Article type: Overview

VOLATILITY FORECASTS EVALUATION AND COMPARISON

Sébastien Laurent Francesco Violante

Maastricht University, The Netherlands and Université catholique de Louvain, CORE, Belgium.

Abstract

This article surveys the most important developments in volatility forecast comparison and model selection. We review a number of evaluation methods and testing procedures for predictive accuracy based on statistical loss functions. We also review recent contributions on the admissible form of loss functions ensuring consistency of the ordering when forecast performances are evaluated with respect to an imperfect volatility proxy. The techniques discussed are illustrated using artificial and EUR/USD exchange rate data.

Keywords

Volatility, GARCH, Loss function, Consistent ranking

JEL Classification: C10, C32, C51, C52, C53, G10.

Traditional regression tools have shown their limitation in the modelling of financial time-series (say y_t). Assuming that only the conditional mean could be changing with covariates while the variance remains constant over time often revealed to be an unrealistic assumption in practice. Indeed, it is now widely accepted that high frequency financial returns are heteroskedastic. As an example, Figure 1 plots the daily returns in % of the EUR/USD exchange rate on the period January 1999 to - April 2011. This figure clearly suggests that the variance of this series is indeed not constant over time and clusters of volatility can be visually detected. Since the seminal

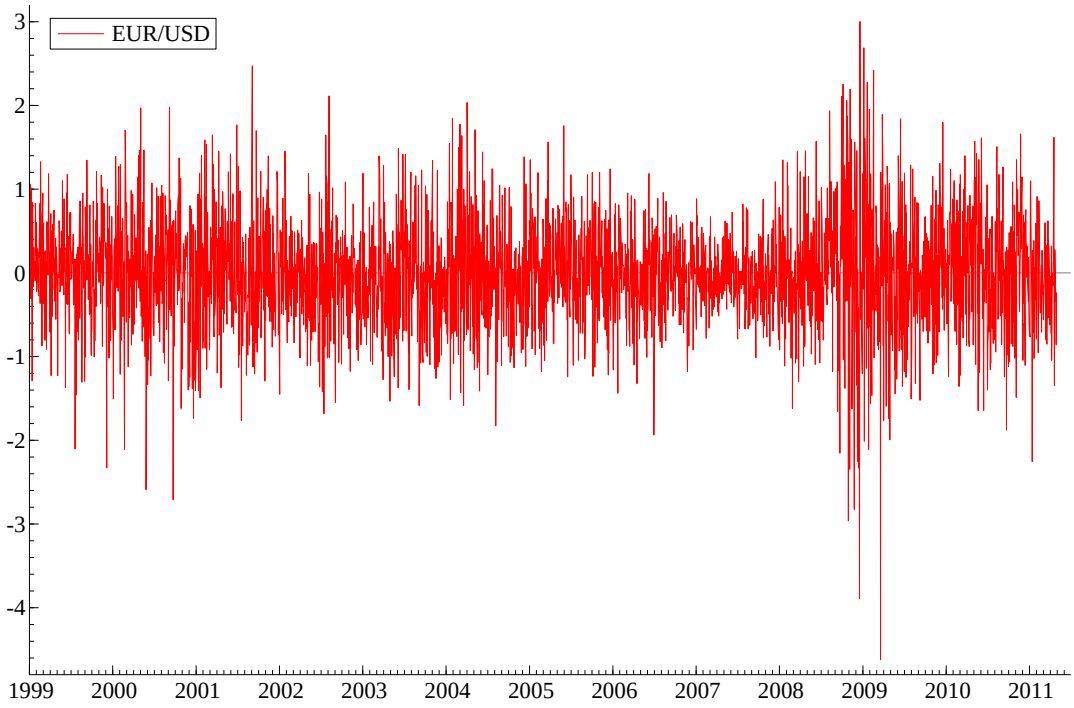


Figure 1: EUR/USD exchange rate returns in % on the period January 1999 to - April 2011.

paper of Engle (1982), autoregressive moving average (ARMA) models have been extended to essentially equivalent models for the variance. Autoregressive Conditional Heteroscedasticity (ARCH) models have been extensively used in the literature. A time series y_t $(t = 1, \ldots, T)$ is said to follow an ARCH-type model when it can be described as follows:

$$y_t = m_t(\eta) + \varepsilon_t \quad (1)$$

$$\varepsilon_t = \sigma_t(\eta) z_t \quad (2)$$

$$m_t(\eta) = c(\eta|\Omega_{t-1}) \quad (3)$$

$$\sigma_t(\eta) = h(\eta|\Omega_{t-1}), \quad (4)$$

where $c(.|\Omega_{t-1})$ and $h(.|\Omega_{t-1})$ are deterministic functions of Ω_{t-1} (the information set at time $t-1$), depending on an unknown vector of parameters η, and z_t is an $(i.i.d.)$ process with $E(z_t) = 0$ and $Var(z_t) = 1$. For modeling the conditional mean $m_t(\eta)$, one usually relies on Autoregressive (AR) and/or Moving Average (MA) specifications. See Brockwell (2011) for an overview of ARMA models. Many parametric specifications have also been proposed for $\sigma_t^2(\eta)$. An extensive review is given in Bollerslev (2010).

In this article we essentially focus on the conditional variance $\sigma_t^2(\eta)$ and, more precisely, we review recent developments on volatility forecasts evaluation and comparison. Once point forecasts are computed from one or more volatility models, models' performances can be measured by contrasting forecasts to realisations by means of a statistical loss function. Then performances can be ordered according to the selected criterion and a ranking of models established. Finally, inference on predictive accuracy based on such ranking can be carried out using a variety of approaches. In this article, we discuss several statistical methods for single, pairwise and multiple forecast evaluation. A critical problem characterising the comparison of volatility forecasts is the fact that the target variable is latent. Typically, this problem is solved by using a conditionally unbiased (and possibly consistent) ex-post estimator, often referred to as a volatility proxy. Some of the most popular proxies used in the literature are mentioned in Section . However, it has been shown in Hansen and Lunde (2006), Patton (2009), Patton and Sheppard (2009) and Laurent, Rombouts, and Violante (2009) that the substitution of the true volatility by a proxy, by definition imperfect, may introduce serious distortions in the ordering of volatility forecasts. To overcome the problem, these authors provide conditions that the loss function has to satisfy in order to ensure a ranking asymptotically robust to the noise in the proxy and propose a number of robust functional forms. The rest of the article is organised as follows. In the next section, we discuss statistical methods for single, pairwise and multiple forecast evaluation. Then, we discuss the problem of forecast evaluation under imperfect volatility proxies and provide an illustration based on artificial and exchange rate data. The last section concludes.

Inference on volatility forecasts

GARCH model

The most popular ARCH-type model is certainly the Generalized ARCH model of Bollerslev (1986). The GARCH (p, q) model specifies the square of σ_t in Equation (4) as follows:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2. \tag{5}$$

Estimation of ARCH-type models is commonly done by maximum likelihood so that one has to make an additional assumption about the innovation process z_t. Weiss (1986) and Bollerslev and Wooldridge (1992) show that the Gaussian quasi-maximum likelihood (QML) estimator is consistent if the conditional mean and the conditional variance are correctly specified.

