ON LOSS FUNCTIONS AND RANKING FORECASTING PERFORMANCES OF MULTIVARIATE VOLATILITY MODELS

Sébastien Laurent[1], Jeroen V.K. Rombouts[2] and Francesco Violante[3]

December 7, 2011

Abstract

A large number of parameterizations have been proposed to model conditional variance dynamics in a multivariate framework. However, little is known about the ranking of multivariate volatility models in terms of their forecasting ability. The ranking of multivariate volatility models is inherently problematic because when the unobservable volatility is substituted by a proxy, the ordering implied by a loss function may result to be biased with respect to the intended one. We point out that the size of the distortion is strictly tied to the level of the accuracy of the volatility proxy. We propose a generalized necessary and sufficient functional form for a class of non-metric distance measures of the Bregman type, suited to vector and matrix spaces, which ensure consistency of the ordering when the target is observed with noise. An application to three foreign exchange rates, where we compare the forecasting performance of 24 multivariate GARCH specifications over two forecast horizons, is provided.

Keywords: Volatility, Multivariate GARCH, Matrix norm, Loss function, Model Confidence Set.
JEL Classification: C10, C32, C51, C52, C53, G10.

[1]Maastricht University, The Netherlands; Université catholique de Louvain, CORE, B-1348, Louvain-la-Neuve, Belgium. Address: Department of Quantitative Economics, Maastricht University, School of Business and Economics, P.O. Box 616, 6200 MD, The Netherlands. Tel: +31 43 3883843; Fax: +31 43 3884874; E-mail: s.laurent@maastrichtuniversity.nl

[2]Institute of Applied Economics, HEC Montréal, CIRANO, CIRPEE; Université catholique de Louvain, CORE, B-1348, Louvain-la-Neuve, Belgium. Address: 3000 Cote Sainte Catherine, Montréal (QC), Canada, H3T 2A7. Tel: +1 514 3406466; Fax: +1 514 3406469; E-mail: jeroen.rombouts@hec.ca

[3]Maastricht University, The Netherlands; Université catholique de Louvain, CORE, B-1348, Louvain-la-Neuve, Belgium. Address: Department of Quantitative Economics, Maastricht University, School of Business and Economics, P.O. Box 616, 6200 MD, The Netherlands. Tel: +31 43 3883843; Fax: +31 43 3884874; E-mail: f.violante@maastrichtuniversity.nl

We would like to thank Luc Bauwens, Mohammed Bouaddi, Raouf Boucekkine, Giovanni Motta, Franz Palm, Andrew Patton, Lars Stentoft, the participants to the 2^{nd} International Workshop of the ERCIM Working Group on Computing and Statistics, the 64^{th} European Meeting of the Econometric Society and the 26^{th} Canadian Econometric Study Group for their helpful comments. Financial support from the HEC Montreal and the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State Prime Minister's Office, science policy programming, is gratefully acknowledged.

1 Introduction

A special feature of economic forecasting compared to general economic modeling is that we can measure a model's performance by comparing its forecasts to the outcomes when they become available. Generally, several forecasting models are available for the same variable and forecasting performances are evaluated by means of a loss function. Elliott and Timmermann (2008) provide an excellent survey on the state of the art of forecasting in economics. Details on volatility and correlation forecasting can be found in Andersen, Bollerslev, Christoffersen, and Diebold (2006).

The evaluation of volatility forecasts raises the problem that the variable of interest is latent. This problem can be solved by replacing the latent conditional variance by a proxy, see e.g. the realized variance estimator of Andersen and Bollerslev (1998). However, as noted by Andersen, Bollerslev, and Meddahi (2005), the evaluation of volatility forecasts using a proxy may not lead, asymptotically, to the same ordering that would be obtained if the true volatility was observed. In a general framework, Hansen and Lunde (2006a) show that when the evaluation is based on a target observed with error, the choice of the loss function becomes critical in order to avoid a distorted outcome and provide conditions on the functional form of the loss function which ensure consistency of the proxy based ordering. Their results can be cast in the invariant decision rules setting, see for instance Ferguson (1967). For univariate volatility, Patton (2011) derives a class of loss functions which are able to order consistently in the presence of noise. Building on these results, Patton and Sheppard (2009) give a direct multivariate analogue but without providing any general expression.

The above results have important implications on testing procedures for superior predictive ability, see among others Diebold and Mariano (1995), West (1996), Clark and McCracken (2001), the reality check by White (2000) and the contributions of Hansen and Lunde (2005) with the superior predictive ability test and Hansen, Lunde, and Nason (2011) with the model confidence set test. In fact, when the target variable is unobservable, an unfortunate choice of the loss function may deliver unintended results even when the testing procedure is formally valid. With respect to the evaluation of multivariate volatility forecasts little is known about the properties of the loss functions. This is the first paper that addresses this issue.

In this paper, we provide a general framework to the evaluation of multivariate volatility forecasts, in the two following ways. First, we cast to the multivariate dimension the

sufficient conditions that a loss function has to satisfy to deliver the same ordering whether the evaluation is based on the true conditional variance matrix or an unbiased proxy of it. Moreover, instead of focussing only on the detection of the optimal forecast as in Patton and Sheppard (2009), we show that when the proxy is sufficiently accurate with respect to the degree of similarity between models' performances, inconsistent loss functions can still deliver an unbiased ranking. This result is important because several commonly used loss functions do not satisfy such conditions. Nevertheless, they may have other desirable properties, like down-weighting extreme forecast errors for example.

