Chapter One

Volatility Forecasts Evaluation and Comparison

Francesco Violante
Maastricht University and CORE, Université catholique de Louvain

Sébastien Laurent
Maastricht University and CORE, Université catholique de Louvain)

1.1 Introduction

In this chapter we review recent developments on volatility forecasts evaluation and comparison based on inference of moments of functions of predictions and predictions errors. We consider one-step ahead forecasts of both univariate and multivariate volatility models, although the results can be easily extended to the multi-step ahead case. Depending on the purpose of the analysis, the forecaster may be interested in evaluating a single model, two or several models. When the object of the analysis is the forecasting accuracy of a single model, the quality of the model can be measured by the correlation between prediction and realizations. A common method that falls in this category is the Mincer-Zarnowitz regression (Mincer and Zarnowitz (1969)), which involves regressing the realization of a variable on a constant and its forecast. Alternatively, the forecaster may aim to compare two or more models. Examples of tests for pairwise comparison have been proposed by Diebold and Mariano (1995), West (1996) and later generalized by Giacomini and White (2006). The multiple comparison problem can be tackled in different ways. We distinguish between two different approaches, the multiple comparison with control (e.g., the reality check for data snooping of White (2000) and the superior predictive ability test of Hansen and Lunde (2005)) where a benchmark forecast is chosen ex-ante and compared to all others to asses whether any of the competing forecasts outperforms the benchmark, and the multiple comparison without control (e.g., the model confidence set test of Hansen et al. (2011)) where all forecasts are compared against each other and poor performing models excluded.

A common problem in the evaluation of point forecasts is the comparison of nested models. In fact, depending on the forecasting scheme used to produce the sequence of out-of-sample observations, most of the tests for predictive ability discussed here may not apply, in the sense that the distribution under the null turns out to be degenerate in some cases and the test may suffer from serious size distortions. In this chapter we consider the three most commonly used forecasting schemes (West (2006)). Let us denote by $\mathcal{T} + T$ the total sample size, where $\mathcal{T}$ is the estimation sample and T is the forecast evaluation sample. In the recursive scheme the sample used to estimate the parameters of the model grows as the forecaster makes predictions for successive observations, i.e. at each step $i = 1, ..., T - 1$ the forecast is based on all past information. In the rolling scheme the sequence of T forecasts is based on parameters estimated using a rolling sample of fixed size $\mathcal{T}$. In the fixed scheme the parameters of the model are estimated only once using data from 1 to $\mathcal{T}$. Then, the estimates are used to generate all forecasts, i.e., at each step, $\mathcal{T} + 1, \mathcal{T} + 2, ..., \mathcal{T} + T - 1$, only the data are updated with the new information.

In this chapter we essentially focus on the rolling and fixed forecasting schemes. In fact, as will be discussed in the following sections, the recursive scheme can be problematic when evaluating forecasts generated by nested models. Furthermore, the rolling and fixed scheme, other than allowing for the comparison of nested models, (Giacomini and White (2006)) also present other advantages. The rolling scheme is, in fact, rather appealing in case of heterogeneneity of the data or parameter drifts that cannot be easily modeled explicitly, whereas the fixed scheme can be useful when it is difficult to carry out parameter estimation. This is often the case, for instance, when comparing multivariate volatility models, where the large number of parameters makes the rolling scheme computationally challenging and time consuming, see Laurent et al. (2011) for an example where a combination of rolling and fixed schemes is used. A number of examples of applications using each of the three schemes can be found in West (2006).

Another problem, which characterizes the comparison of volatility forecasts, is the fact that the target variable is latent. Thus, the evaluation of forecasts or forecasts errors has to be done with respect to some ex-post estimator. Typically, this problem is solved by using a conditionally unbiased (and possibly consistent) estimator as, for example, the squared innovations, the realized volatility or kernels, see Andersen and Bollerslev (1998) and the further developments by Barndorff-Nielsen and Shephard (2002), Zhang et al. (2004), Zhou (1996), Barndorff-Nielsen et al. (2008a) among the others, and their multivariate extension, see Andersen et al. (2003), Barndorff-Nielsen et al. (2008b) and Hansen and Lunde (2006b), or yet range based variance estimators, see Parkinson (1980), Garman and Klass (1980) and Brandt and Diebold (2006). In the reminder of the chapter we refer to the ex-post volatility estimator as the volatility proxy. However, it is not always true that using a conditionally unbiased proxy will lead, asymptotically, to the same outcome that would be obtained if the true volatility was observed. Hansen and Lunde (2006a) show that when the evaluation is based on a target observed with error, the choice of the evaluation

criterion becomes critical in order to avoid a distorted outcome. The problem of consistency, sometimes referred to as robustness, of the ordering between two or more volatility forecasts has been further developed in Patton (2009), Patton and Sheppard (2009) and Laurent et al. (2009).

Finally, since in most methods discussed here, the evaluation of volatility forecasts relies, more or less explicitly, on the ranking implied by a statistical loss function and on the choice of a volatility proxy, we discuss the properties of a number of admissible loss functions and elaborate on the value of high precision proxies. In particular, 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 ranking is robust, the presence of noise in the proxy is likely to affect the power of the test, but not its asymptotic size.

The rest of the chapter is organized as follows. In Section 1.2, we introduce the basic notation used throughout the chapter. In Section 1.3, we discuss the evaluation of the predictive accuracy of single forecasts. In Section 1.4, we introduce the problem of forecast evaluation under imperfect volatility proxies and provide a number of admissible loss functions. In Sections 1.5 and 1.6, we discuss methods for pairwise and multiple forecast comparison respectively. In Section 1.7 we illustrate using artificial data the latent variable problem and its impact on the inference on forecast accuracy. In Sections 1.8, we conclude and discuss directions for further research.

1.2 Notation

We now introduce the basic notation and definitions used throughout the chapter. Let define $t = 1, ..., T$ the time index of the forecast sample of size T. Let r_t be a scalar random variable whose conditional variance, $E[r_t^2|\Im_{t-1}] = E_{t-1}[r_t^2] = \sigma_t$, is of interest (to simplify the exposition, we assume that $E[r_t|\Im_{t-1}] = E_{t-1}[r_t] = 0$). The set $\Im_{t-1}$ denotes the information set at time $t-1$ and contains, for instance, past realizations of r_t. In financial applications, r_t typically represents a sequence of returns, i.e. first difference of logarithmic prices, of some financial asset. We also assume that $r_t|\Im_{t-1} \sim F(0, \sigma_t)$, where F is some distribution with zero mean and finite variance. The (set of) variance forecast(s) (sometimes referred to as models) is denoted by h_t ($h_{k,t} \in \mathcal{H}$, $k = 1, ..., m$ when there are more than one model).

In the multivariate case the variable of interest is the variance matrix, denoted $\Sigma_t = E_{t-1}[\mathbf{r}_t\mathbf{r}_t']$ where $\mathbf{r}_t$ is a $(N \times 1)$ random vector with $\mathbf{r}_t|\Im_{t-1} \sim F(\mathbf{0}, \Sigma_t)$ and Σ_t, whose typical element indexed by $i, j = 1, ..., N$ is denoted $\sigma_{ij,t}$, is symmetric and positive definite. The volatility forecasts are denoted $H_{k,t} \in \mathcal{H}^{N\times N}$, with typical element $h_{ij,k,t},\ i, j = 1, ..., N$, where $\mathcal{H}^{N\times N}$ is a compact subset of the space of symmetric and positive definite matrices $R_{++}^{N\times N}$.

The forecast accuracy is usually evaluated by means of a loss function denoted as $L : R_{++} \times \mathcal{H} \to R^+$ where R_+ and R_{++} correspond respectively to

the non-negative and positive portions of the real line and $\mathcal{H}$ is a compact subset of R_{++} identifying the set of volatility forecasts. In the multivariate case, the loss is defined as $L : R_{++}^{N\times N} \times \mathcal{H}^{N\times N} \to R^{+}$. Note that, both in the univariate and multivariate cases, the first argument of the loss function is the true variance or some proxy of it, whereas the second is a volatility forecast.