After the estimation of the parameters of the model, h-step-ahead forecasts of the conditional variance, i.e. $\sigma^2_{t+h|t}$ are obtained as follows:

$$\sigma^2_{t+h|t} = \hat{\alpha}_0 + \sum_{i=1}^{q} \hat{\alpha}_i \varepsilon^2_{t+h-i|t} + \sum_{j=1}^{p} \hat{\beta}_j \sigma^2_{t+h-j|t}, \tag{6}$$

where $\varepsilon^2_{t+i|t} = \sigma^2_{t+i|t}$ for $i > 0$ while $\varepsilon^2_{t+i|t} = \varepsilon^2_{t+i}$ and $\sigma^2_{t+i|t} = \sigma^2_{t+i}$ for $i \leq 0$. Equation (6) is usually computed recursively, even if a closed form solution of $\sigma^2_{t+h|t}$ can be obtained by recursive substitution in Equation (6). Similarly, one can easily obtain the h-step-ahead forecast of the conditional variance of more complicated ARCH-type models.

Single forecast evaluation

A simple method for evaluating the accuracy of the volatility forecasts of an ARCH-type model, say model k, is the regression based evaluation proposed by Mincer and Zarnowitz (1969) (hereafter MZ). This approach requires the estimation of the coefficients of a regression of the target on a constant and the forecast under evaluation (denoted $\sigma^2_{t,k}$),

$$\sigma^2_t = a + b\sigma^2_{t,k} + u_t. \tag{7}$$

The MZ regression allows to evaluate two different aspects of the volatility forecast. First, by testing the joint hypothesis $H^0 : a = 0 \cup b = 1$, it allows to test the presence of systematic over- or under-predictions, i.e., whether the forecast is biased. Second, being the R^2 of (7) an indicator of the correlation between the realisation and the forecast, it can be used as evaluation criterion of the accuracy of the forecast.

Pairwise comparison

The first approach to pairwise comparison that we consider is the test of equal predictive ability proposed by Diebold and Mariano (1995) and further refined by West (1996), McCracken (2000), Clark and McCracken (2001), Corradi, Swanson, and Olivetti (2001), Clark and West (2006), Clark and West (2007), McCracken (2007) and Clark and McCracken (2005) among others (hereafter DMW). The DMW test is a very general procedure[1] designed to compare two rival forecasts in terms of their forecasting accuracy using a general

[1]The test does not require zero-mean forecast errors (hence the forecasts can be biased), specific distributional assumptions nor zero-serial correlation for the forecast errors.

loss function, $L(.) : R_{++} \times \mathcal{H} \rightarrow R^{+}, \mathcal{H} \subset R_{++}$. The loss function, i.e. the measure of predictive accuracy, can be specified according to the definition of optimality adopted by the forecaster.

Define the loss differential between model k and j as

$$d_t = L(\sigma_t^2, \sigma_{t,k}^2) - L(\sigma_t^2, \sigma_{t,j}^2) \tag{8}$$

or using a more compact notation, $d_t = L_{t,k} - L_{t,j}$. Under stationarity of d_t, $\mathrm{E}[d_t]$ is well defined and the null hypothesis of equal predictive ability takes the form $H^0 : \mathrm{E}[d_t] = 0$. The test statistic is

$$DM - T = \frac{\sqrt{T}\bar{d}}{\sqrt{\omega}} \overset{a}{\sim} \mathrm{N}(0,1), \tag{9}$$

where $\bar{d} = T^{-1}\sum_t d_t$ and $\omega = \lim_{t\rightarrow\infty} \mathrm{Var}(\sqrt{T}\bar{d})$ is its asymptotic variance. A natural estimator of ω is the sample variance of d_t, though this estimator is consistent only if the loss differentials are serially uncorrelated. Since this is not generally the case, a suitable HAC estimator, such as the Newey-West variance estimator, is preferable.

It is worth noting that the aim of these tests is to infer about $\mathrm{E}[d_t(\theta_0)]$ using $T^{-1}\sum_t d_t(\theta_0)$, where θ_0 represents the models parameters population values, and thus require asymptotics based on the size of the estimation sample $\mathcal{T}$ and the size of the forecast evaluation sample T to grow to infinity at the same rate.[2] Since this type of asymptotics relies on parameter population values, the comparison of nested models is obviously not allowed, because the asymptotic distribution of the statistic under the null turns out to be degenerate (identically zero) when the restricted model is true. A solution to this problem has been provided by McCracken (2007) and Clark and McCracken (2005) (CM), which argue that, although $T^{-1}\sum_t d_t(\hat{\theta}) - \mathrm{E}[d_t(\theta_0)] \underset{p}{\rightarrow} 0$ when models are nested, $T^{-1}\sum_t d_t(\hat{\theta})$ is a non-degenerate random variable. Based on this argument, they suggest a variety of statistics, suited for testing equal predictive accuracy, which depart from the standard Gaussian asymptotics of DWM and whose distribution depends entirely on the parameters uncertainty. To obtain the null distribution Clark and McCracken (2009) develop an asymptotically valid procedure based on bootstrap sampling.

Giacomini and White (2006) develop a test of finite-sample predictive ability. They construct a test for conditional equal predictive accuracy based on asymptotics in which the estimation error is a permanent component of the forecast error. Rather than focussing on unconditional expectations, their approach aims at inferring about conditional expectations of forecast errors, i.e. inferring about $\mathrm{E}[d_t(\hat{\theta})]$ using $T^{-1}\sum_t d_t(\hat{\theta})$. The null hypothesis of equal predictive ability can be expressed as

$$\mathrm{E}[L(\sigma_t^2, \sigma_{t,k,\tau_k}^2(\hat{\theta}_{k,t,\tau_k})) - L(\sigma_t^2, \sigma_{t,j,\tau_j}^2(\hat{\theta}_{j,t,\tau_j}))] \equiv \mathrm{E}[d_{\mathcal{T},t}(\hat{\theta})] = 0, \tag{10}$$

[2] Such asymptotics apply naturally under a recursive forecasts scheme, where the sample used to estimate the parameters of the model grows at the same rate as the forecast sample, i.e. at each step t the forecast is based on all available information up to $t-1$. Additional assumptions for asymptotics based on rolling and fixed schemes, where the estimation sample increases with the overall sample size, are given in West (1996).

where, for $i = k, j$, τ_i is size of the estimation window, possibly different for each model (explaining the third index in σ^2_{t,k,τ_k} and $\hat{\theta}_{k,t,\tau_k}$) and $\mathcal{T} = \max(\tau_k, \tau_j)$. Given that, under the null hypothesis, $\{d_{\mathcal{T},t}, \Im_t\}$ is a martingale difference sequence, (10) is equivalent to $\mathrm{E}[\delta_{t-1} d_{\mathcal{T},t}] = 0$, where δ_{t-1}, referred to as the test function, is a $\Im_{t-1}$-measurable vector of dimension q. Contrary to CM, in this case, standard asymptotic normality arguments hold. The GW test takes the form of a Wald-type statistic

$$GW - T^{\delta}_{\mathcal{T}} \quad = \quad T \left(T^{-1} \sum_{t=1}^{T} \delta_{t-1} d_{\mathcal{T},t} \right)' \hat{\Omega}^{-1} \left(T^{-1} \sum_{t=1}^{T} \delta_{t-1} d_{\mathcal{T},t} \right), \tag{11}$$

where $\hat{\Omega}$ is a consistent estimator of the variance of $\delta_{t-1} d_{\mathcal{T},t}$. The statistic is asymptotically χ^2_q under the null hypothesis.