Second, we propose a generalized necessary and sufficient functional form for a class of non-metric distance measures suited to vector and matrix spaces which ensure consistency of the ordering. The resulting class of distance functions represents a particular instance of Bregman divergences, see Bregman (1967), and is related to the family of linear exponential densities of Gourieroux, Monfort, and Trognon (1984) and Banerjee, Merugu, Dhillon, and Ghosh (2005). As in Patton (2011), we focus on homogeneous statistical loss functions expressed as sample means of each period loss as homogeneity ensures invariance of the ordering under rescaling of the data. As we shall point out, unlike the univariate case, parametric expressions for the entire class of consistent loss functions cannot be derived. However, from the generalized functional form, we identify a range of Bregman-type functions, such as the Kullback-Leibler divergence, the Itakura-Saito distance, the von Neumann divergence and the Stein distance. Nevertheless, we are able to provide a parametric expression for the entire subset of homogeneous loss functions based on the difference between volatility forecasts and proxies. This function nests a number of squared error-type loss functions, e.g. weighted Euclidean distance, Mahalanobis distance and Frobenius distance.

To make our theoretical results concrete, we evaluate the predictive accuracy of 24 multivariate GARCH models using a portfolio of three exchange rates (Euro, UK pound and Japanese yen against the US dollar) at two forecast horizons. Models' superior predictive accuracy is inferred using the model confidence set test of Hansen, Lunde, and Nason (2011). The advantage of choosing a consistent loss function to evaluate model performances is striking. The ranking based on an inconsistent loss function, together with an uninformative proxy, is found to be severely biased. As the quality of the proxy deteriorates, inferior models emerge and outperform models which are otherwise preferred when the comparison is based

on a more accurate proxy. The results also clearly demonstrate the value of high precision proxies. Although the loss function ensures consistency of the ordering, only a high precision proxy allows to efficiently discriminate between models.

The rest of the paper is organized as follows. In Section 2 we introduce the notation and the general setup. In Section 3 we provide conditions for consistency of the ranking and derive the admissible functional form of the loss function. In Section 4 we present an empirical application using three exchange rates. In Section 5 we conclude and discuss directions for further research. All proofs are provided in the appendix.

2 Notation and setup

We denote $\mathbb{R}_{++}^{N\times N}$ the space of $N \times N$ symmetric positive definite matrices and $\dot{H} \subset \mathbb{R}_{++}^{N\times N}$ a compact subset of $\mathbb{R}_{++}^{N\times N}$. $\dot{H}$ represents the set of variance matrix forecasts with typical elements $H_{m,t} \in \dot{H}$, where m denotes the mth model and with $t = 1, ..., T$ (where T is the forecast sample size). The matrix $\Sigma_t \in \mathbb{R}_{++}^{N\times N}$ denotes the true but unobservable conditional variance matrix and $\hat{\Sigma}_t$ a proxy. The minimum requirement for $\hat{\Sigma}_t$ is conditional unbiasedness and we always assume that at least one such proxy is available. We define $L(\cdot, \cdot)$ an integrable loss function $L : \mathbb{R}_{++}^{N\times N} \times \dot{H} \to \mathbb{R}_+$, such that $L(\Sigma_t, H_{m,t})$ is the loss of model m with respect to Σ_t. $\mathbb{R}_+$ denotes the positive part of the real line. We refer to the ordering based on the expected loss, $\mathrm{E}[L(\Sigma_t, H_{m,t})] < \infty$ as the true ordering. Similarly, $L(\hat{\Sigma}_t, H_{m,t})$ is the loss with respect to the proxy $\hat{\Sigma}_t$, and $\mathrm{E}[L(\hat{\Sigma}_t, H_{m,t})] < \infty$ determines the approximated ranking over $\dot{H}$. When needed, we also refer to the empirical ranking as the one based on the sample evaluation of $L(\hat{\Sigma}_t, H_{m,t})$, i.e., $T^{-1}\sum_t L(\hat{\Sigma}_t, H_{m,t})$. The set, $\Im_{t-1}$ denotes the information at time $t-1$ and $\mathrm{E}_{t-1}(\cdot) \equiv \mathrm{E}(\cdot|\Im_{t-1})$ the conditional expectation. The elements, $\sigma_{i,j,t}$, $\hat{\sigma}_{i,j,t}$ and $h_{i,j,t}$ indexed by $i, j = 1, ..., N$, refer to the elements of the matrices Σ_t, $\hat{\Sigma}_t$, H_t respectively. Furthermore, $\sigma_{k,t}$, $\hat{\sigma}_{k,t}$ and $h_{k,t}$ are the elements, indexed by $k = 1, ..., N(N+1)/2$, of the vectors $\sigma_t = \mathrm{vech}(\Sigma_t)$, $\hat{\sigma}_t = \mathrm{vech}(\hat{\Sigma}_t)$ and $h_t = \mathrm{vech}(H_t)$ respectively, where $\mathrm{vech}(\cdot)$ is the operator that stacks the lower triangular portion of a matrix into a vector. Finally, the vectorized difference between the true variance matrix and its proxy is denoted by $\xi_t = (\hat{\sigma}_t - \sigma_t)$.

The following assumptions ensure that the loss function $L(\cdot, \cdot)$ is able to correctly order with respect to the true variance matrix.

A1.1 $L(\cdot,\cdot)$ is continuous and uniquely minimized at H_t^*, which represents the optimal forecast. If $H_t^* \in int(\dot{H})$, $L(\cdot,\cdot)$ is strictly convex on $\dot{H}$.