As underlined in the previous section, due to the latent nature of the variable of interest, the evaluation of the model forecasts has to rely on a volatility proxy, denoted $\hat{\sigma}_t$ and $\hat{\Sigma}_t$ respectively in the univariate and multivariate cases. The only property that we require for the volatility proxy is conditional unbiasedness, i.e., $E_{t-1}[\hat{\sigma}_t] = \sigma_t$ and $E_{t-1}[\hat{\Sigma}_t] = \Sigma_t \; \forall t$, respectively. Throughout the chapter, we consider the forecasts as observable. However, the forecasts may be biased or inaccurate in any way (e.g., due to parameter uncertainty, misspecification, etc.). About the volatility proxy, if not otherwise stated, we only assume that at least one conditionally unbiased proxy is available. In some specific cases we also require the stronger assumption of consistency or the availability of a sequence of proxies that can be ordered in terms of their accuracy.

A simple variance proxy commonly used in the financial literature is the squared return, or the outer product of the return vector in the multivariate case. However, we discourage the use of such estimator for two reasons. First, although, under mild hypotheses, such proxy is conditionally unbiased for the latent variance, it is extremely noisy, which makes it unsuited in many situations. In fact, the scarce informative content of the volatility proxy could render difficult to asses the statistical relevance of the forecast performances and thus to discriminate between them. Second, even for the smallest multivariate dimension, $N = 2$, this proxy violates the positive definiteness requirement for the volatility proxy. Other variance proxies based on realized moments are discussed in Andersen and Bollerslev (1998), Barndorff-Nielsen and Shephard (2002), Zhang et al. (2004), Zhou (1996), Andersen et al. (2003), Hansen and Lunde (2006b), Barndorff-Nielsen et al. (2008a) and Barndorff-Nielsen et al. (2008b), among the others. Range based variance estimators can be found in Parkinson (1980), Garman and Klass (1980) and Brandt and Diebold (2006).

1.3 Single forecast evaluation

A simple method for evaluating the accuracy of a volatility forecast is the well known Mincer-Zarnowitz (MZ) regression, see Mincer and Zarnowitz (1969). This approach requires the estimation of the coefficients of a regression of the target on a constant and a time series of forecasts, i.e.

$$\sigma_t = \alpha + \beta h_t + \epsilon_t \qquad \forall t = 1, ..., T. \tag{1.1}$$

The null hypothesis of optimality of the forecast can be written as $H^0 : \alpha = 0 \cup \beta = 1$. Given the latent nature of the target variable, the regression in (1.1) is unfeasible. Substituting the true variance by some conditionally unbiased

proxy, $\hat{\sigma}_t = \sigma_t + \eta_t$ with $E_{t-1}[\eta_t] = 0$, we can rewrite (1.1) as

$$\hat{\sigma_t} = \alpha + \beta h_t + e_t, \tag{1.2}$$

where the innovation are $e_t = \eta_t + \epsilon_t$. Since $\hat{\sigma_t}$ is a conditionally unbiased estimator of the true variance then (1.2) yields unbiased estimates of α and β. The MZ regression allows to evaluate two different aspects of the volatility forecast. First, the MZ regression allows to test the presence of systematic over- or under-predictions, that is whether the forecast is biased, by testing the joint hypothesis $H^0 : \alpha = 0 \cup \beta = 1$. Second, being the R^2 of (1.2) an indicator of the correlation between the realization and the forecast, it can be used as evaluation criterion of the accuracy of the forecast.

Since the variance of the innovation term e_t in (1.2) depends on the accuracy of the volatility proxy, when a high quality proxy is available, the regression parameters are estimated more accurately. Also the R^2 of the regression in (1.2), $1 - Cov(\hat{\sigma_t}, h_t)^2/(Var(\hat{\sigma_t})Var(h_t))$, results penalized as the quality of the proxy deteriorates, see Andersen and Bollerslev (1998) for an analytical example.

The R^2 of the MZ regression has often been used as a criterion for ordering over a set of volatility forecasts, see Andersen and Bollerslev (1998) and Andersen et al. (2003) for instance. Furthermore, to respond to the concern that few extreme observations can drive the forecast evaluation, many authors have argued in favor of MZ regressions on transformations of σ_t (and consequently $\hat{\sigma_t}$ and h_t), for instance $\log(\hat{\sigma_t})$ on $\log(h_t)$ or $|r_t|$ on $\sqrt{h_t}$, see Pagan and Schwert (1990), Jorion (1995), Bollerslev and Wright (2001) among others for some examples.

Although appealing, this approach suffers from an important weakness. In fact, as pointed out by Andersen et al. (2005), transformed unbiased forecasts for the latent variance are not generally unbiased for the transformed proxy, $\hat{\sigma}_t$. However, allowing for $\alpha \neq 0$ and/or $\beta \neq 1$ in the MZ regression of the volatility proxy on the transformed forecasts explicitly corrects what would appear as signal of bias in the forecasts. Analytical examples under different distributional assumption for the volatility proxy can be found in Patton and Sheppard (2009). It is important to point out that these drawbacks are only due to the substitution of the true volatility by the proxy. For the unfeasible transformed regression, i.e., if the true volatility was observable, the null $H^0 : \alpha = 0 \cup \beta = 1$ would still apply for the transformed regression.

Also related to the use of transformations of the variables of interest, Hansen and Lunde (2006a) show that, due to the latent variable problem, the R^2 of the MZ regression based on transformed variables is not always adequate and may lead to a perverse ordering. They establish that a sufficient conditions for the R^2 to be valid criterion is $E_{t-1}[\sigma_t - \hat{\sigma}_t](\partial^i \phi(\sigma_t)/\partial \sigma_t^i) = c_i$ for some constant c_i, $\forall t = 1, 2, ...$ and $i \in N$ 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 et al. (2005), the quadratic transformation of an un-

biased forecasts will not generally result to be unbiased for $\hat{\sigma}_t^2$, but rejects, for example, the log-regression.[1]

Another suitable property for a good forecast is that the forecasts or forecast errors are uncorrelated with other series or more generally with any other information available at the time the forecast is made. If this is not the case, then it would be possible to use such information to produce superior forecasts, see Mincer and Zarnowitz (1969), Figlewsky and Wachtel (1981), Zarnowitz (1985) and Keane and Runkle (1990) among the others. Furthermore, including additional variables, such as lagged values of the volatility or of the standardized volatility, sign indicators or yet transformations and combinations of these variables, allows to detect whether nonlinearities, asymmetries and persistence have been neglected. This approach is called augmented MZ regression, where the augmentation consists in adding to the right hand side of (1.2) the term $\mathbf{z}_t\gamma$, where $\mathbf{z}_t$ represents the set of measurable additional regressors. The relevant null hypothesis becomes $H^0 : \alpha = 0 \cup \beta = 1 \cup \gamma = 0$.

Other than a test for unbiasedness and forecast accuracy, the MZ regression can also be viewed as a test of efficiency, i.e. $E[h_t(\hat{\sigma_t} - h_t)] = 0$. In fact, if forecasts and forecast errors are correlated then it would be possible to produce superior forecasts by exploiting this relationship. From (1.2) we have

$$\hat{\sigma_t} - h_t = \alpha + (\beta - 1)h_t + e_t \tag{1.3}$$

and therefore

$$E[h_t(\hat{\sigma_t} - h_t)] = \alpha E[h_t] + (\beta - 1)E[h_t^2] + E[h_t e_t] = 0, \tag{1.4}$$

when $\alpha = 0$ and $\beta = 1$.