An example of test function suggested by Giacomini and White (2006) is $\delta_t = (1, d_{\mathcal{T},t})'$ which allows to test jointly for equal predictive ability and lack of serial correlation in the loss differentials.

Clearly, the GW asymptotics hold when the size of the estimation sample is fixed as the forecasts sample grows, i.e., $\mathcal{T}$ fixed, $T \to \infty$, but also under a rolling scheme[3] and in general to any limited memory estimator.

Multiple comparison

When multiple alternative forecasts are available, it may be of interest to test whether a specific model, selected independently from the data, produces systematically superior performances with respect to the other models. The difference with the approaches discussed in the previous section is twofold: first, the multiple comparison allows to recognize the multiplicity effect by testing multiple hypotheses, and second, the choice of a benchmark requires a test of superior predictive ability which requires testing composite hypotheses, i.e. (weak) inequalities. Consequently, the asymptotic distribution of these tests is typically non-standard.

The first approach that we consider is the reality check for data snooping of White (2000) (hereafter RC). Let us define the loss differential between the benchmark, $\sigma^2_{t,0}$, and some rival forecast, $\sigma^2_{t,k}$ $k = 1, ..., m$ as

$$d_{t,k} = L(\sigma^2_t, \sigma^2_{t,0}) - L(\sigma_t, \sigma^2_{t,k}) \tag{12}$$

and $\mathbf{d}_t = (d_{1,t}, ..., d_{m,t})$. Provided that $\mathbf{d}_t$ is (strictly) stationary, $\mathrm{E}[\mathbf{d}_t]$ is well defined and the null hypothesis of interest takes the form

$$H^0 : \max_k \mathrm{E}[d_{k,t}] \leq 0 \tag{13}$$

i.e., the benchmark is superior to the best alternative. Clearly, the null hypothesis in (13) is a multiple hypothesis, i.e., the intersection of the one-sided individual hypotheses $\mathrm{E}[d_{t,k}] \leq 0$. The RC test statistic

[3] The sequence of T parameters is generated using the most recent information, e.g. a rolling sample of fixed size $\mathcal{T}$.

takes the form

$$RC - T = \max_k(\sqrt{T}\,\bar{d}_k), \tag{14}$$

where $\bar{d}_k = T^{-1}\sum_{t=1}^{T} d_{t,k}$.

Given strict stationary of $\mathbf{d}_t$, White (2000) invokes conditions provided in West (1996) that lead to $\sqrt{T}(\bar{\mathbf{d}} - \mathrm{E}[\mathbf{d}_t]) \xrightarrow{d} \mathrm{N}(0, \Omega)$. However, (14) has an asymptotic distribution under the null which is unknown and that depends on the nuisance parameters $\mathrm{E}[\mathbf{d}_t]$ and Ω.[4]

White (2000) suggests two procedures to obtain the distribution under the null, namely the 'Monte Carlo Reality Check' (simulated inference) and the 'Bootstrap Reality Check' (bootstrap inference), see White (2000) for further details.

Note that, as in Diebold and Mariano (1995), White's (2000) asymptotics are based on population values. Using similar arguments as Giacomini and White (2006), Hansen (2005) generalize the procedure to the comparison of nested models. Using a similar approach, Hansen (2005) proposes a new test for superior predictive ability (henceforth SPA). The SPA statistic takes the form

$$SPA - T = \max\left[\max_k \frac{\sqrt{T}\bar{d}_k}{\sqrt{\hat{\omega}_k}}, 0\right], \tag{15}$$

where $\hat{\omega}_k$ is some consistent estimator of $\omega_k = \lim_{t\to\infty} \mathrm{Var}(\sqrt{T}\bar{d}_k)$. The null distribution of the SPA statistic is based on $\sqrt{T}\bar{\mathbf{d}} \xrightarrow{d} \mathrm{N}(\hat{\boldsymbol{\mu}}^c, \hat{\Omega})$, where $\hat{\boldsymbol{\mu}}^c$ is a consistent estimator of $\boldsymbol{\mu} = \mathrm{E}[\mathbf{d}_t]$ that conforms with the null hypothesis. Hansen (2005) also provides a detailed description of the bootstrap scheme used to obtain the distribution under the null hypothesis.

The SPA test differs from the RC in two ways. First, he proposes a different statistic based on studentized quantities to reduce the loss of power that the RC can suffer when poor and irrelevant forecasts are considered. Second, he employs a sample dependent distribution under the null. The latter is based on a procedure that incorporates additional sample information in order to identify the relevant alternatives. In fact, while the procedure based on the principle of the least favourable configuration to the alternative adopted by White (2000) implicitly relies on an asymptotic distribution under the null that assumes $\mathrm{E}[d_{t,k}] = 0$ for all k, Hansen (2005) points out that all negative values of $\mathrm{E}[d_{t,k}]$ should be considered because they also conform with the null. He provides lower and upper bounds to the distribution of (15) corresponding to a liberal test under the null hypothesis that the models with worse performance than the benchmark are poor models in the limit and a conservative one under the least favourable configuration to the alternative, respectively.

Clearly, in many applications the choice of a benchmark may not be obvious or an objective benchmark may not exist. In other applications a single model that is significantly superior to all the alternatives may

[4] $\mathrm{E}[\mathbf{d}_t]$ is estimated using the least favorable configuration for the alternative which in this case correspond to $\mathrm{E}[\mathbf{d}_t] = 0$, i.e., all alternatives are as good as the benchmark.

not emerge especially when the set of competing models is large and/or the data may not be sufficiently informative to give a univocal answer. In these cases, the forecaster may aim to reduce the set of competing models to a smaller set that is guaranteed to contain the best forecasting model at a given confidence level. This approach is known as multiple comparison without control.

Within this category we find the model confidence set (MCS) of Hansen, Lunde, and Nason (2009). The MCS is a sequential test of equal predictive ability. It differs from the RC and the SPA because it does not require a benchmark to be specified. It has the additional advantage of relying on simple hypotheses (equalities), allowing to derive standard asymptotics.

Given an initial set of forecasts, M^0, the starting hypothesis is that all models in M^0 have equal forecasting performances. The relative performance of each pair of forecasts is measured by $d_{t,k,j} = L(\sigma_t^2, \sigma_{t,k}^2) - L(\sigma_t^2, \sigma_{t,j}^2)$, for all $k, j \in M^0$ with $k \neq j$. Under the assumption that $d_{t,k,j}$ is stationary, the null hypothesis of equal predictive ability takes the form

$$H^0 : \mathrm{E}[d_{t,k,j}] = 0 \qquad \forall k, j \in M^0. \tag{16}$$

If the null of equal predictive ability is rejected at a given confidence level α, then an elimination rule is called to remove the worst performing model. The equal predictive ability test is then repeated until the non-rejection of the null, while keeping the confidence level α fixed at each iteration, thus allowing to construct a $(1-\alpha)$-confidence set, $M^* \equiv \{k \in M_0 : E(d_{t,k,j}) \leq 0 \; \forall \; j \in M^0\}$, for the best model(s) in M^0.