A1.2 $L(\cdot,\cdot)$ is such that the optimal forecast equals the true conditional variance Σ_t, i.e.,

$$H_t^* = \underset{H_t \in \dot{H}}{\arg\min} L(\Sigma_t, H_t) \Leftrightarrow H_t^* = \Sigma_t. \tag{1}$$

Unless otherwise stated, in the remainder of the paper we always assume that the loss function meets all these requirements. The function $L(\cdot,\cdot)$ does not need to be a metric: symmetry and triangular inequality do not need to hold. Throughout the paper, we also normalize the loss function such that it implies zero loss if and only if $\Sigma_t = H_t$.

We rely on the definition of consistency of the ranking introduced by Hansen and Lunde (2006a) and adopted by Patton (2011). That is, the ranking between any two models $H_{l,t}, H_{m,t} \in \dot{H}$, $l \neq m$, is consistent when it is the same whether it is based on the true conditional variance or a conditionally unbiased proxy. More formally

$$\mathrm{E}(L(\Sigma_t, H_{l,t})) \geq \mathrm{E}(L(\Sigma_t, H_{m,t})) \Leftrightarrow \mathrm{E}(L(\hat{\Sigma}_t, H_{l,t})) \geq \mathrm{E}(L(\hat{\Sigma}_t, H_{m,t})). \tag{2}$$

3 Consistent ranking and distance functions

3.1 Consistency of the ranking and proxy accuracy

Starting from the general framework for the analysis of the ordering of stochastic sequences elaborated in Hansen and Lunde (2006a), we assess the role of the proxy in the appearance of the objective bias. We point out that the appearance of a distortion in the ranking is strictly tied to the level of the accuracy of the volatility proxy. To be precise, consider a sequence of volatility proxies $\hat{\Sigma}_t^{(s)}$ indexed by s and denote $H_t^{*(s)}$ such that

$$H_t^{*(s)} = \underset{H_t \in int\dot{H}}{\arg\min} \mathrm{E}_{t-1}[L(\hat{\Sigma}_t^{(s)}, H_t)]. \tag{3}$$

The index s can be thought of, for instance, as the sampling frequency used to compute the realized covariance proxy in Andersen, Bollerslev, Diebold, and Labys (2003).
We make use of the following assumptions:

A2.1 Σ_t and H_t are $\Im_{t-1}$ measurable.

A2.2 $L(\cdot,\cdot)$ is twice continuously differentiable with respect to $\hat{\sigma}_t$ and h_t.

A2.3 **a)** $\xi_t^{(s)} = (\hat{\sigma}_t^{(s)} - \sigma_t^{(s)})$ is a vector martingale difference sequence with respect to $\Im_t$ $\forall s$ with conditional variance matrix $V_t^{(s)} = \mathrm{E}_{t-1}[\xi_t^{(s)}\xi_t^{(s)\prime}] < \infty$ $\forall s$,

b) $V_t^{(s)} \xrightarrow{p} 0$ as $s \to \infty$.

By Assumption A2.1, Σ_t and H_t are considered as observable, which implies that conditionally on $\Im_{t-1}$, the difference between $L(\Sigma_t, H_t)$ and $L(\hat{\Sigma}_t^{(s)}, H_t)$ depends only on $\hat{\Sigma}_t^{(s)}$, $\forall H_t \in \dot{H}$ and all s. Assumption A2.3 a) implies the mild requirement of conditional unbiasedness, which together with b) implies consistency of the covariance proxy respectively.

Proposition 1 *Under Assumptions A2.1 to A2.3 a)*

i) if $\frac{\partial^3 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t' \partial h_{k,t}} = 0$ $\forall k$, *then* $H_t^{*(s)} = \Sigma_t$ $\forall s$.

Under Assumptions A2.1 to A2.3

ii) if $\frac{\partial^3 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t' \partial h_{k,t}} \neq 0$ *for some* k, *then* $H_t^{*(s)} \xrightarrow{p} \Sigma_t$ *as* $s \to \infty$.

The proof is given in the appendix. Point *i)* recovers Hansen and Lunde's (2006a) result and identifies loss functions ensuring consistency of the ordering regardless of the quality of the proxy. It is worth noting that, although in order to ensure (2) we require the volatility proxy to be conditionally unbiased, this is not the case for the approximated loss, $L(\hat{\Sigma}_t, H_t)$. We can illustrate this point by considering the second order Taylor expansion of $L(\hat{\Sigma}_t, H_t)$ around the true value Σ_t:

$$L(\hat{\Sigma}_t^{(s)}, H_t) \cong L(\Sigma_t, H_t) + \left(\frac{\partial L(\Sigma_t, H_t)}{\partial \sigma_t}\right)' (\hat{\sigma}_t^{(s)} - \sigma_t) + \frac{1}{2}\left[(\hat{\sigma}_t^{(s)} - \sigma_t)' \frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t'} (\hat{\sigma}_t^{(s)} - \sigma_t)\right].$$

Taking conditional expectations with respect to $\Im_{t-1}$ and recalling A2.1 and A2.3 we get

$$\mathrm{E}_{t-1}[L(\hat{\Sigma}_t^{(s)}, H_t)] = L(\Sigma_t, H_t) + \frac{1}{2}\left[\mathrm{E}_{t-1}\left(\xi_t^{(s)\prime} \frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t'} \xi_t^{(s)}\right)\right], \tag{4}$$

which under *i)* holds with equality because the last term depends on the accuracy of the volatility proxy but not on H_t. Hence, by the law of iterated expectations, for all s and any two models $H_{l,t}$,$H_{m,t} \in \dot{H}$

$$\mathrm{E}[L(\hat{\Sigma}_t^{(s)}, H_{l,t})] - \mathrm{E}[L(\hat{\Sigma}_t^{(s)}, H_{m,t})] = \mathrm{E}[L(\Sigma_t, H_{l,t})] - \mathrm{E}[L(\Sigma_t, H_{m,t})].$$

