.0687)	52.693 (<0.0001)
5	134.68 (<0.0001)	10.615 (0.0011)	2658.3 (<0.0001)	57.475 (<0.0001)	7.5661 (0.0059)	65.201 (<0.0001)
10	109.09 (<0.0001)	19.605 (<0.0001)	2714.6 (<0.0001)	49.234 (<0.0001)	3.3138 (0.0687)	52.693 (<0.0001)

Note: 250 out of sample forecasts of the SP500 index (from 20/12/1956 to 19/12/1957) are computed for three different horizons (2, 5 and 10) both at a 5% and 1% coverage rate. The evaluation results of the corresponding interval forecasts are reported both for our GMM-based tests and Christoffersen (1998)'s LR tests. For this objective, a block size N=25 was used. For all tests, the numbers in the parentheses denote the corresponding p-values.

Table 7. Interval Forecast Evaluation, (Nikkei)

			Coverage rate 5%			
	GMM-based tests			LR tests		
Horizon	J_{UC}	J_{IND} (2)	J_{CC} (2)	LR_{UC}	LR_{IND}	LR_{CC}
1	2.5263 (0.1120)	3.9132 (0.0479)	12.060 (0.0024)	1.7470 (0.1863)	0.2521 (0.6156)	2.1382 (0.3433)
5	1.7544 (0.1853)	3.8728 (0.0491)	9.6337 (0.0081)	1.1744 (0.2785)	0.4005 (0.5268)	1.7072 (0.4259)
10	1.7544 (0.1853)	3.8728 (0.0491)	9.6337 (0.0081)	1.1744 (0.2785)	0.4005 (0.5268)	1.7072 (0.4259)
			Coverage rate 1%			
	GMM-based tests			LR tests		
Horizon	J_{UC}	J_{IND} (2)	J_{CC} (2)	LR_{UC}	LR_{IND}	LR_{CC}
1	109.09 (0.0000)	3.9132 (0.0479)	1279.3 (0.0000)	45.258 (<0.0001)	0.2521 (0.6100)	45.649 (<0.0001)
5	97.306 (0.0000)	3.8728 (0.0491)	1073.6 (0.0000)	41.384 (<0.0001)	0.4005 (0.5268)	41.916 (<0.0001)
10	97.306 (0.0000)	3.8728 (0.0491)	1073.6 (0.0000)	41.384 (<0.0001)	0.4005 (0.5268)	41.916 (<0.0001)

Note: 250 out of sample forecasts of the Nikkei index (from 27 January 1987 to 21 February 1992) are computed for three different horizons (2, 5 and 10) both at a 5% and 1% coverage rate. The evaluation results of the corresponding interval forecasts are reported both for our GMM-based tests and Christoffersen (1998)'s LR tests. For this objective, a block size N=25 was used. For all tests, the numbers in the parentheses denote the corresponding p-values.

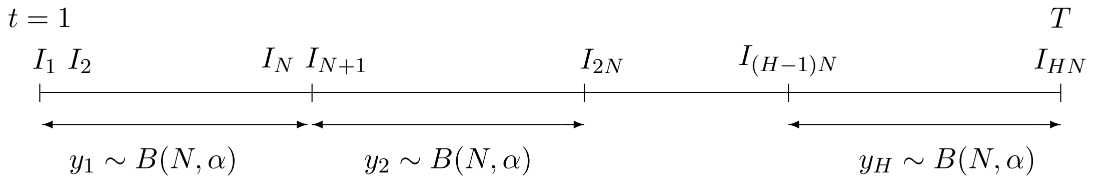

FIGURE 1 – Partial sums y_h and block size N

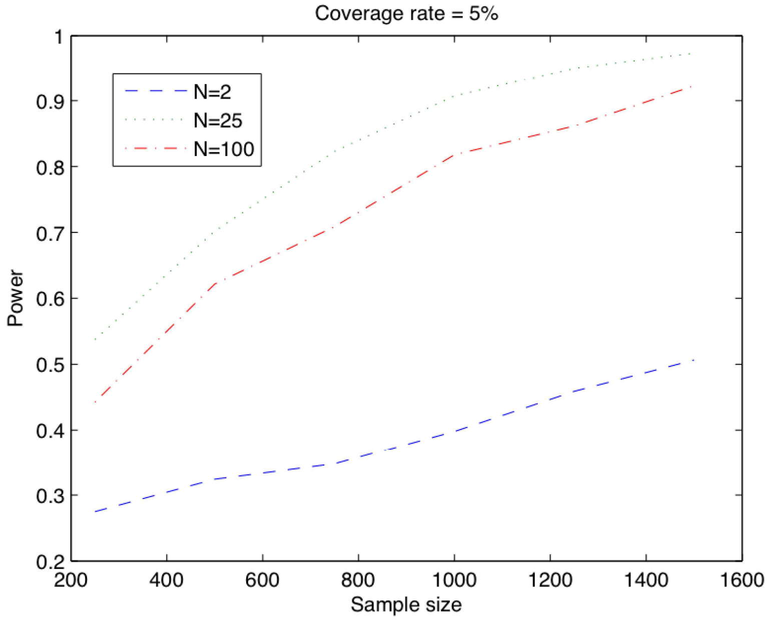

FIGURE 2 – Corrected power of the $J_{CC}\,(2)$ test statistic as function of the sample size T(coverage rate $\alpha = 5\%$)

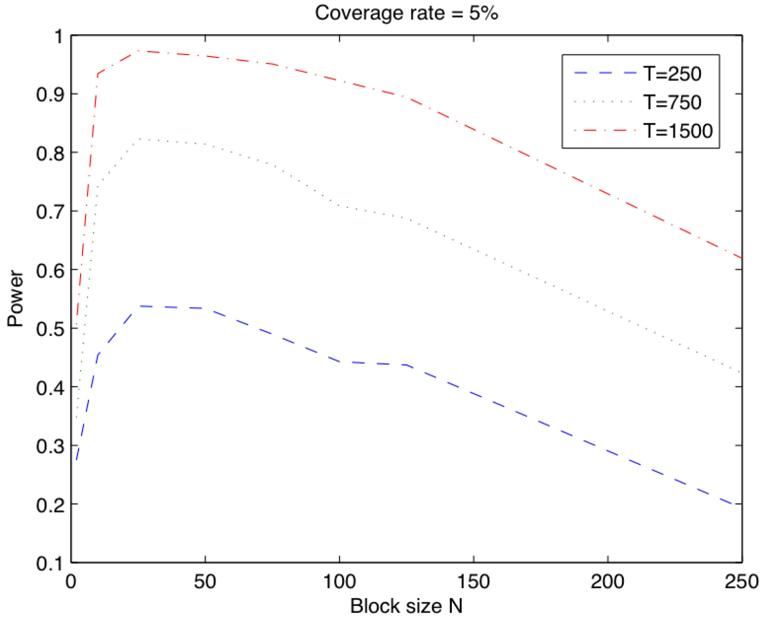

FIGURE 3 – Corrected power of the $J_{CC}\,(2)$ test statistic as function of the block size N (coverage rate $\alpha = 5\%$)