Let $\mathbf{L}_t$ be the $(m \times 1)$ vector of sample performances $L(\sigma_t^2, \sigma_{t,k}^2)$, $k \in M$ and $\iota_\perp$ the $(m \times (m-1))$ orthogonal complement of a m-dimensional vector of ones, where m is the dimension of M. Then, the vector $\iota_\perp' \mathbf{L}_t$ can be viewed as $m-1$ relevant contrasts as each element can be obtained as a linear combination of $d_{k,j,t},\ k, j \in M$ which has mean zero under the null (16). Hence, (16) is equivalent to $\mathrm{E}[\iota_\perp' \mathbf{L}_t] = 0$ and, under strict stationarity of $d_{k,j,t}$, it holds that $T^{-1/2} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t$ is asymptotically Gaussian with mean 0 and covariance matrix $\Omega = \lim_{t \to \infty} \mathrm{Var}\left(T^{-1/2} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t\right)$. Thus, it seems natural to employ traditional quadratic-form type of tests as

$$MCS - T_Q = T \left(T^{-1} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t \right)' \hat{\Omega}^+ \left(T^{-1} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t \right) \tag{17}$$

and

$$MCS - T_F = \frac{T-q}{q(T-1)} \, MCS - T_Q, \tag{18}$$

where $\hat{\Omega}$ is some consistent estimator of Ω, $q = \mathrm{rank}(\hat{\Omega})$ denotes the number of linearly independent contrasts and $\hat{\Omega}^+$ denotes the More-Penrose pseudo-inverse of $\hat{\Omega}$. The statistic in (17) is asymptotically χ_q^2, whereas (18) is asymptotically $\mathrm{F}_{q,T-q}$ under the null hypothesis, as the subscripts Q (quadratic) and F (F-distributed) suggest.

The main pitfall of these asymptotic tests is that, when m is large, it might be problematic to obtain a well conditioned estimate of Ω. Alternatively, Hansen, Lunde, and Nason (2009) propose three simpler statistics expressed as functions of studentized quantities. The first statistic is expressed as a sum of deviations from the common average (hence the subscript). Under the null hypothesis $H^0 = \mathrm{E}[\bar{d}_k] = 0 \ \forall k \in M$ the statistic takes the form[5]

$$MCS - T_D = \frac{1}{m} \sum_{k \in M} t_k^2, \tag{19}$$

where $t_k = \sqrt{T}\bar{d}_k / \sqrt{\hat{\omega}_k^D}$, $k = 1, ..., m$, and $\bar{d}_k = m^{-1} \Sigma_{j \in M} \bar{d}_{k,j}$ is the contrast of model i's sample loss with respect to the average across all models and $\bar{d}_{k,j} = T^{-1} \Sigma_{t=1}^T d_{k,j,t}$ is the sample loss differential between models k and j. The variances $\hat{\omega}_k^D$ are consistent estimators of $\omega_k^D = lim_{t \to \infty} \mathrm{Var}(\sqrt{T}\bar{d}_k)$. The remaining two statistics, dubbed range and semi-quadratic, take the form

$$MCS - T_R = \max_{k,j \in M} |t_{k,j}| \qquad \text{and} \qquad MCS - T_{SQ} = \frac{1}{m} \sum_{k,j \in M} t_{k,j}^2 \tag{20}$$

respectively, where $t_{k,j} = \sqrt{T}\bar{d}_{k,j} / \sqrt{\hat{\omega}_s^R}$, $k, j = 1, ..., m \ k \neq j$ and $s = 1, ..., m(m-1)$ and the variance $\hat{\omega}_s^R$ is a consistent estimator of $\omega_s^R = lim_{t \to \infty} \mathrm{Var}(\sqrt{T}\bar{d}_{k,j})$.

Note that the distribution of (19) and (20) is non-standard and depend on the nuisance parameters ω_k^D and ω_s^R, respectively. Hansen, Lunde, and Nason (2009) also provide details on the bootstrap scheme employed to solve the nuisance parameter problem and obtain the distribution under the null hypothesis.

If the null hypothesis is rejected, then Hansen, Lunde, and Nason (2009) suggest the use of the following elimination rule $\mathcal{E}_M = \arg \max_{k \in M} t_k$ which excludes the model with the largest standardised excess loss relative to the average across models. The iterative testing procedure ends when the first non rejection occurs, or obviously if all models but one have been recursively eliminated. Finally, the MCS p-value is equal to $p_i = \max(p_{i-1}, p(i))$, $i = 1, ..., m$, where p_i is the p-value of the test under the null hypothesis $H_{M^i}^0$, i.e., at the ith step of the iteration process. By convention the p-value when there is only one surviving model is $p_m = 1$.

Interestingly, the SPA and MCS tests are implemented in the free Ox software package MULCOM of Hansen and Lunde (2010).

Loss functions and the latent variable problem

A critical problem, which characterises the comparison of volatility forecasts, is the fact that the target variable is latent. Typically, this problem is solved by using a conditionally unbiased (and possibly consistent) ex-post estimator, often referred to as volatility proxy and denoted $\hat{\sigma}_t^2$. It is worth noting that the

[5] Note that the null hypothesis is equivalent to (16).

only property that we require for the volatility proxy is conditional unbiasedness, i.e., $\mathrm{E}_{t-1}[\hat{\sigma}_t^2] = \sigma_t^2$. if not otherwise stated, we assume that at least one conditionally unbiased proxy is available. In some specific cases we will also require the stronger assumption of consistency or the availability of a variety of proxies that can be ordered in terms of their level of accuracy. A simple variance proxy commonly used in the financial literature is the squared return, although such proxy is known to be extremely noisy. However, its scarce informative content makes it unsuited for the purpose of assessing the accuracy of volatility forecasts, in that an uninformative volatility proxy makes difficult to asses the statistical relevance of the forecast performances. Other volatility proxies based on realised moments are discussed in Barndorff-Nielsen and Shephard (2002), Zhang, Mykland, and Ait-Sahalia (2004), Zhou (1996), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) among the others. Range based variance estimators can be found in Parkinson (1980), Garman and Klass (1980) and Brandt and Diebold (2006).

As first noted by Andersen and Bollerslev (1998) and Andersen, Bollerslev, and Meddahi (2005), conditional unbiasedness alone does not suffice to ensure, asymptotically, the same outcome that would be obtained if the true volatility was observable. It has been shown in Hansen and Lunde (2006), Patton (2009), Patton and Sheppard (2009) and Laurent, Rombouts, and Violante (2009) that the substitution of the true volatility by a proxy, that by definition is imperfect, may introduce serious distortions in the ordering of volatility forecasts. More formally, given two model based forecasts, $\sigma_{t,k}^2$ and $\sigma_{t,j}^2$, it may be the case that a given loss function $L(.)$ is such that the true ordering between model k and model j implies $\mathrm{E}[L(\sigma_t^2, \sigma_{t,k}^2)] < \mathrm{E}[L(\sigma_t^2, \sigma_{t,j}^2)]$, while the ordering based on the proxy reveals $\mathrm{E}[L(\hat{\sigma}_t^2, \sigma_{t,k}^2)] \geq \mathrm{E}[L(\hat{\sigma}_t^2, \sigma_{t,j}^2)]$. Since the distortion in the ordering does not disappear asymptotically, when the evaluation is based on a target observed with error the choice of the evaluation criteria becomes critical in order to avoid a biased outcome. To overcome the problem, these authors define conditions that the loss function has to satisfy in order to ensure a ranking asymptotically robust to the noise in the proxy and propose a number of robust functional forms.