Thus, conditional unbiasedness of $L(\hat{\Sigma}_t^{(s)}, H_{l,t})$ is not required, but it suffices that the bias is constant across models to ensure an ordering over $\dot{H}$ asymptotically invariant to the noise in the proxy. Point *ii)* of Proposition 1, shows that, if $L(.,.)$ is such that the bias term in (4)

depends on H_t, the distortion introduced in the ordering disappears and consistency of the ordering is recovered as long as the quality of the proxy improves, i.e., $s \to \infty$. This result becomes particularly relevant when ordering over a discrete set of models. Indeed, when the variance of the proxy is small with respect to the loss differential between any pair of models, the distortion induced by the variability of the proxy becomes negligible, leaving the ordering unaffected. This is important since commonly used loss functions although having other desirable properties, e.g., down-weighting extreme forecast errors or penalizing positive errors more than negative errors, do not satisfy *i)*. This is the case, for instance, for loss functions based on proportional forecast error or functional transformations of forecasts and realizations. Examples are: the entrywise 1-norm, $||\Sigma_t - H_t||_1$, the proportional Frobenius distance, $Tr(\Sigma_t H_t^{-1} - I)^2$ and log-Frobenius distances, $\left(\log \left|\Sigma_t H_t^{-1}\right|\right)^2$ or $\left(\log \frac{Tr[\Sigma_t \Sigma_t]}{Tr[H_t H_t]}\right)^2$. We denote a loss function that satisfies (violates) *i)* as "consistent" ("inconsistent").

An illustration of the results stated in Proposition 1 based on artificial data can be found in the online discussion paper. In a realistic setting, we observe that when the realized variance of Andersen, Bollerslev, Diebold, and Labys (2003) is used as a proxy, inconsistent loss functions preserve the ranking of one-step-ahead volatility forecasts only if the proxy is sufficiently accurate, i.e., when it is computed using data sampled at a frequency of 30 minutes or higher. When the quality of the proxy deteriorates, while the consistent loss functions rank properly, the ranking implied by the inconsistent loss functions appears heavily biased. However, this result is only illustrative and, as mentioned above, the minimum rank preserving sampling frequency depends on the choice of the loss function and ultimately on the degree of similarity between models' performances. Indeed, we also observe that, as loss differentials widen, which is the case when the forecast horizon increases (5 to 20-steps ahead), distortions in the ranking induced by inconsistent loss functions become less noticeable and only marginally affect the ranking even when the evaluation is based on a very noisy proxy. Similar evidence is discussed in the empirical application in Section 4.

3.2 Functional form of the consistent loss function

In this section, we propose a generalized necessary and sufficient functional form for a class of non-metric distance measures suited to vector and matrix spaces which ensure consistency of the ordering. These functions represent a particular instance of the class of Bregman

divergences, see Bregman (1967) and are related to the class of linear exponential densities of Gourieroux, Monfort, and Trognon (1984). To this regard, it can been shown that there is a unique Bregman divergence associated with every member of the exponential family of density functions, see Banerjee, Merugu, Dhillon, and Ghosh (2005). Our results generalize the work of Patton (2011) and Patton and Sheppard (2009).

In order to proceed, we need the following assumptions:

A3.1 $\hat{\Sigma}_t|\Im_{t-1} \sim F_t \in F$ the set of absolutely continuous distribution functions of $\mathbb{R}_{++}^{N\times N}$.

A3.2 $\exists H_t^* \in int(\dot{H})$ such that $H_t^* = \mathrm{E}_{t-1}(\hat{\Sigma}_t)$.

A3.3 $\mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right] < \infty$ for some $H \in \dot{H}$, $\left|\mathrm{E}_{t-1}\left[\frac{\partial L(\hat{\Sigma}_t,H_t)}{\partial h_t}\Big|_{H_t=\Sigma_t}\right]\right| < \infty$ and $\left|\mathrm{E}_{t-1}\left[\frac{\partial L(\hat{\Sigma}_t,H_t)}{\partial h_t \partial h_t'}\Big|_{H_t=\Sigma_t}\right]\right| < \infty \ \forall t$ where the last two inequalities hold elementwise.

Note that A3.2 follows directly from A1.2 and A2.3 because $H_t^* \in int(\dot{H})$ implies $H_t^* = \Sigma_t$ by A1.2 while $\mathrm{E}_{t-1}(\hat{\Sigma}_t) = \Sigma_t$ results from A2.3. Assumption A3.3 allows to interchange differentiation and expectation, see L'Ecuyer (1990) and L'Ecuyer (1995) for details.