The results outlined above can be directly extended, with few exceptions to the multivariate case. A simple approach is to consider the unique elements of the true variance matrix (proxy) and of the covariance forecast. The feasible MZ regression can be written as

$$vech(\hat{\Sigma}_t) = \boldsymbol{\alpha} + diag(\boldsymbol{\beta})vech(H_t) + \boldsymbol{e}_t, \tag{1.5}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(N(N+1)/2 \times 1)$ vectors of parameters, $vech(.)$ is the half-vector operator and $diag(.)$ is the operator that transforms a vector into a square diagonal matrix with the vector along the diagonal. Equation (1.5) can be estimated by seemingly unrelated regression techniques. Then a joint

[1] Note that, although according to Hansen and Lunde (2006a) the R^2 of the quadratic MZ regression is a robust criterion in the sense that it leaves the ranking between volatility forecasts unaffected when the latent variance is substituted by a proxy, the quadratic transformation of an unbiased forecasts does not generally result to be unbiased for the quadratic transformation of the volatility proxy, $\hat{\sigma}_t$, see Andersen et al. (2005). As an example assume $r_t|\Im_{t-1} \sim N(0, \sigma_t)$ and consider the volatility proxy $\hat{\sigma}_t = r_t^2$. The quadratic MZ regression ($\phi(x) = x^2$) can be written as $(\hat{\sigma}_t)^2 = \alpha + \beta h_t^2 + e_t$. Under the null $H^0 : h_t = \sigma_t \ a.s. \ \forall t$ the population values of the parameters are $\alpha = 0$ and $\beta = 3$. For the unfeasible transformed regression, i.e., if the true volatility was observable, the null $H^0 : \alpha = 0 \cup \beta = 1$ would still apply for the transformed regression.

test that $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{1}$ can be performed. As pointed out by Patton and Sheppard (2009) the large dimension of the system my adversely affect the finite sample properties of the joint test. The solution proposed to reduce the parameter space is to impose in (1.5) the parameter constrains $\alpha_i = \alpha$ and $\beta_i = \beta \; \forall i = 1, ..., N(N+1)/2$.

1.4 Loss functions and the latent variable problem

A common approach to the evaluation of forecast performances is the comparison of expected losses evaluated with respect to the true variance. However, as noted in Section 1.3, the latent nature of the conditional variance makes it difficult to evaluate the performances of volatility forecasts. The latent variable problem can be solved, at least partly, by substituting the true conditional variance by some ex-post estimator based on observed quantities as they become available. Examples of volatility proxies have been provided in Section 1.1.

Obviously a good volatility proxy must be conditionally unbiased. However, as first noted by Andersen and Bollerslev (1998) and Andersen et al. (2005), it is not always the case that the evaluation of forecast performances using a conditionally unbiased proxy will lead, asymptotically, to the same outcome that would be obtained if the true volatility was used. Hansen and Lunde (2006a), focussing on a qualitative assessment (ordering) of volatility forecasts, show that when the evaluation is based on a target observed with error, the choice of the evaluation criterion becomes critical in order to avoid a perverse outcome. They define the theoretical framework for the analysis of the ordering of stochastic sequences and provide conditions on the functional form of the loss function which ensure consistency between the ordering based on a volatility proxy and the one based on the true, but latent, variance.

Let us define the precision, measured in terms of expected loss, of some generic volatility forecast, $h_{k,t}$, with respect to the true variance as $E[L(\sigma_t, h_{k,t})]$. The aim is to seek conditions that ensure consistency of the ranking (equivalence of the ordering) between any two forecasts k and j when a conditionally unbiased proxy is substituted to the true variance, that is

$$E[L(\sigma_t, h_{k,t})] \leq E[L(\sigma_t, h_{j,t})] \Leftrightarrow E[L(\hat{\sigma}_t, h_{k,t})] \leq E[L(\hat{\sigma}_t, h_{j,t})], \qquad (1.6)$$

where k and j refer to two competing volatility forecasts. The violation of (1.6) is defined as objective bias. A sufficient condition to ensure (1.6) is the following

$$\frac{\partial^2 L(\sigma_t, h_t)}{(\partial \sigma_t)^2} \quad \text{exists and does not depend on } h_t. \qquad (1.7)$$

It follows immediately that (1.7) rejects evaluation criteria commonly used in applied work such as absolute deviations, root of squared errors, or proportional error loss functions, whereas it validates the use of squared errors. Numerous

examples of loss functions violating (1.7) are discussed by Hansen and Lunde (2006a), Patton (2009), Patton and Sheppard (2009) and Laurent et al. (2009). See also Section 1.7 for an analytical illustration.

Focussing on the univariate dimension, Patton (2009) provides analytical results for the undesirable outcome that arises when using a loss function that violates (1.7), under different distributional assumption for the returns and considering different volatility proxies and for a number of commonly used loss functions. Furthermore, building upon Hansen and Lunde (2006a), he provides necessary and sufficient conditions on the functional form for 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 represents the entire subset of consistent homogeneous loss functions, with degree of homogeneity given by ξ:[2]

$$L(\hat{\sigma}_t, h_t) = \begin{cases} \frac{1}{(\xi-1)\xi}(\hat{\sigma}_t^{\xi} - h_t^{\xi}) - \frac{1}{\xi-1} h_t^{\xi-1}(\hat{\sigma}_t - h_t) & \text{for } \xi \notin (0,1) \\ h_t - \hat{\sigma}_t + \hat{\sigma}_t \log \frac{\hat{\sigma}_t}{h_t} & \text{for } \xi = 1 \\ \frac{\hat{\sigma}_t}{h_t} - \log \frac{\hat{\sigma}_t}{h_t} - 1 & \text{for } \xi = 0 \end{cases} . \tag{1.8}$$

The loss function in (1.8) can take a variety of shapes: symmetric, ($\xi = 2$ corresponds to the mean squared prediction error loss function) and asymmetric with penalty on overpredictions ($\xi > 2$) or underpredictions ($\xi < 2$). The set of consistent loss functions in (1.8) relates to the class of linear exponential densities of Gourieroux et al. (1984) and, as noted by Laurent et al. (2009) it partially coincides with the subset of homogeneous loss functions associated with the most important linear exponential densities.[3] In fact, for $\xi = 0, 1, 2$, the function can be alternatively derived from the objective functions of the Gaussian, Poisson and Gamma densities respectively, see Gourieroux and Monfort (1995) for details.

A first of generalization to the multivariate case has been proposed by Patton and Sheppard (2009). Laurent et al. (2009) complete this setting and provide a general framework for the evaluation of variance matrices. They identify a number of robust vector and matrix loss functions and provide insight on their properties, interpretation and geometrical representation. In the multivariate case, the sufficient condition in (1.6) becomes

$$\frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_{l,t} \partial \sigma_{m,t}} \quad \text{finite and independent of } H_t \ \forall l, m = 1, ..., N(N+1)/2, \tag{1.9}$$

[2] Recall that a function is homogeneous if it has a multiplicative scaling behavior, i.e., if $f : X \to Y$ and ξ integer, then f is homogeneous of degree ξ if $f(kx) = k^{\xi} f(x), \forall k > 0$.

[3] Recall that a function $f(x)$ is homogeneous of degree α if it satisfies $f(kx) = k^{\alpha} f(x)$ for all non-zero k.

where $\sigma_{l,t}$ is the lth element of the vector $\boldsymbol{\sigma}_t = vech(\Sigma_t)$. Given (1.9), a generalized necessary and sufficient functional form for the loss functions is

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)' vech(\hat{\Sigma}_t - H_t), \tag{1.10}$$

where $\tilde{C}(.) : R_{++}^{N\times N} \rightarrow R_+$ with

$$C(H_t) = \begin{bmatrix} \frac{\partial \tilde{C}(H_t)}{\partial h_{1,t}} \\ \vdots \\ \frac{\partial \tilde{C}(H_t)}{\partial h_{K,t}} \end{bmatrix}, \quad C'(H_t) = \begin{bmatrix} \frac{\partial \tilde{C}(H_t)}{\partial h_{1,t}\partial h_{1,t}} & \cdots & \frac{\partial \tilde{C}(H_t)}{\partial h_{1,t}\partial h_{K,t}} \\ \vdots & \ddots & \\ \frac{\partial \tilde{C}(H_t)}{\partial h_{K,t}\partial h_{1,t}} & & \frac{\partial \tilde{C}(H_t)}{\partial h_{K,t}\partial h_{K,t}} \end{bmatrix}, \tag{1.11}$$

where $C(.)$ and $C'(.)$ are the gradient and the hessian of $\tilde{C}(.)$ with respect to the $K = N(N+1)/2$ unique elements of H_t, denoted $\mathbf{h}_t = vech(H_t)$ and $C'(H_t)$ is negative definite. The general form in (1.10) is related to the linear exponential matrix densities and represents the family of the Bregman matrix divergences criteria. Well known loss functions belonging to this family are the Frobenius distance, the von Neumann divergence and the Stein loss function (also called LogDet or Burg divergence).