Given the latent nature of the variable of interest and since the type of evaluation and inference on forecasts accuracy that we have in mind relies, more or less explicitly, on the ordering implied by a predefined loss function, e.g., squared, absolute, relative forecast error or yet correlation between forecasts and realisations, if the ranking is non-robust to the noise in the proxy (i.e. is subject to potential distortions) the inference on models' predictive accuracy will be incorrect even if the testing procedure is formally valid. If instead the loss function ensures robustness of the ranking, the variability of the volatility proxy is only likely to reduce the power of the test but not its asymptotic size. See Laurent, Rombouts, and Violante (2009) for an illustration.

We first consider the evaluation based on the MZ approach. Obviously the latent nature of the target variable makes the regression in (7) unfeasible. Substituting the true variance by some conditionally unbiased proxy,

$\hat{\sigma}_t^2 = \sigma_t^2 + \lambda_t$ with $\mathrm{E}_{t-1}[\lambda_t] = 0$ and $\mathrm{Var}_{t-1}[\lambda_t] \neq 0$ and finite, we can rewrite (7) as

$$\hat{\sigma}_t^2 = a + b\sigma_{t,k}^2 + e_t, \tag{21}$$

where $e_t = \lambda_t + u_t$. Since $\hat{\sigma}_t^2$ is a conditionally unbiased estimator of the true variance then (21) yields unbiased estimates of a and b.

As mentioned, the R^2 of the MZ regression has been used as a criterion for ordering over a set of volatility forecasts, see Andersen and Bollerslev (1998) and Andersen, Bollerslev, Diebold, and Labys (2003) for examples. Hansen and Lunde (2006) show that, due to the latent variable problem, this criterion is not always adequate to the scope and may lead to a perverse outcome. They derive sufficient conditions under which the ordering of volatility forecasts is unaffected when the true variance is substituted by a proxy. They establish that the R^2 is a valid criterion if $\mathrm{E}_{t-1}[\sigma_t^2 - \hat{\sigma}_t^2](\partial^i \phi(\sigma_t^2)/\partial(\sigma_t^2)^i) = c_i$ for some constant c_i, $\forall t = \mathcal{T}+1, \ldots, \mathcal{T}+T$ and $i \in \mathbb{N}$ and where $\phi(.)$ represents the transformation of the dependent variable and the regressor, e.g., log, square, square root, etc. This condition validates the use of the MZ regression in level but also, for example, of the quadratic transformation, i.e., $\phi(x) = x^2$, although in the latter case, as pointed out by Andersen, Bollerslev, and Meddahi (2005), the quadratic transformation of an unbiased forecasts will not generally result to be unbiased for $(\hat{\sigma}_t^2)^2$, but rejects, for example, the log-regression. Analytical examples under different distributional assumptions for the volatility proxy can be found in Patton and Sheppard (2009).

Given (21), it is also interesting to elaborate on the role played by the level of accuracy of the proxy. Clearly, the variance of the innovations in (21) depends on the accuracy of the volatility proxy. Thus, if a high quality proxy is available, the regression parameters are estimated more accurately. Similarly, as the quality of the proxy deteriorates, the R^2 of the regression in (21), $\mathrm{Cov}(\hat{\sigma}_t^2, \sigma_{t,k}^2)^2/(\mathrm{Var}(\hat{\sigma}_t^2)\mathrm{Var}(\sigma_{t,k}^2))$, results penalised. See Andersen and Bollerslev (1998) for an analytical example.

When the ordering is based on a statistical loss function, a sufficient condition to ensure consistency of the ordering is that $\partial^2 L(\sigma_t^2, \sigma_{t,k}^2)/(\partial\sigma_t^2)^2$ exists and does not depend on $\sigma_{t,k}^2$. It follows immediately that many evaluation criteria commonly used in applied works, e.g., forecast errors of square roots and log transformations or proportional error loss functions, are rejected whereas the squared forecast error is a valid criterion. Numerous examples of loss functions violating this condition are discussed by Hansen and Lunde (2006) and Patton (2009).

Focussing on the univariate dimension, Patton (2009) provides analytical results for the undesirable outcome that arises when using a loss function that violates Hansen and Lunde's (2006) conditions, under different distributional assumption for the returns, different volatility proxies and a number of commonly used loss functions. Furthermore, building upon Hansen and Lunde (2006), he provides necessary and sufficient conditions on the functional form of the loss function (defined within the class of homogeneous

statistical loss functions that can be expressed as means of each period loss) ensuring consistency of the ordering when using a proxy. The following family of functions

$$L(\hat{\sigma}_t^2, \sigma_{t,k}^2) = \begin{cases} \frac{1}{(\xi-1)\xi}\left[(\hat{\sigma}_t^2)^\xi - (\sigma_{t,k}^2)^\xi\right] - \frac{1}{\xi-1}(\sigma_{t,k}^2)^{\xi-1}(\hat{\sigma}_t^2 - \sigma_{t,k}^2) & \text{for } \xi \notin (0,1) \\ \sigma_{t,k}^2 - \hat{\sigma}_t^2 + \hat{\sigma}_t^2 \log\frac{\hat{\sigma}_t^2}{\sigma_{t,k}^2} & \text{for } \xi = 1 \\ \frac{\hat{\sigma}_t^2}{\sigma_{t,k}^2} - \log\frac{\hat{\sigma}_t^2}{\sigma_{t,k}^2} - 1 & \text{for } \xi = 0 \end{cases} \tag{22}$$

represents the subset of consistent homogeneous loss functions. The parameter ξ represents the degree of homogeneity and determines the shape of the function: symmetric ($\xi = 2$) or asymmetric ($\xi \neq 2$). Note that $\xi = 2$ corresponds to the squared forecast error, while $\xi > 2$ (resp. $\xi < 2$) implies that over- (resp. under-) predictions are more heavily penalised.

A generalisation to the multivariate case has been proposed by Patton and Sheppard (2009) and Laurent, Rombouts, and Violante (2009). The latter also show that, under the higher level assumption of consistency of the volatility proxy, the distortion introduced in the ordering when using an inconsistent loss function tends to disappear as the quality of the proxy improves. Since non-robust loss functions might have other desirable properties, as for example down-weighting extreme forecast errors, they may still be used provided that the volatility proxy can be assumed to be sufficiently accurate relative to the degree of similarity between models performances.

Consistency of the ordering and inference on forecast performances

In this section, using a Monte Carlo simulation devoted to illustrate asymptotics that are solely based on $T \to \infty$, we assess to what extent the latent variable problem induces distortions and discuss the role of the quality of the proxy. Although we focus on univariate volatility models, a similar exercise based on the comparison of multivariate models is presented in Laurent, Rombouts, and Violante (2009).

The forecast performances are measured by the loss functions in Table 1. The MSE and the QLIKE loss functions represent the robust loss functions as they satisfy Hansen and Lunde's (2006) condition (column 3) discussed in the previous section. They belong to the family of functions in (22) with $\xi = 2$ and 0 respectively. The other two loss functions, namely Log-MSE and MSE-SD, are based on transformations of the variables of interest. Frequently used in applied work, their use is often justified using the argument that these transformations avoid an excessive penalisation of models that exhibit few extreme forecast errors. The violation of Hansen and Lunde's (2006) condition is shown in column 3.