Proposition 2 *Under Assumptions A2.1, A2.3 and A3.1 to A3.3, $L(.,.)$ is consistent if and only if it is a Bregman-type distance function, that is*

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)' vech(\hat{\Sigma}_t - H_t), \tag{5}$$

where $\tilde{C}(\cdot)$ is a scalar valued function from the space of $N \times N$ positive definite matrices to $\mathbb{R}$, three times continuously differentiable with

$$\begin{aligned} C(H_t) &= \nabla\tilde{C}(H_t) = \begin{bmatrix} \frac{\partial\tilde{C}(H_t)}{\partial h_{1,t}} \\ \vdots \\ \frac{\partial\tilde{C}(H_t)}{\partial h_{K,t}} \end{bmatrix} \\ C'(H_t) &= \nabla^2\tilde{C}(H_t) = \begin{bmatrix} \frac{\partial\tilde{C}(H_t)}{\partial h_{1,t}\partial h_{1,t}} & \cdots & \frac{\partial\tilde{C}(H_t)}{\partial h_{1,t}\partial h_{K,t}} \\ \vdots & \ddots & \\ \frac{\partial\tilde{C}(H_t)}{\partial h_{K,t}\partial h_{1,t}} & & \frac{\partial\tilde{C}(H_t)}{\partial h_{K,t}\partial h_{K,t}} \end{bmatrix}, \end{aligned}$$

where $C(\cdot)$ and $C'(\cdot)$ are the gradient and the Hessian of $\tilde{C}(\cdot)$ with respect to the $K = N(N+1)/2$ unique elements of H_t and $C'(H_t)$ is negative definite.

The proof is given in the appendix. Intuitively (5) can be thought of as the difference between the value of $\tilde{C}(\cdot)$ evaluated at H_t and the value of the first-order Taylor expansion of $\tilde{C}(\cdot)$ around point $\hat{\Sigma}_t$ evaluated at H_t. An alternative expression for the loss function defined in Proposition 2 is provided in the following corollary.

Corollary 1 *The family of Bregman functions defined in (5) is isometric to*

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + Tr[\bar{C}(H_t)(\hat{\Sigma}_t - H_t)], \tag{6}$$

with $\tilde{C}(\cdot)$ *defined as in Proposition 2 and*

$$\bar{C}(H_t) = \begin{bmatrix} \frac{\partial \tilde{C}(H)}{\partial h_{1,1,t}} & \frac{1}{2}\frac{\partial \tilde{C}(H)}{\partial h_{1,2,t}} & \cdots & \frac{1}{2}\frac{\partial \tilde{C}(H)}{\partial h_{1,N,t}} \\ \frac{1}{2}\frac{\partial \tilde{C}(H)}{\partial h_{1,2,t}} & \frac{\partial \tilde{C}(H)}{\partial h_{2,2,t}} & & \\ \vdots & & \ddots & \\ \frac{1}{2}\frac{\partial \tilde{C}(H)}{\partial h_{1,N,t}} & & & \frac{\partial \tilde{C}(H)}{\partial h_{N,N,t}} \end{bmatrix},$$

where the derivatives are taken with respect to all N^2 *elements of* H_t*.*

The proof is provided in the appendix. As we illustrate next, the general functional form admits several well known distance functions on the real line and on vector and matrix spaces. In the 1-dimensional space, the class of consistent loss functions corresponds the set of divergence criteria identified by Patton (2011) for the evaluation of univariate volatility forecasts. In particular, the *squared error*, the *relative entropy* (Kullback-Leibler divergence/I-divergence) and *Itakura-Saito distance* (homogeneous of degree zero, one and two respectively) correspond to the objective functions of the Gaussian, Poisson/Multinomial and Exponential/Gamma distributions respectively.

In the space of symmetric positive definite matrices, canonical examples employing different concave functions $\tilde{C}(.)$ include:

a) Squared Frobenius distance: obtained by setting $\tilde{C}(H_t) = -\mathrm{Tr}(H_t^2)$. This function, of the squared error type, relates to the matrix Gaussian density. This loss function can be expressed as

$$L(\hat{\Sigma}_t, H_t) = \sum_{i=1}^{N} \lambda_i, \tag{7}$$

where λ_i are the eigenvalues of $(\hat{\Sigma}_t - H_t)^2$.

b) von Neumann divergence: obtained by setting $\tilde{C}(H_t) = -\mathrm{Tr}(H_t \log(H_t) - H_t)$, where $\log(H_t)$ is the matrix logarithm (principal logarithm) of H_t. The distance function can be written as

$$L(\hat{\Sigma}_t, H_t) = \sum_{i,j=1}^{N} (v_i' u_j)^2 \alpha_i \log(\beta_j) - \sum_{i=1}^{N} \alpha_i \log(\alpha_i) + \sum_{i=1}^{N} (\alpha_i - \beta_i), \tag{8}$$

where u_i and v_i are the i-th column eigenvectors of H_t and $\hat{\Sigma}_t$ respectively and α_i and β_i are their corresponding i-th eigenvalues.

c) Stein loss: (also known as Burg divergence or LogDet divergence) is given by $\tilde{C}(H_t) = \log(\left|H_t^2\right|)$, i.e., the Burg entropy of the eigenvalues of H_t. It corresponds to the scale invariant loss function introduced by James and Stein (1961). It can be obtained as the objective function of the Wishart distribution. It is asymmetric, i.e., under-predictions (in matrix sense) are overweighed. When expressed using the eigendecomposition presented above, this loss function takes the form

$$L(\hat{\Sigma}_t, H_t) = \sum_{i,j=1}^{N} \frac{\beta_i}{\alpha_j}(v_i' u_j)^2 - \sum_{i=1}^{N} \frac{\beta_i}{\alpha_i}. \tag{9}$$

In the univariate case, Patton (2011) provides parametric expressions for the entire class of consistent loss functions. In higher dimensional spaces such a generalization is unfeasible because there are many functions $\tilde{C}(.)$ that can be used to weight forecasts and forecasts errors. Nevertheless, Proposition 3 identifies the entire subset of homogeneous loss functions based on forecast errors.