From (1.10), Laurent et al. (2009) identify the entire subset of homogeneous ($\xi = 2$) loss functions based on forecast errors, i.e., $(\Sigma_t - H_t)$, which can be expressed as

$$L(\hat{\Sigma}_t, H_t) = L(\hat{\Sigma}_t - H_t) = vech(\hat{\Sigma}_t - H_t)' \hat{\Lambda} vech(\hat{\Sigma}_t - H_t), \tag{1.12}$$

where $\hat{\Lambda}$ is a positive definite matrix of constants which defines the weights assigned to the elements of the forecast error matrix. The loss function in (1.12) nests a number of MSE-type loss functions, defined on both vector and matrix spaces, e.g. (weighted) Euclidean distance on the half-vectorization of the forecast error matrix or Frobenius distance on the difference of realized and predicted variance matrices, Σ_t and H_t.

Unlike the univariate case, where an analytical expression is available for the entire class of consistent loss functions, in the multivariate case such generalization is unfeasible because there are infinite combinations forecasts and forecasts errors, hence of functions $\tilde{C}(.)$, that satisfy (1.10). However, given (1.10), application-specific loss functions can be easily derived. Laurent et al. (2009) provide and discuss a number of examples.

Finally, Laurent et al. (2009) 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 often non robust loss functions have other desirable properties which are useful in applications, e.g. down-weighting extreme forecast errors, they may still be used provided that the volatility proxy is sufficiently accurate.

In the following sections, we review a number of tests for forecast evaluation where performances are evaluated by means of a statistical loss function. Although most of the methodologies discussed are valid under a general loss

function, we remind that, in empirical applications, when the true variance is substituted by a proxy, the loss function should be chosen, depending on the setting, according to (1.8) and (1.10) respectively.

1.5 Pairwise comparison

The first approach to pairwise comparison that we consider is the test of equal predictive ability proposed by Ashley et al. (1980) as a generalization of the approach introduced by Granger and Newbold (1977). The test is based on the comparison of the mean square forecast errors (MSE) of a pair of forecasts with respect to the target. Let us define $e_{k,t} = \sigma_t - h_{k,t}$ the forecast error and $L_{k,t}^{MSE} = T^{-1}\sum_t e_{k,t}^2$ the mean square forecast error of some model k with respect to σ_t. Then, when comparing the performance of model k to some other model j, simple algebra yields

$$L_{k,t}^{MSE} - L_{j,t}^{MSE} = (Var(e_{k,t}) - Var(e_{j,t})) + (\bar{e}_k^2 - \bar{e}_j^2), \tag{1.13}$$

where $\bar{e}_i = T^{-1}\sum_t e_{i,t}$. Let us now define $D_t = e_{k,t} - e_{j,t}$, $S_t = e_{k,t} + e_{j,t}$ and $\bar{D}$, $\bar{S}$ their empirical means. Then, (1.13) can be rewritten as

$$L_{k,t}^{MSE} - L_{j,t}^{MSE} = Cov(D_t, S_t) + \bar{D}\bar{S}. \tag{1.14}$$

A test of equal predictive ability, or more precisely equal MSE, corresponds to testing the null hypothesis

$$H^0 : Cov(D_t, S_t) = 0 \cup \bar{D} = 0. \tag{1.15}$$

Note that (1.15) implies that the forecasts can be biased. In fact, $\bar{D}_t = 0$ does not require $e_{k,t} = e_{j,t} = 0$ but only that the biases are equal in size and sign. The null hypothesis in (1.15) is equivalent to testing the null hypothesis $H^0 : \alpha = 0 \cup \beta = 0$ in the following regression

$$D_t = \alpha + \beta(S_t - \bar{S}) + \epsilon_t. \tag{1.16}$$

If the forecast errors have zero-mean, i.e., they are both unbiased, and under the additional assumption that they are normally distributed and uncorrelated, the test of equal MSE is equivalent to the test proposed by Granger and Newbold (1977), henceforth GN, that is

$$GN - T = \frac{\rho}{\sqrt{(T-1)^{-1}(1-\rho)^2}} \sim t_{T-1}, \tag{1.17}$$

where $\rho = Cov(D_t, S_t)/\sqrt{Var(D_t)Var(S_t)}$ and t_{T-1} is the student-t distribution with $T-1$ degrees of freedom.

The extension to the multivariate case is straightforward. In fact, the MSE can computed using the Euclidean distance, $L_{k,t}^{E} = T^{-1}\sum_t \left[\sum_{i\leq j} e_{ij,k,t}^2\right]$,

$i, j = 1, ..., N$ or the Frobenius distance, $L^F_{k,t} = T^{-1} \sum_t \left[\sum_{i,j} e^2_{ij,k,t} \right]$, $i, j = 1, ..., N$, although it is worth noting that in the latter case the covariance forecast errors are double weighted. Given that these loss functions can be expressed as a linear combination of MSEs on the unique or all elements of the forecast error matrix respectively, a joint test on the coefficient of the pooled regression $D_{ij,t}$ on $(S_{ij,t} - \bar{S})$ can be performed using standard panel data techniques.

Another approach to pairwise comparison that we consider is the test of equal predictive ability proposed by Diebold and Mariano (1995), henceforth DM, and further refined by West (1996), McCracken (2000), Clark and McCracken (2001), Corradi et al. (2001), Clark and West (2006), Clark and West (2007), McCracken (2007) and Clark and McCracken (2005). The DM test is a very general procedure[4] designed to compare two rival forecasts in terms of their forecasting accuracy using a general loss function. The measure of predictive accuracy, i.e. the loss function, can be specified according to the definition of optimality adopted by the forecaster.

Consider a loss function as defined in Section 1.2 and define the loss differential between model k and j as

$$d_t = L(\sigma_t, h_{k,t}) - L(\sigma_t, h_{j,t}), \tag{1.18}$$

in the univariate case, and

$$d_t = L(\Sigma_t, H_{k,t}) - L(\Sigma_t, H_{j,t}), \tag{1.19}$$

in the multivariate case. Since in either case the loss function is scalar valued, we can more generally refer to the notation $d_t = L_{i,t} - L_{j,t}$. Under the assumption that d_t is stationary, $E[d_t]$ is well defined and allows for the formulation of the null hypothesis of equal predictive ability $H^0 : E[d_t] = 0$. The test takes the form of a t-statistic, i.e.,

$$DM - T = \frac{\sqrt{T}\bar{d}}{\sqrt{\omega}} \overset{a}{\sim} \mathrm{N}(0, 1), \tag{1.20}$$

where $\bar{d} = T^{-1} \sum_t d_t$ and $\omega = \lim_{t \to \infty} Var(\sqrt{T}\bar{d})$ 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 the DM type tests is to infer about $E[d_t(\theta_0)]$ using $T^{-1} \sum_t d_t(\theta_0)$, where θ_0 represents the models parameter population values, and thus are based on asymptotics requiring the size of the estimation sample $\mathcal{T}$ and the forecast evaluation sample T to grow to infinity at the same rate.[5]

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

[5] 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

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 $T^{-1}\sum_t d_t(\hat{\theta}) - E[d_t(\theta_0)] \underset{p}{\to} 0$ when models are nested, $T^{-1}\sum_t d_t(\hat{\theta})$ is a non-degenerate random variable. Based on this argument, they suggest several statistics suited for testing equal predictive accuracy, whose distribution is non standard and depends on the parameter uncertainty. To obtain the distribution under the null hypothesis, Clark and McCracken (2009) propose an asymptotically valid procedure based on bootstrap sampling.