We generate artificial data at a daily frequency from a non-linear asymmetric GARCH (Engle and Ng, 1993), i.e.

$$y_t = \sigma_t z_t \tag{23}$$

Table 1 Loss functions.

Name	$L(\sigma_t^2, h_t)$	$\partial^2 L(\sigma_t^2, h_t)/(\partial\sigma_t^2)^2$	Status
MSE	$(\sigma_t^2 - h_t)^2$	2	robust
Log-MSE	$\big(\log(\sigma_t^2) - \log(h_t)\big)^2$	$2\frac{1-\log(\sigma_t^2/h_t)}{(\sigma_t^2)^2}$	non-robust
QLIKE	$\frac{\sigma_t^2}{h_t} - \log\frac{\sigma_t^2}{h_t} - 1$	$\frac{1}{(\sigma_t^2)^2}$	robust
MSE-SD	$(\sigma_t - \sqrt{h_t})^2$	$\frac{\sqrt{h_t}}{2\sqrt{(\sigma_t^2)^3}}$	non-robust

$$\sigma_t^2 \quad = \quad \omega + \alpha_1(\epsilon_{t-1} + \gamma\sigma_{t-1})^2 + \beta_1\sigma_{t-1}^2, \tag{24}$$

where $z_t \overset{i.i.d.}{\sim} N(0,1)$ and with parameters $\omega = 0.05$, $\alpha = 0.05$ $\gamma = -0.12$ and $\beta = 0.93$. Following Visser (2010), we also generate intraday returns compatible with model (23)-(24) when aggregated at the daily frequency by setting $z_t = \sum_{i=1}^{N} z_{t,i}$, where $z_{t,i} \overset{i.i.d.}{\sim} N(0, 1/N)$, which satisfies $\text{Var}(z_t) = 1$. The N intraday returns of day t are obtained by assuming that the intraday volatility is constant over the day, i.e., $y_{t,i} = \sigma_t z_{t,i}$. At the highest frequency, we simulated $N = 256$ returns per day. We further aggregate returns, by summation, at 7 lower frequencies, i.e. 128, 64, 32, 16, 8, 4, 2 observations per day.

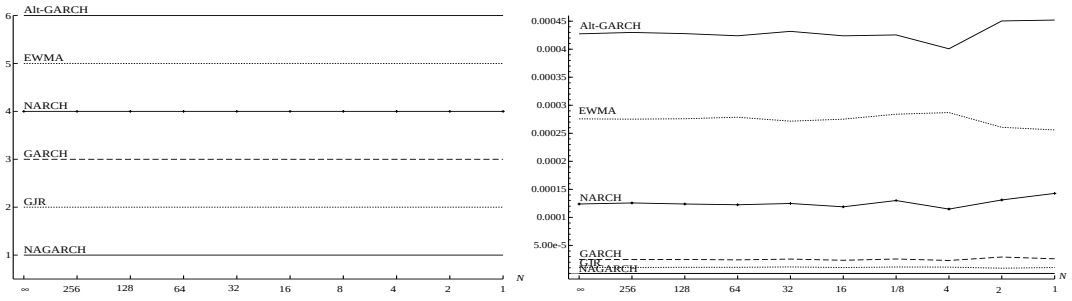
In this setting, and following Andersen and Bollerslev (1998), we dispose of 9 unbiased proxies of the true volatility, denoted $RV_{t,N} = \sum_{i=1}^{N} y_{t,i}^2$ for $N = 256, 128, 64, 32, 16, 8, 4, 2, 1$, ordered in terms of their level of accuracy.

The set of competing models includes, the GARCH (Bollerslev, 1986), the GJR (Glosten, Jagannathan, and Runkle, 1992), the exponential weighted moving average (EWMA) with fixed parameters (J.P.Morgan, 1996), the alternative GARCH (Alt-GARCH) (Knight and Satchell, 2002) and the non-linear ARCH (NARCH) (Higgins and Bera, 1992) models. The models are estimated by QMLE using the first 4000 data points at the daily frequency. 1000 one-step-ahead forecasts are computed using a fixed scheme. The simulations are based on 1000 replications.[6]

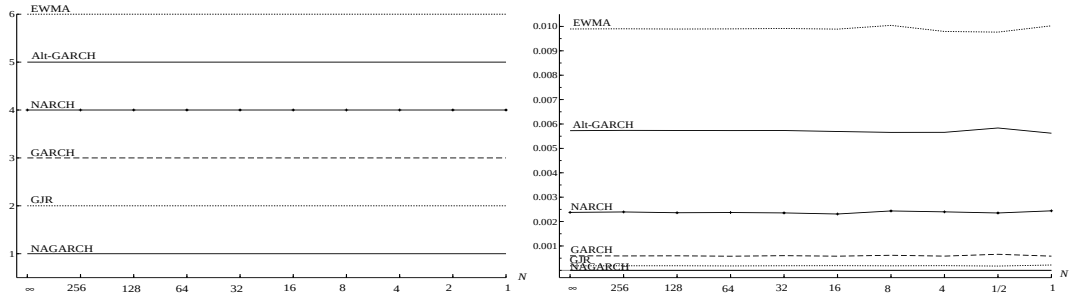
The underlying ordering implied by a given loss function, whether it is robust or not, is identified by ranking forecasts with respect to the true variance, σ_t^2 (indicated by $N = \infty$ in Figures 2 and 3.

[6] All programs have been written by the authors using OxMetrics 6 (Doornik, 2009) and G@RCH 6 (Laurent, 2009).

Figure 2 represents the ranking based on the average sample performances (over the 1000 replications) implied by the two robust loss functions for the true variance ($N = \infty$) and various levels of precision for the proxy ($N = 256$ to $N = 1$). The ranking appears stable and loss differentials between models



a) MSE loss function



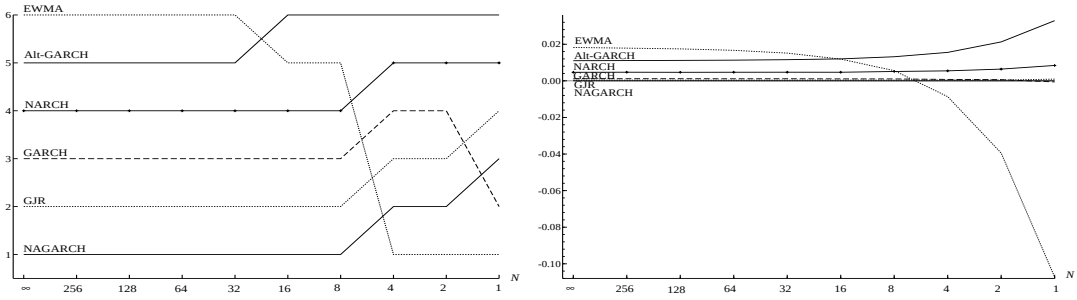
b) QLIKE loss function

Figure 2: Ranking implied by MSE and QLIKE. Ranking based on avg. performances (left) and avg. loss differentials from nagarch (right).