Proposition 3 *The family of consistent loss function based on the forecast errors $\hat{\Sigma}_t - H_t$ is defined by the quadratic form*

$$L(\hat{\Sigma}_t, H_t) = vech(\hat{\Sigma}_t - H_t)' \hat{\Lambda} vech(\hat{\Sigma}_t - H_t), \tag{10}$$

and the loss function has the following properties:

1) homogeneous of degree 2,

2) $\nabla^2 \tilde{C}(H_t) = -2\hat{\Lambda} = \Lambda$ is a matrix of constants defined according to Proposition 2,

3) $\hat{\Lambda}$ defines the weights assigned to the elements of the forecast error matrix $\hat{\Sigma}_t - H_t$,

4) symmetric under 180° *rotation about the origin.*

The proof is given in the appendix. Proposition 3 defines a family of quadratic loss functions, i.e. squared error type, which depends on the choice of the matrix of weights $\hat{\Lambda}$. Formally, the quadratic polynomial in (10) defines a family of quadric surfaces, i.e., elliptic paraboloids, and $\hat{\Lambda}$ defines the shape of the surface. In the univariate case, this loss function is symmetric, i.e., equally penalizes positive and negative forecast errors. An interesting feature of the multivariate case is that the notion of symmetry can be analyzed from different aspects, e.g. symmetry with respect to the origin, axes and planes depending on the particular choices of $\hat{\Lambda}$. It corresponds to the objective function of the multivariate Gaussian distribution and the matrix Λ plays a similar role as the correlation matrix, e.g., positive (negative) weights

$\lambda_{k,l}, k \neq l$ imply that systematic over/under predictions are penalized less (more). Different choices of $\hat{\Lambda}$ lead to:

a) Euclidean distance ($\hat{\Lambda} = I_K$): equally weights variances and covariances.

b) Squared weighted Euclidean distance ($\hat{\Lambda}$ defined as $\hat{\lambda}_{i,i} > 0$ and $\hat{\lambda}_{i,j} = 0$, $i, j = 1, ..., K$): allows to differently weight variances and covariances.

c) Mahalanobis distance ($\hat{\lambda}_{i,j} \in \mathbb{R}$, $i, j = 1, ..., K$): allows to penalize systematic over/under predictions.

d) Frobenius distance (L_F) (being a quadratic function it can be also represented as in (10) with $\hat{\Lambda} = \text{diag}(\text{vech}(V))$ (a diagonal matrix with $\text{vech}(V)$ on the main diagonal), $V : v_{ij} = 1$ if $i = j$, $v_{ij} = 2$ if $i \neq j$, $i, j = 1, ..., K$): assigns double weights to the covariance forecast errors (special case of weighted Euclidean distance).

4 Empirical application

4.1 Data description

The empirical application is based on the Euro, British Pound and the Japanese Yen exchange rates expressed in US dollars (denoted EUR, GBP and JPY respectively). The estimation sample ranges from January 6, 1987 to December 28, 2001 (3666 trading days), while the out-of-sample forecast evaluation sample runs until August 26, 2004 (660 trading days).

The proxy for the conditional variance matrix is the realized covariance estimator $\hat{\Sigma}_{t,\delta}$ of Andersen, Bollerslev, Diebold, and Labys (2003), i.e., the sum of outer products of intra-day returns, computed using data sampled at 17 different frequencies ranging from $\delta = 5$ minutes, i.e., 288 intervals/day (most accurate) to $\delta = 1$ day, i.e., 1 interval/day (least accurate).

A large number of parameterizations have been proposed to model conditional covariances, see Bauwens, Laurent, and Rombouts (2006), Silvennoinen and Terasvirta (2009) and Bollerslev (2010). In this application, the forecasting models set includes 24 specifications, as summarized in Table 1. We do not model the conditional mean since it turns out that, using standard tests, the return data do not exhibit any significant auto-correlation.

The models are estimated by quasi maximum likelihood using programs written by the authors using OxMetrics 6 (Doornik, 2007) and G@RCH 6 (Laurent, 2009). One and ten-step ahead forecasts are compared to the proxy $\hat{\Sigma}_{t,\delta}$ using the Frobenius distance, denoted L_F (consistent) and an absolute error type loss function, i.e., sum of elementwise absolute