To allow for a unified treatment of nested and non-nested models, Giacomini and White (2006) (henceforth GW) suggest to approach the problem of the forecast evaluation as a problem of inference about conditional (rather than unconditional) expectations of forecast errors. The GW is a test of finite-sample predictive ability. Giacomini and White (2006) 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 $E[d_t(\hat{\theta})]$ using $T^{-1}\sum_t d_t(\hat{\theta})$. The GW approach tests the null hypothesis of equal predictive ability

$$E[L(\sigma_t, h_{k,t}^{\tau_k}(\hat{\theta}_{k,t}^{\tau_k})) - L(\sigma_t, h_{j,t}^{\tau_j}(\hat{\theta}_{j,t}^{\tau_k}))|\Im_{t-1}] \equiv E[d_{\mathcal{T},t}|\Im_{t-1}] = 0, \quad (1.21)$$

where, for $i = k, j$, $h_{i,t}^{\tau_i}(\hat{\theta}_{i,t}^{\tau_i})$ are $\Im_{t-1}$-measurable forecasts, τ_i is size of the estimation window, possibly different for each model and $\mathcal{T} = \max(\tau_k, \tau_j)$. Since under the null hypothesis, $\{d_{\mathcal{T},t}, \Im_t\}$ is a martingale difference sequence, (1.21) is equivalent to $E[\delta_{t-1}d_{\mathcal{T},t}] = 0$, where δ_{t-1}, referred to as the test function, is a $\Im_{t-1}$-measurable vector of dimension q. By invoking standard asymptotic normality arguments, the GW test takes the form of a Wald-type statistic

$$GW - T_{\mathcal{T}}^{\delta} = T\left(T^{-1}\sum_{t=1}^{T}\delta_{t-1}d_{\mathcal{T},t}\right)'\hat{\Omega}^{-1}\left(T^{-1}\sum_{t=1}^{T}\delta_{t-1}d_{\mathcal{T},t}\right), \quad (1.22)$$

where $\hat{\Omega}$ is a consistent estimator of the variance of $\delta_{t-1}d_{\mathcal{T},t}$. The statistic is asymptotically χ_q^2 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. Note that, in the case where $\tau_k = \tau_j$ and $\delta_t = 1\ \forall t$, then the GW test is equivalent to a 'conditional' DM

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).

test with forecasts evaluated using the rolling window forecast scheme. Apart from this simple case we are not aware of any other application of the GW approach (for instance allowing for more sophisticated test functions, $\tau_k \neq \tau_j$, time dependent estimation windows, different forecasting rules/methods or yet different estimation procedures for each model).

Clearly, the GW asymptotics hold when the size of the estimation sample is fixed and the forecast sample grows, i.e., $\mathcal{T}$ fixed, $T \to \infty$, but also under a rolling scheme[6] and in general to any limited memory estimator.

1.6 Multiple comparison

When multiple alternative forecasts are available, it is of interest to test whether a specific forecast (hereafter the benchmark), selected independently from the data, produces systematically superior (at least equivalent) performances with respect to the rival models. In this case, we aim to test the null hypothesis that the benchmark is not inferior to any other alternative. This approach, called multiple comparison with control, differs from the techniques discussed in Section 1.5 for two reasons. First, the multiple comparison allows to recognize the multiplicity effect, i.e., statistical relevance of all comparisons between the benchmark and each of the alternative models, and calls for a test of multiple hypotheses to control for the size of the overall testing procedure. Second, while Section 1.5 involves tests of equal predictive ability, the choice of a control requires a test of superior predictive ability. The distinction is crucial because, while the former lead to simple null hypotheses, i.e., testing equalities, the latter leads to composite hypotheses, i.e. testing (weak) inequalities. The main complications in composite hypotheses testing is that (asymptotic) distributions typically depend on nuisance parameters, hence the distribution under the null is not unique.

To simplify the exposition, the notation used in this section only refers the univariate dimension. Since all the techniques discussed hereafter are based on comparisons of forecast performances measured by some statistical loss function, the extension to the multivariate case, as noted in Section 1.5, is straightforward and only involves an appropriate redefinition of the loss function, namely $L : R_{++}^{N\times N} \times \mathcal{H}^{N\times N} \to R^{+}$. Issues related to the choice of the loss function and to the latent variable problem have been discussed in Section 1.4.

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, $h_{0,t}$, and some rival forecast, $h_{k,t}$ $k = 1, ..., m$ as

$$d_{k,t} = L(\sigma_t, h_{0,t}) - L(\sigma_t, h_{k,t}) \tag{1.23}$$

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

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

$$H^0 : E[\mathbf{d}_t] \leq \mathbf{0}, \quad (\text{or equivalently } H^0 : \max_k E[d_{k,t}] \leq 0) \tag{1.24}$$

that is, the benchmark is superior to the best alternative. Clearly, the null hypothesis in (1.24) is a multiple hypothesis, i.e., the intersection of the one-sided individual hypotheses $E[d_{k,t}] \leq 0$. The RC statistic takes the form

$$RC - T = \max_k(\sqrt{T}\,\bar{d}_k), \tag{1.25}$$

where $\bar{d}_k = T^{-1}\sum_{t=1}^{T} d_{k,t}$. Note that, as in Diebold and Mariano (1995), the RC test is based on asymptotics which require the parameters of the model based forecasts to converge to their population values, thus not allowing for the comparison of nested models. Using similar arguments of Giacomini and White (2006), Hansen (2005) extends the procedure to the comparison of nested models. The framework defined in Hansen (2005) is well suited when parameters are estimated once, i.e., fixed scheme, or using a moving window (of fixed or time dependent size or yet of different size for each model), i.e., rolling schemes, whereas the comparison of models with parameters that are estimated recursively is not accommodated.

Given strict stationary of $\mathbf{d}_t$, White (2000) invokes conditions provided in West (1996) that lead to

$$\sqrt{T}(\bar{\mathbf{d}} - E[\mathbf{d}_t]) \xrightarrow{d} \mathrm{N}(0, \Omega). \tag{1.26}$$

The challenge when implementing the RC test is that (1.25) has an asymptotic distribution under the null hypothesis depending on the nuisance parameters $E[\mathbf{d}_t]$ and Ω. One way to proceed is to substitute a consistent estimator for Ω and employ the least favorable configuration (LFC) over the values of $E[\mathbf{d}_t]$ that satisfy the null hypothesis. From (1.24), it is clear that the least favorable value to the alternative is $E[\mathbf{d}_t] = 0$, which presumes that all alternatives are as good as the benchmark. However, the distribution of (1.25), i.e., the extreme value of a vector of correlated normal variables, is unknown. White (2000) suggests two ways to obtain the distribution under the LFC for the alternative, namely the Monte Carlo Reality Check (simulated inference) and the Bootstrap Reality Check (bootstrap inference). We refer to White (2000) for further details on the two methods.

Using a similar approach, Hansen (2005) proposes a new test for superior predictive ability (henceforth SPA). His framework differs from White (2000) in two ways. First, he proposes a different statistic based on studentized quantities to alleviate the substantial loss of power that the RC can suffer due to the inclusion of poor and irrelevant forecasts. 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 LFC suggested in White

(2000) implicitly relies on an asymptotic distribution under the null hypothesis that assumes $E[d_{k,t}] = 0$ for all k, Hansen (2005) points out that all negative values of $E[d_{k,t}]$ have also to be considered since they conform with the null hypothesis.