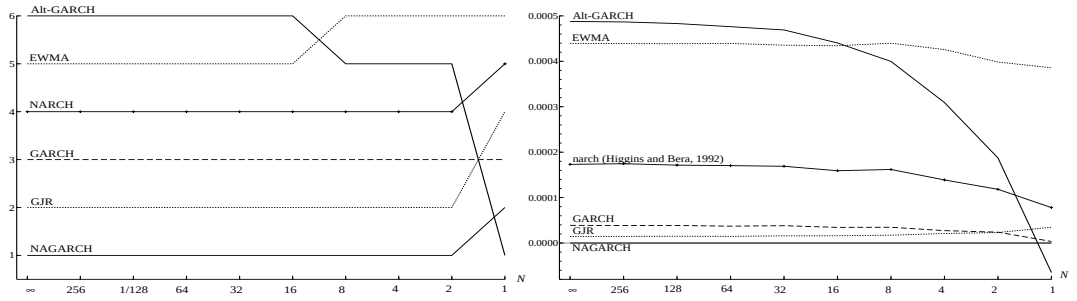
remain constant independently of the level of accuracy of the proxy. Thus, the noise in the volatility proxy is asymptotically irrelevant to the ordering, i.e., the ranking obtained under $RV_{t,N}$ is consistent for the one under the true, latent, conditional variance σ_t^2, for any value of N.

When considering the non-robust loss functions the objective bias becomes striking. Indeed, Figure 3 suggests that for non-robust loss functions, inferior models emerge as the quality of the proxy deteriorates. The relative performance of inferior models begins to improve rapidly and we observe major distortions at all levels of the ranking. For instance, under the Log-MSE, the EWMA model, which ranks last when using the true variance, raises to the top of the ranking when the proxy used in the evaluation is computed using $N = 4$ (or less) intraday returns, while under the MSE-SD, the Alt-GARCH model raises from the last to the first position of the ranking when the evaluation is based on the least accurate proxy ($N = 1$).

In conclusion, for a robust loss function, even when the relative performances are extremely close, the ordering remains unaffected under a noisy proxy and it is always possible to recover asymptotically the



a) Log-MSE loss function



b) MSE-SD loss function

Figure 3: Ranking implied by Log-MSE and MSE-SD. Ranking based on avg. performances (left) and avg. loss differentials from nagarch (right).

true ranking. For a non-robust loss function, it is possible to recover this result only if the volatility proxy is sufficiently accurate relative to the degree of similarity between model performances. However, as the quality of the proxy deteriorates the relative performances of some models appear to improve with respect to the others.

Empirical application

Data description

The empirical application is based on the EUR/USD exchange rate. The models' parameters are estimated using the first 3666 trading days (January 6, 1987 to December 28, 2001). The parameter estimates are then used to compute 1-step ahead forecasts for the following 660 trading days (January 2, 2002 to August 26, 2004). The volatility proxy for the evaluation period is the realised variance of Andersen, Bollerslev, Diebold, and Labys (2003) computed using intra-day returns sampled at the 5-minute frequency ($N = 288$). The forecasting models set includes the six specifications used in the previous section. Model performances are evaluated using two robust loss functions, namely the MSE and the QLIKE.

Table 2 Sample evaluation of forecasting performances.

MSE					
NAGARCH	GARCH	EWMA	Alt-GARCH	GJR	NARCH
0.0388	0.0393	0.0439	0.0383	0.0392	0.0393

QLIKE					
NAGARCH	GARCH	EWMA	Alt-GARCH	GJR	NARCH
0.0774	0.0799	0.114	0.0750	0.0794	0.0767

The sample evaluation of the six competing models is reported in Table 2. Focusing on the MSE loss function, the model exhibiting the best sample performance is the Alt-GARCH, followed by the NAGARCH and the GJR. The worst performing model is the EWMA. A similar ranking is obtained when the evaluation is based on the QLIKE loss function.

The GW tests (with test function $\delta_t = 1$) supports the hypothesis of superior predictive accuracy of the Alt-GARCH with the null hypothesis of zero loss differentials being rejected in favour of this model in

Table 3 GW test (test function $\delta_t = 1$).

MSE

	NAGARCH	GARCH	EWMA	Alt-GARCH	GJR	NARCH
NAGARCH	-	**-2.585**	**-5.399**	1.317	**-2.445**	-0.984
GARCH		-	**-5.384**	**2.271**	**2.916**	0.019
EWMA			-	**4.620**	**5.403**	**4.653**
Alt-GARCH				-	**-2.129**	-1.295
GJR					-	-0.176
NARCH						-

QLIKE

	NAGARCH	GARCH	EWMA	Alt-GARCH	GJR	NARCH
NAGARCH	-	**-3.625**	**-6.691**	**2.445**	**-3.493**	0.673
GARCH		-	**-6.692**	**3.751**	**4.069**	**2.164**
EWMA			-	**6.466**	**6.705**	**6.638**
Alt-GARCH				-	**-3.587**	-1.045
GJR					-	1.916
NARCH						-

Note: Significant values at the 5% confidence level (two-tailed test) in bold indicate the rejection of the null hypothesis of equal predictive ability. The results suggest a preference for the model reported in the row (resp. column) if negative (resp. positive).

Table 4 SPA test.

	MSE			QLIKE		
Benchmark	p_l	p_c	p_u	p_l	p_c	p_u
NAGARCH	**0.12**	**0.21**	**0.35**	0.03	0.03	0.04
GARCH	0.01	0.01	0.02	0.00	0.00	0.00
EWMA	0.00	0.00	0.00	0.00	0.00	0.00
Alt-GARCH	**0.58**	**0.94**	**0.97**	**0.49**	**0.83**	**0.96**
GJR	0.04	0.04	0.06	0.00	0.00	0.00
NARCH	**0.16**	**0.16**	**0.23**	**0.17**	**0.26**	**0.37**

Note: P-values in bold indicate the non-rejection of the SPA null hypothesis for the corresponding benchmark. p_l and p_u denote respectively the lower and the upper bounds for the consistent p-value (p_c).

Table 5 MCS test.

	MSE			QLIKE		
M^0	T_R	T_D	T_{SQ}	T_R	T_D	T_{SQ}
NAGARCH	**0.22**	**0.32**	**0.18**	**0.17**	**0.06**	**0.05**
GARCH	0.01	**0.06**	0.01	0.00	0.00	0.00
EWMA	0.01	0.00	0.00	0.00	0.00	0.00
Alt-GARCH	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
GJR	0.03	**0.17**	**0.09**	0.00	0.01	0.00
NARCH	**0.22**	**0.32**	**0.18**	**0.29**	**0.29**	**0.29**

Note: The models corresponding to the figures in bold represent the MCS at the 5% confidence level.

three cases and non-rejected in two. The results also suggest a preference for the NAGARCH. When we consider the QLIKE loss function the results are similar and the test suggests a preference for Alt-GARCH, NAGARCH but also NARCH. However, as mentioned above, interpreting jointly results based on the pairwise approach may be misleading. When applying the SPA test, we find that 3 models (Alt-GARCH, NAGARCH and NARCH) under the MSE loss function and 2 models (Alt-GARCH and NARCH) under the QLIKE loss function cannot be rejected as the one outperforming all the others.

Finally, the ambiguous result of the GW test, corroborated by the outcome of the SPA test which does not provide a clear identification of a single superior model, supports the use of inference based on the more general MCS. Indeed, the MCS finds different sets depending on the loss function and the statistic used. When performances are evaluated using the MSE loss function, under T_R and T_{SQ} the MCS consists of the Alt-GARCH, the NAGARCH and the NARCH while under T_D the MCS appears to be more conservative, including all models except the EWMA. The results appear to be somewhat homogeneous when using the QLIKE loss function. In this case the three statistics deliver the same MCS which consists of the Alt-GARCH, the NAGARCH and the NARCH models.