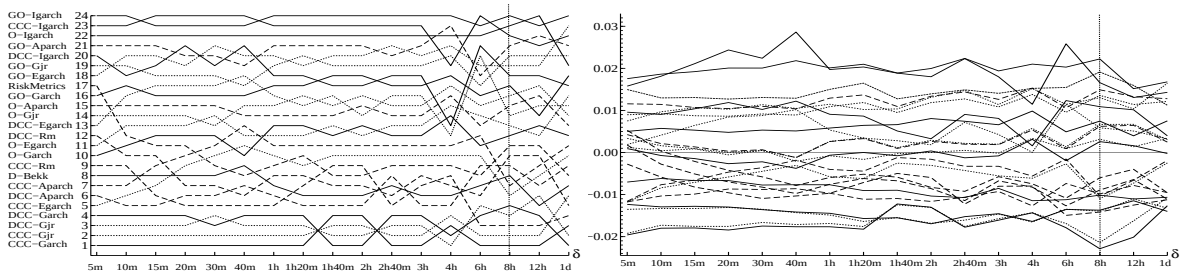
Table 1: Forecasting models set

Conditional correlation		
Correlation	Variance	Acronym
CCC (Bollerslev 1990)	Garch (Bollerslev, 1986)	CCC/DCC-Garch
DCC (Engle, 2002)	Asym. Power Garch (Ding, Granger, and Engle, 1993)	CCC/DCC-Aparch
	Exponential Garch (Nelson, 1991)	CCC/DCC-Egarch
	Gjr (Glosten, Jagannathan, and Runkle, 1992)	CCC/DCC-Gjr
	Integrated Garch (Engle and Bollerslev, 1986)	CCC/DCC-Igarch
	RiskMetrics (univ.) (J.P.Morgan, 1996)	CCC/DCC-Rm
Orthogonal		
	Variance	Acronym
Orth. (Alexander, 2000)	Garch	O/GO-Garch
Generalized orth.	Asym. Power Garch	O/GO-Aparch
(van der Weide, 2002)	Exponential Garch	O/GO-Egarch
	Gjr	O/GO-Gjr
	Integrated Garch	O/GO-Igarch
Linear		
	Variance	Acronym
diag. BEKK (Engle and Kroner, 1995)		D-Bekk
RiskMetrics (multiv.) (J.P.Morgan, 1996)		RM

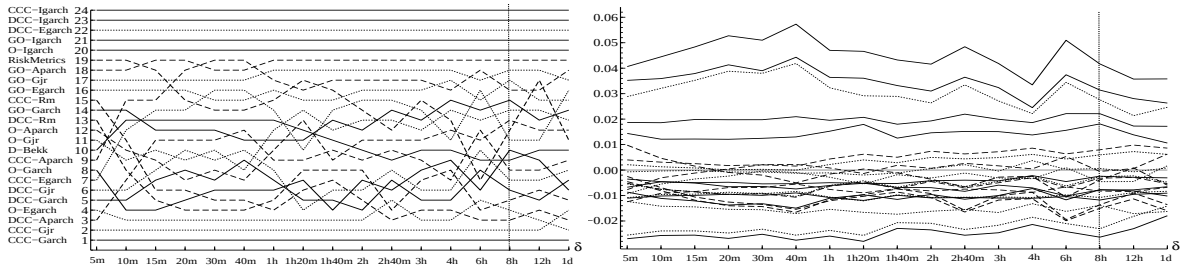
forecast errors, denoted L_{1M} (inconsistent). Note that other volatility proxies can be used instead, examples are multivariate realized kernels, see Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b), Hansen and Lunde (2006b) and Zhou (1996), or the range based covariance estimators of Brandt and Diebold (2006).

4.2 Model comparison

The empirical ranking of the 24 MGARCH models, as a function of the level of aggregation of the data used to compute $\hat{\Sigma}_{t,\delta}$, is reported in Figures 1 and 2 for the two loss functions. The vertical line at $\delta = 8$ hours denotes the lowest sampling frequency that ensures positive definiteness of $\hat{\Sigma}_{t,\delta}$. With respect to the one step ahead forecast evaluation, the consistent loss function, see Figure 1(a-left), points to the CCC-Garch as the best forecasting model at almost all frequencies. More generally, the subset given by the CCC and the DCC, both with Garch and Gjr variances, clearly outperforms all the other models. These models exhibit particularly stable and relatively close sample performances (Figure 1(a-right)). This result is not surprising being the two pairs of models nested and since we do not find any evidence of leverage effect. More generally, for all the models which capture the asymmetric impact



a) 1-step



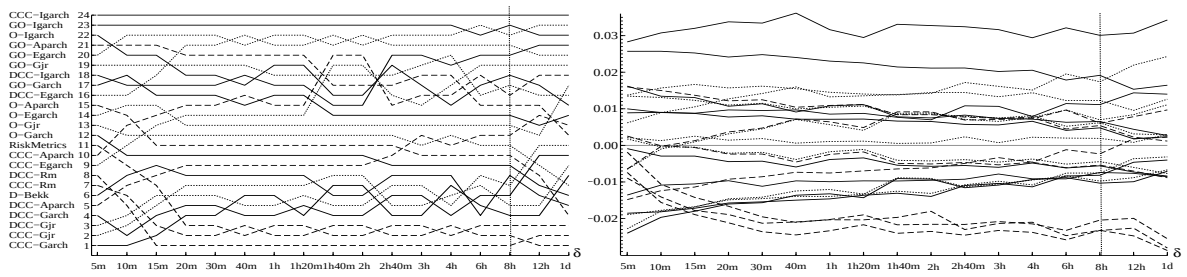
b) 10-step

Figure 1: Ranking implied by L_F (consistent). Ranking based on sample performances (left) and loss differentials from average between models (right).

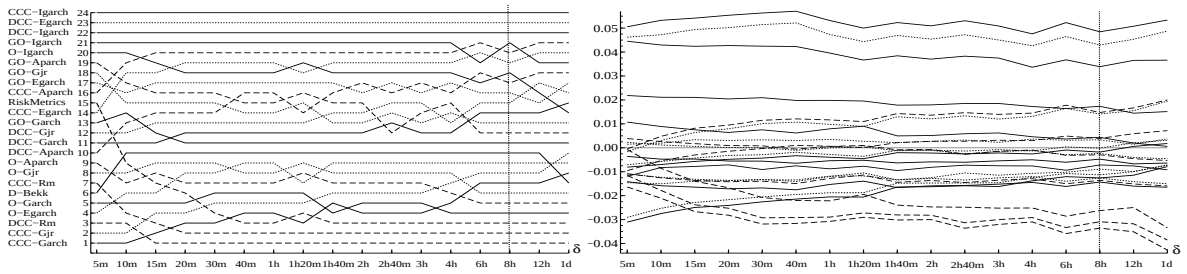
of shocks to the variance the null of symmetric variances cannot be rejected at standard significance levels on the basis of the estimated parameters. The worst performing models are the ones allowing for non-stationarity, with the exception of the three specifications based on the RiskMetrics approach which rank in the middle of the classification. Although the overall ranking is well preserved across all frequencies, it appears particularly stable when $\hat{\Sigma}_{t,\delta}$ is computed using 5-minute to 1-hour returns. As the quality of the proxy deteriorates, the ranking becomes more volatile. This is because the loss function becomes less informative, thus making it more difficult to effectively order models' performances.