The new statistic takes the form

$$SPA - T = \max \left[\max_k \frac{\sqrt{T}\bar{d}_k}{\sqrt{\hat{\omega}_k}}, 0 \right], \tag{1.27}$$

where $\hat{\omega}_k$ is some consistent estimator of $\omega_k = \lim_{t\to\infty} Var(\sqrt{T}\bar{d}_k)$, i.e. the kth diagonal element of Ω. 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 $\boldsymbol{\mu} = E[\mathbf{d}_t]$ that conforms with the null hypothesis. The suggested estimator is

$$\hat{\mu}_k^c = \bar{d}_k \, 1_{\left\{\sqrt{T}\,\bar{d}_k/\hat{\omega}_k \leq -(2 \log \log T)^{1/2}\right\}}, \tag{1.28}$$

where $1_{\{.\}}$ denotes the indicator function. The threshold $(2 \log \log T)^{1/2}$ in (1.28) represents the slowest rate that captures all alternatives with $\mu_k = 0$. More generally, any threshold in the interval $\left[(2 \log \log T)^{1/2},\ T^{1/2-\epsilon}\right]$, for any $\epsilon > 0$ also produces a valid test and guarantees that all poor models are discarded asymptotically. For instance, Hansen (2005) proposes the value $0.25\,T^{0.25}$. Furthermore, since different threshold rates lead to different p-values in finite samples, Hansen (2005) also provides a lower and upper bound for the SPA p-values. These p-values can be obtained by using $\hat{\mu}_k^l = \min(\bar{d}_k, 0)$ and $\hat{\mu}_k^u = 0$, where the latter yields a distribution under the null based on the LFC principle.[7] Hansen (2005) also provide a detailed description of the bootstrap scheme used to obtain the distribution under the null hypothesis.

Clearly, when comparing volatility models, the choice of a benchmark is not obvious. Furthermore, especially when the set of competing models is large, such applications may not yield a single model that is significantly superior to the alternatives because the data may not be sufficiently informative to give an univocal answer. In this case, 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, known as multiple comparison without control, suggests a comparison of all models with each other. It differs from the techniques discussed above for two reasons. First, the procedure does not require a benchmark to be specified. Second, the testing procedure generally relies on simple hypotheses, i.e., equalities.

Hansen et al. (2011) construct a sequential test of equal predictive ability, dubbed model confidence set (MCS), which, given an initial set of forecasts M^0, allows to: *i)* test the null that no forecast is distinguishable from any other,

[7] In the latter case the distribution under the null is obtained using the same arguments as in White (2000). The difference here stands in the fact that the variable of interest is the maximum of studentized quantities, whereas in White (2000) it is the maximum of non-studentized quantities.

ii) discard any inferior forecasts if they exist, *iii)* characterize the set of models that are (equivalent to each other and) superior to all the discarded models. The set of surviving model is called model confidence set and can be interpreted as a confidence interval for the forecasts in that it is the set containing the best forecast at some confidence level.

Designed around the testing principle of Pantula (1989) to ensure that sequential testing does not affect the overall size of the test, the MCS test involves a sequence of tests for equal predictive ability. Given 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 $d_{k,j,t} = L(\sigma_t, h_{k,t}) - L(\sigma_t, h_{j,t})$, for all $k, j \in M^0$ and $k \neq j$. Under the assumption that $d_{k,j,t}$ is stationary, the null hypothesis of equal predictive ability takes the form

$$H^0 : E[d_{k,j,t}] = 0 \qquad \forall k, j \in M^0. \tag{1.29}$$

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, $M^* \equiv \{k \in M_0 : E(d_{k,j,t}) \leq 0 \, \forall \, j \in M^0\}$, for the best model in M^0.

Let $\mathbf{L}_t$ be the $(m \times 1)$ vector of sample performances $L(\sigma_t, h_{k,t})$, $k \in M$ 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 (1.29). Hence, (1.29) is equivalent to $E[\iota'_\perp \mathbf{L}_t] = 0$ 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 Normal with mean 0 and covariance matrix $\Omega = \lim_{t\to\infty} Var\left(T^{-1/2} \sum_{t=1}^{T} \iota'_\perp \mathbf{L}_t\right)$. Thus, it seems natural to employ traditional quadratic-form type tests as

$$MCS - T_Q = T \left(T^{-1} \sum_{t=1}^{T} \iota'_\perp \mathbf{L}_t \right)' \hat{\Omega}^+ \left(T^{-1} \sum_{t=1}^{T} \iota'_\perp \mathbf{L}_t \right) \tag{1.30}$$

and

$$MCS - T_F = \frac{T-q}{q(T-1)} \, MCS - T_Q, \tag{1.31}$$

where $\hat{\Omega}$ is some consistent estimator of Ω, $q = \text{rank}(\hat{\Omega})$ denotes the number of linearly independent contrasts and $\hat{\Omega}^+$ denotes the More-Penrose pseudo-inverse of $\hat{\Omega}$. The statistic in (1.30) is asymptotically χ^2_q, whereas (1.31) is asymptotically $\text{F}_{q,T-q}$ under the null hypothesis, as the subscripts Q (quadtratic) and F (F-distributed) suggest.

However, when m is large, it might be difficult to obtain an accurate estimate of Ω. Alternatively, Hansen et al. (2011) also propose three simpler statistics which only require the estimation of the diagonal elements of Ω. The

drawback is that their distribution under the null becomes non-standard. To this respect, Hansen et al. (2011) provide a detailed description of the bootstrap scheme employed to solve the nuisance parameter problem and to obtain the distribution under the null hypothesis. Similarly to the SPA, the three statistics are expressed as functions of studentized quantities.

The first statistic is a sum of deviations (hence the subscript) from the common average. Under the null hypothesis $H^0 = E[\bar{d}_k] = 0, \forall k \in M$, the statistic takes the form[8]

$$MCS - T_D = \frac{1}{m} \sum_{k \in M} t_k^2, \tag{1.32}$$

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 $\omega_k^D = lim_{t \to \infty} Var(\sqrt{T}\bar{d}_k)$. The remaining two statistics, dubbed range and semi-quadratic, take the form

$$MCS - T_R = \max_{k,j \in M} |t_{k,j}| \qquad \text{and} \qquad MCS - T_{SQ} = \frac{1}{m} \sum_{k,j \in M} t_{k,j}^2 \tag{1.33}$$

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 variances $\hat{\omega}_s^R$ are consistent estimators of $\omega_s^R = Avar(\sqrt{T}\bar{d}_{k,j})$.

If the null hypothesis is rejected, then Hansen et al. (2011) 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 standardized excess loss relative to the average across models. The iterative testing procedure ends as soon there is the first non rejection, or obviously if all forecasts 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 $H^0_{M^i}$, i.e., at the ith step of the iteration process. By convention the p-value when there is only one surviving model is $p_m = 1$. The tests for multiple comparison mentioned in this section are implemented in Hansen and Lunde (2010).

1.7 Consistency of the ordering and inference on forecast performances

In this section we illustrate, using a Monte Carlo simulation, to what extent the latent variable problem induces distortions in the ranking and affects the inference on forecast accuracy.

[8] Note that this null hypothesis is equivalent to (1.29).

We focus on univariate volatility models, whereas a similar exercise based on the comparison of multivariate models is presented in Laurent et al. (2009) and Laurent et al. (2011).

The forecast performances are measured by the following two loss functions

1. L_{MSE}: $L(\hat{\sigma}_t, h_{k,t}) = (\hat{\sigma}_t - h_{k,t})^2$ (mean squared error)

2. L_{LMSE}: $L(\hat{\sigma}_t, h_{k,t}) = \left(log(\hat{\sigma}_t) - log(h_{k,t})\right)^2$ (mean squared error of the log transform).