Thus, given the composition of the MCS we can conclude that more sophisticated and flexible models are required to fit the dynamics of the conditional variance of the EUR/USD, with emphasis given to the non-linear relationship between conditional variance and innovations.

Conclusion

In this article we provide an overview of methods for volatility forecast evaluation and comparison. We discuss a large variety of methodologies that can be classified in three groups, namely methods for the evaluation of the forecasting accuracy of single forecast, methods for pairwise comparison and methods for multiple comparison.

We pay particular attention to the problems that arise due to the latent nature of the conditional variance. In fact, being the variance unobservable the actual evaluation of the volatility forecasts, usually involving a loss function, requires the use of some proxy. Since this substitution may introduce dramatic distortions in the ordering between forecasts under evaluation, which can be avoided by an appropriate choice of the loss function, we elaborate on the admissible functional form of the loss function and discuss some examples. Using artificial data, we illustrate the danger of combining an inconsistent loss function and a noisy proxy of the true volatility. In this article we focus on methodologies for forecasts evaluation and comparison where the forecast accuracy is measured by a statistical criterion, i.e., means of functions of predictions and predictions errors. At some point, the forecaster may be interested in the economic evaluation of the forecasts, for instance by means of an utility or profit function or yet any other economically meaningful application-specific evaluation criteria. However, to date a comprehensive investigation of the

properties of economic loss function has not been addressed yet. In particular the robustness of the ordering when the evaluation is based on an imperfect volatility proxy remains an open issue and should be further investigated.

References

ANDERSEN, T., AND T. BOLLERSLEV (1998): "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, 885–905.

ANDERSEN, T., T. BOLLERSLEV, F. DIEBOLD, AND P. LABYS (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

ANDERSEN, T., T. BOLLERSLEV, AND N. MEDDAHI (2005): "Correcting the Errors: Volatility Forecast Evaluation Using High-frequency Data and Realized Volatility," *Econometrica*, 73, 279–296.

BARNDORFF-NIELSEN, O., P. HANSEN, A. LUNDE, AND N. SHEPHARD (2008): "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76(6), 1481–1536.

BARNDORFF-NIELSEN, O. E., AND N. SHEPHARD (2002): "Estimating Quadratic Variation using Realised Volatility," *Journal of Applied Econometrics*, 17, 457–477.

BOLLERSLEV, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

——— (2010): "Glossary to ARCH (GARCH)," in *Volatility and Time Series Econometrics - Essays in Honor of Robert Engle*, ed. by T. Bollerslev, J. Russell, and M. Watson. Oxford University Press.

BOLLERSLEV, T., AND J. WOOLDRIDGE (1992): "Quasi-maximum Likelihood Estimation and Inference in Dynamic Models with Time-varying Covariances," *Econometric Reviews*, 11, 143–172.

BRANDT, M., AND F. DIEBOLD (2006): "A No-Arbitrage Approach to Range-Based Estimation of Return Covariances and Correlations," *Journal of Business*, 79, 61–74.

BROCKWELL, P. (2011): "Autoregressive Processes," *Wiley Interdisciplinary Reviews: Computational Statistics*, 3, 316–331.

CLARK, T., AND M. MCCRACKEN (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

——— (2005): "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 24, 369–404.

——— (2009): "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," Federal Reserve of St. Louis Working Paper.

CLARK, T., AND K. WEST (2006): "Using Out-of-sample Mean Squared Prediction Errors to test the Martingale Difference Hypothesis," *Journal of Econometrics*, 135, 155–186.

——— (2007): "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138, 291–311.

CORRADI, V., N. SWANSON, AND C. OLIVETTI (2001): "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315–358.

DIEBOLD, F., AND R. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

DOORNIK, J. (2009): *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press.

ENGLE, R. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1007.

ENGLE, R., AND V. NG (1993): "Measuring and Testing the Impact of News on Volatility," *Journal of Finance*, 48, 1749–1778.

GARMAN, M., AND M. KLASS (1980): "On the Estimation of Securities Price Volatilities," *Journal of Business*, 53, 67–78.

GIACOMINI, G., AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

GLOSTEN, L., R. JAGANNATHAN, AND D. RUNKLE (1992): "On the Relation Between the Expected Value and Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 46, 1779–1801.

HANSEN, P. (2005): "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365–380.

HANSEN, P., AND A. LUNDE (2006): "Consistent Ranking of Volatility Models," *Journal of Econometrics*, 131, 97–121.

——— (2010): "MULCOM 2.00, an Ox^{tm} Software Package for Multiple Comparisons," http://mit.econ.au.dk/vip_htm/alunde/MULCOM/MULCOM.HTM.

HANSEN, P., A. LUNDE, AND J. NASON (2009): "Model Confidence Sets," Forthcoming in *Econometrica*.

HIGGINS, M., AND A. BERA (1992): "A Class of Nonlinear ARCH Models," *International Economic Review,*.

J.P.MORGAN (1996): *Riskmetrics Technical Document, 4th ed.* J.P.Morgan, New York.

KNIGHT, J., AND S. SATCHELL (2002): "Forecasting Volatility in the Financial Markets," in *GARCH processes some exact results, some difficulties and a suggested remedy*, ed. by J. Knight, and S. Satchell. Butterworth-Heinemann.

LAURENT, S. (2009): *G@RCH 6. Estimating and Forecasting Garch Models*. Timberlake Consultants Ltd.

LAURENT, S., J. ROMBOUTS, AND F. VIOLANTE (2009): "On Loss Functions and Ranking Forecasting Performances of Multivariate Volatility Models," Cirano discussion paper 2009-45.

MCCRACKEN, M. (2000): "Robust Out-of-sample Inference," *Journal of Econometrics*, 99, 195–223.

——— (2007): "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics*, 140, 719–752.

MINCER, J., AND V. ZARNOWITZ (1969): "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations*, ed. by J. Mincer.

PARKINSON, M. (1980): "The Extreme Value Method for Estimating the Variance of the Rate of Return," *Journal of Business*, 53, 61–65.

PATTON, A. (2009): "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Forthcoming in Journal of Econometrics*.

PATTON, A., AND K. SHEPPARD (2009): "Evaluating Volatility and Correlation Forecasts," in *Handbook of Financial Time Series*, ed. by T. Andersen, R. Davis, J. Kreiss, and T. Mikosch. Springer.

VISSER, M. (2010): "GARCH Parameter Estimation Using High-Frequency Data," *Journal of Financial Econometrics*, 9, 162–197.

WEISS, A. (1986): "Asymptotic Theory for ARCH Models: Estimation and Testing," *Econometric Theory*, 2, 107–131.

WEST, K. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

WHITE, H. (2000): “Reality Check for Data Snooping,” *Econometrica*, 68, 1097–1126.

ZHANG, L., P. MYKLAND, AND Y. AIT-SAHALIA (2004): “A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data,” *Journal of The American Statistical Association*, 100, 1394–1411.

ZHOU, B. (1996): “High-frequency Data and Volatility in Foreign Exchange Rates,” *Journal of Business and Economic Statistics*, 14, 45–52.