Note that, while L_{MSE} belongs to the family defined in (1.8) with $\xi = 2$ (henceforth referred to as 'robust'), it is straightforward to show that L_{LMSE} violates (1.7) (henceforth 'non-robust'), that is

$$\begin{aligned} L'_{LMSE} &= \frac{\partial L(\sigma_t, h_t)}{\partial \sigma_t} = 2\frac{\log(\sigma_t/h_{k,t})}{\sigma_t} \\ L''_{LMSE} &= \frac{\partial^2 L(\sigma_t, h_t)}{(\partial \sigma_t)^2} = 2\frac{1 - \log(\sigma_t/h_{k,t})}{\sigma_t^2}, \end{aligned}$$

with the second derivative depending on $h_{k,t}$. The choice of L_{LMSE} is not coincidental. Patton (2009) quantifies, under different assumption on the distribution of the returns, the bias with respect to the optimal forecast when using this loss function. To illustrate the centrality of the role of the quality of the volatility proxy when the evaluation of forecast performances is based on a loss function that violates (1.7), consider the conditional expectation of the second order Taylor expansion of L_{LMSE} around the true value σ_t, i.e.

$$\begin{aligned} E[L_{LMSE}(\hat{\sigma}_t, h_{k,t}) \mid \Im_{t-1}] &\approx L_{LMSE}(\sigma_t, h_{k,t}) + L^{'}_{LMSE} E[\eta_t \mid \Im_{t-1}] \\ &+ 0.5 L^{''}_{LMSE}(\sigma_t, h_t) E[\eta_t^2 \mid \Im_{t-1}], \end{aligned}$$

where $\eta_t = (\hat{\sigma}_t - \sigma_t)$, σ_t and $h_{k,t}$ are $\Im_{t-1}$ measurable and, since the volatility proxy is required to be conditionally unbiased, $E[\eta_t \mid \Im_{t-1}] = 0$ and $E[\eta_t^2 \mid \Im_{t-1}] < \infty$ is the conditional variance of the proxy. Let us now define

$$\begin{aligned} \Delta(h_{k,t}) &= E[L_{LMS}(\hat{\sigma}_t, h_{k,t}) \mid \Im_{t-1}] - L_{LMSE}(\sigma_t, h_{k,t}) \\ &= 0.5 L^{''}_{LMSE}(\sigma_t, h_{k,t}) E[\eta_t^2 \mid \Im_{t-1}] \\ \Delta(h_{j,t}) &= E[L(\hat{\sigma}_t, h_{j,t}) \mid \Im_{t-1}] - L_{LMSE}(\sigma_t, h_{j,t})] \\ &= 0.5 L^{''}_{LMSE}(\sigma_t, h_{j,t}) E[\eta_t^2 \mid \Im_{t-1}] \end{aligned}$$

for a pair of forecast k and j. Then we have

$$\begin{aligned} \Delta(h_{k,t}) - \Delta(h_{j,t}) &= 0.5 \left(L^{''}_{LMSE}(\sigma_t, h_{k,t}) - L^{''}_{LMSE}(\sigma_t, h_{j,t}) \right) E[\eta_t^2 \mid \Im_{t-1}] \\ &\neq 0. \end{aligned}$$

Since, apart from coincidental cancelation, $L^{''}_{LMSE}(\sigma_t, h_{k,t}) \neq L^{''}_{LMSE}(\sigma_t, h_{j,t})$, then the order implied by the proxy is likely to differ from the one implied by

the true variance and the bias in the ranking is more likely to appear as the quality of the proxy deteriorates. On the other hand, the true ordering is likely to be preserved as the proxy becomes nearly perfect, i.e., $E[\eta_t^2 \mid \Im_{t-1}] \to 0$.

We generate artificial data from an Exponential GARCH(0,1) diffusion, see Nelson (1991a) for details, that is

$$\begin{bmatrix} \mathrm{dp(t)} \\ \mathrm{d}\log(\sigma(\mathrm{t})) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 - 0.05\log(\sigma_(t)) \end{bmatrix} \mathrm{dt} + \begin{bmatrix} \sigma_(t) & -0.1\sqrt{\sigma(t)} \\ -0.1\sqrt{\sigma(t)} & 0.01 + 0.04(1 - 2/\pi) \end{bmatrix}^{1/2} \begin{bmatrix} \mathrm{dW}_1(\mathrm{t}) \\ \mathrm{dW}_2(\mathrm{t}) \end{bmatrix}, \quad (1.34)$$

where $\mathrm{dW}_i(t)$, $i = 1, 2$ are two independent Brownian motions. The simulation is based on 500 replications. Using an Euler discretization scheme of (1.34), we approximate the continuous time process by generating 7200 observation/day. All the competing models are estimated by QMLE using data aggregated at daily frequency. The estimation is based on a fixed sample of size equal to 1500 days. The forecast evaluation sample amounts to 1000 days which are used for the one-step ahead forecasts evaluation. All programs have been written using OxMetrics (Doornik (2009)) by the authors. The estimation of the models and the MCS have been performed using the G@RCH (Laurent (2009)) and MULCOM (Hansen and Lunde (2010)) respectively.

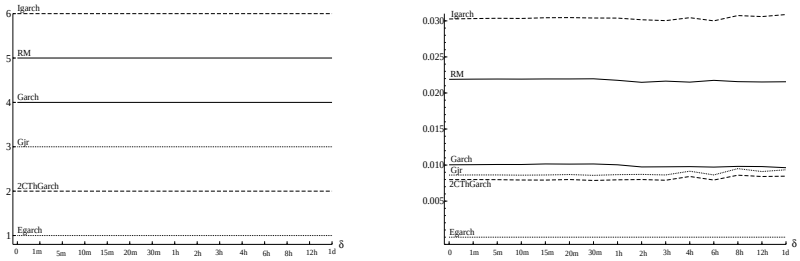
The set of competing models includes, Exponential (Egarch) (Nelson (1991b)), Garch (Bollerslev (1986)), Gjr (Glosten et al. (1992)), Integrated (Igarch) (Engle and Bollerslev (1986)), RM (J.P.Morgan (1996)) and 2-Components Threshold Garch (2CThGarch) (Engle and Lee (1999)) models. The latent variance is computed as $\sigma_t = \int_{t-1}^{t} \sigma(u)\ \mathrm{du}$, $t \in \mathbb{N}$. The proxy is the realized variance of Andersen and Bollerslev (1998), i.e., the sum of intraday squared returns, and is computed using returns sampled at 14 different frequencies ranging from 1-minute to 1-day. The proxy is denoted $\hat{\sigma}_{t,\delta}$, where δ =1m,5m,...,1h,...,1d represents the sampling frequency. In this setting the realized variance estimator is conditionally unbiased, allows to control for the accuracy of the proxy (through the level of aggregation of the data δ) and it is also consistent, i.e., $\hat{\sigma}_{t,\delta} \underset{p}{\to} \sigma_t$ as $\delta \to 0$. 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 (denoted as $\delta = 0$ in Figure 1.1).

Figure 1.1(a) represents the ranking based on the average sample performances (over the 500 replications) implied by the robust loss function, L_{MSE}, for the true variance ($\delta = 0$) and various levels of precision for the proxy (δ =1m to δ =1d). The ranking appears stable and loss differentials between models remain constant independently of the level of accuracy of the proxy. Thus, the ranking obtained under $\hat{\sigma}_{t,\delta}$ is consistent for the one under the true conditional variance σ_t, for all values of δ.

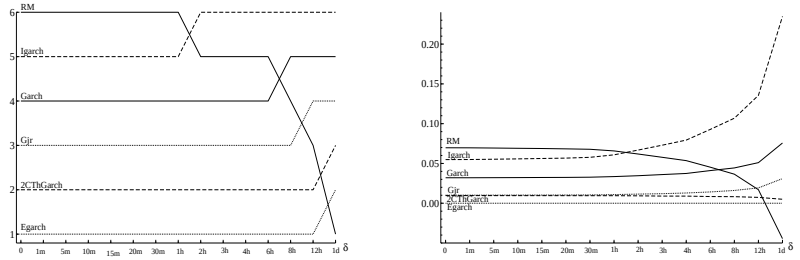
When looking at the non-robust loss function, L_{LMSE}, the evidence of the objective bias is striking. In fact, although the consistency of the proxy ensures

convergence of the proxy-based ordering to the true one as $\delta \to 0$, which is the case when the ranking is based on $\hat{\sigma}_{t,\delta}$ computed using returns sampled at frequency higher than 1-hour (Figure 1.1(b)), as the quality of the proxy deteriorates inferior models emerge. The relative performances of inferior models seems to improve rapidly and we observe major distortions at all levels of the ranking.

We now compare the forecast performances of our set of models using the MCS test. Ideally the MCS, i.e., the set of superior models, should be a singleton containing the true process, i.e., the Egarch. However, as the quality of the proxy deteriorates, losses and thus loss differentials, become less informative, which in turn makes more difficult to efficiently discriminate between models. Consequently we expect the set of superior models to grow in size as δ increases. Figure 1.2 reports two statistics, the frequency at which the Egarch is in the MCS, which shows the size properties of the test (left) and the average number of models in the MCS, which is informative about the power properties of the test (right). As before, the results are reported as a function of the precision of the proxy, δ. The levels of confidence considered are $\alpha = [0.25, 0.1]$. The statistic considered is the $MCS - T_D$. The number of bootstrap samples used to obtain the distribution under the null is set to 1000.



a) L_{MSE} loss function

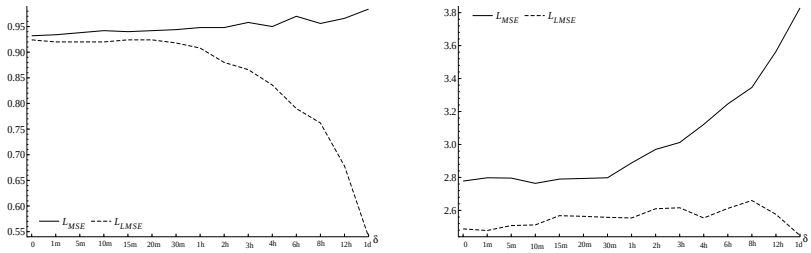


b) L_{LMSE} loss function

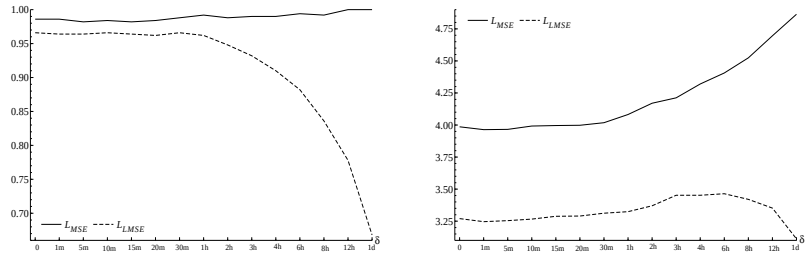
Figure 1.1 Ranking implied by L_{MSE} and L_{LMSE}. Ranking based on avg. performances (left) and avg. loss differentials from Egarch (right).

When considering the robust L_{MSE}, and the evaluation is based on an accurate proxy, the MCS approach is able to correctly separate between superior and

poor performing models. The deterioration of the precision of the proxy only translates into a loss of power, i.e., a larger MCS. In fact, the MCS includes the true model with probability that converges to one. These results clearly demonstrate the value of high precision proxies. Estimators based on relatively high frequency returns provide sensible gains in power. The inference based on the non-robust L_{LMSE} is reliable only when a highly accurate proxy is available. In this case, as the quality of the proxy deteriorates we identify on average a smaller MCS but the probability that the set of superior models contains the true model reduces dramatically. As expected, the threshold, in terms of accuracy of the proxy, after which the MCS under L_{LMSE} breaks down coincides with $\delta =$1h, i.e., when the objective bias starts affecting the ranking, see Figure 1.1 (b).



a) $\alpha = 0.25$



b) $\alpha = 0.10$

Figure 1.2 Size (left) and power (right) indicators for the MCS test under L_{MSE} (solid) and L_{LMSE} (dashed).

Concluding, although the MCS testing procedure is formally valid, an unfortunate choice of the loss function can lead to undesired outcomes and results in an incorrect identification of the set of superior models.

1.8 Conclusion

In this chapter we provide an overview of methods for volatility forecast evaluation and comparison. We consider both the univariate and multivariate setting.

We discuss a large variety of methodologies that can be classified into 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.

We also emphasize the importance of high precision proxies. In fact, even if the forecast under evaluation is highly informative, the variable of interest is always some measure of forecast error. The informativeness of the latter, that allows to efficiently distinguish between superior and poor models, depends crucially on the quality of the proxy.

We show, using artificial data, how both size and power properties of some of the most well known tests for predictive accuracy behave under both robust and non-robust loss functions and different levels of accuracy of the volatility proxy.

In this chapter we focused 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 might 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 criterion. 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.

Chapter 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.

Ashley, R., C. Granger, and R. Schmelensee (1980). Advertising and aggregate consumption: An analysis of causality. *Econometrica 48*, 1149–1168.

Barndorff-Nielsen, O., P. Hansen, A. Lunde, and N. Shephard (2008a). 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., P. Hansen, A. Lunde, and N. Shephard (2008b). 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.

Bollerslev, T. and J. Wright (2001). High-frequency data, frequency domain inference and volatility forecasting. *Review of Economics and Statistics 83*, 596–602.

Brandt, M. W. and F. X. Diebold (2006). A no-arbitrage approach to range-based estimation of return covariances and correlations. *Journal of Busines 79*, 61–73.

Clark, T. and M. McCracken (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105*, 85–110.

Clark, T. and M. McCracken (2005). Evaluating direct multistep forecasts. *Econometric Reviews 24*, 369–404.

Clark, T. and M. McCracken (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.

Clark, T. and K. West (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*. London: Timberlake Consultants Press.

Engle, R. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric Reviews 5*, 1–50.

Engle, R. and G. Lee (1999). *A Permanent and Transitory Component Model of Stock Return Volatility,* , pp. 475–497. Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger. Oxford University Press - R. Engle and H. White eds.

Figlewsky, S. and P. Wachtel (1981). The formation of inflationary expectation. *Review of Economics and Statistics 63*, 1–10.

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.

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models*. Cambridge University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods theory. *Econometrica 52*, 681–700.

Granger, C. and P. Newbold (1977). *Forecasting Economic Time Seies*. Academic Press.

Hansen, P. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics 23*, 365–380.

Hansen, P. and A. Lunde (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1). *Journal of Applied Econometrics 20*, 873–889.

Hansen, P. and A. Lunde (2006a). Consistent ranking of volatility models. *Journal of Econometrics 131*, 97–121.

Hansen, P. and A. Lunde (2006b). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics 24*, 127–218.

Hansen, P. and A. Lunde (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 (2011). The model confidence set. *Econometrica 79*, 453–497.

Jorion, P. (1995). Predicting volatility in the foreign exchange market. *Journal of Finance 50*, 507–528.

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

Keane, M. and D. Runkle (1990). Testing the rationality of price forecasts: New evidence from panel data. *American Economic Review 80*, 714–735.

Laurent, S. (2009). *G@RCH 6. Estimating and Forecasting Garch Models*. London: 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.

Laurent, S., J. Rombouts, and F. Violante (2011). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*. 26: n/a. doi: 10.1002/jae.1248.

McCracken, M. (2000). Robust out-of-sample inference. *Journal of Econometrics 99*, 195–223.

McCracken, M. (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 J. Mincer (Ed.), *Economic Forecasts and Expectations*.

Nelson, D. (1991a). Arch models as a diffusion approximation. *Journal of Econometrics 45*, 7–38.

Nelson, D. (1991b). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica 59*, 349–370.

Pagan, A. and W. Schwert (1990). Alternative models for conditional stock volatility. *Journal of Econometrics 45*, 267–290.

Pantula, S. (1989). Testing for unit roots in time series data. *Econometric Theory 5*, 256–271.

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 T. Andersen, R. Davis, J. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Springer.

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica 64*, 1067–1084.

West, K. (2006). Forecasts evaluation. In G. Elliot, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*.

White, H. (2000). Reality check for data snooping. *Econometrica 68*, 1097–1126.

Zarnowitz, V. (1985). Rational expectations and macroeconomic forecasts. *Journal of The American Statistical Association 100*, 1394–1411.

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.