As the forecast horizon increases (Figure 1(b)) model performances tend to cluster and loss differentials between clusters broaden. Also models' performances are more stable than in the 1-step ahead forecast horizon (Figure 1(b-right)) because long horizons forecasts are typically smoother. However the clustering induces a large variability of the ranking in the middle of the classification (Figure 1(b-left)).

Figure 2 illustrates to what extent the presence of the objective bias can affect the ranking when using an inconsistent loss function. As soon as the quality of the proxy deteriorates, the



a) 1-step



b) 10-step

Figure 2: Ranking implied by L_{1M} (inconsistent). Ranking based on sample performances (left) and loss differentials from average between models (right).

three RiskMetrics type models, characterized by a dynamic in the variance imposed ex-ante and independent from the data, steadily improve with respect to the other models. When the proxy is computed using data sampled every 2 hours or less, the three Riskmetrics models show by far the best forecasting performances (Figure 2(a-right)). Clearly, conclusions based on such evidence would be fallacious because this result is only due to the deterioration of the quality of the proxy. Similar conclusions hold when we consider the 10-step ahead forecast horizon (Figure 2(b)).

4.3 Model confidence set

The model confidence set (MCS) test of Hansen, Lunde, and Nason (2011) is a procedure that allows to identify a subset of superior models (in terms of predictive ability) containing the best one at a given level of confidence. Because the selection of the superior models by the MCS approach depends on the orderings implied by a loss function (e.g., the ranking given in Figures 1(a) and 2(a) - left panels), an incorrect choice of the loss function, although not invalidating per se the test, may result in an incorrect identification of the set of superior

models. Table 2 reports the MCS obtained using L_F and L_{1M}, with respect to five volatility proxies $\hat{\Sigma}_{t,\delta}$ (δ=5m, 20m, 1h20m, 2h40m, 8h) and two forecast horizon (1 and 10 days).

The MCSs with a confidence level $\alpha = 0.1$ are reported in Table 2. The MCS test is implemented in the free Ox software package MULCOM of Hansen and Lunde (2010).

Under the consistent loss function, L_F, the sets of superior models appear to be consistent across sampling frequencies (δ). Consistency of the ranking implies that the set of superior

Table 2: Model Confidence Set

1-step ahead forecast horizon

Frobenius distance (consistent)					Entrywise-1 norm (inconsistent)				
δ=5m	δ=20m	δ=1h20m	δ=2h40m	δ=8h	δ=5m	δ=20m	δ=1h20m	δ=2h40m	δ=8h
CCC-Garch	CCC-Garch	CCC-Garch	CCC-Garch	CCC-Garch		RiskMetrics	RiskMetrics	RiskMetrics	RiskMetrics
CCC-Gjr	CCC-Gjr	CCC-Gjr	CCC-Gjr	CCC-Gjr		DCC-Rm	DCC-Rm	DCC-Rm	DCC-Rm
	DCC-Garch	DCC-Garch	DCC-Garch	DCC-Garch		CCC-Rm	CCC-Rm	CCC-Rm	CCC-Rm
	DCC-Gjr	DCC-Gjr	DCC-Gjr	DCC-Gjr	CCC-Garch	CCC-Garch	CCC-Garch		
		DCC-Aparch	DCC-Aparch	DCC-Aparch	CCC-Gjr	CCC-Gjr	CCC-Gjr		
		D-Bekk	D-Bekk	D-Bekk	DCC-Aparch	DCC-Aparch			
			CCC-Egarch	CCC-Egarch	DCC-Garch	DCC-Garch			
			CCC-Rm	CCC-Rm	DCC-Gjr	DCC-Gjr			
			DCC-Rm	DCC-Rm	D-Bekk				
			RiskMetrics	RiskMetrics					
			O-Garch	O-Egarch					
				CCC-Aparch					

10-step ahead forecast horizon

Frobenius distance (consistent)					Entrywise-1 norm (inconsistent)				
δ=5m	δ=20m	δ=1h20m	δ=2h40m	δ=8h	δ=5m	δ=20m	δ=1h20m	δ=2h40m	δ=8h
CCC-Garch	CCC-Garch	CCC-Garch	CCC-Garch	CCC-Garch		DCC-Rm	DCC-Rm	DCC-Rm	DCC-Rm
		CCC-Gjr	CCC-Gjr	CCC-Gjr		CCC-Rm	CCC-Rm		
		O-Egarch	O-Egarch	O-Egarch	CCC-Garch	CCC-Garch	CCC-Garch		
			O-Garch	O-Garch		CCC-Gjr			
			O-Aparch	O-Aparch		O-Garch			
			DCC-Garch	DCC-Garch		RiskMetrics			
			DCC-Gjr	DCC-Gjr					
			D-Bekk	D-Bekk					
			CCC-Rm	CCC-Rm					
			DCC-Rm	DCC-Rm					
			RiskMetrics	RiskMetrics					
				CCC-Egarch					
				CCC-Aparch					
				DCC-Aparch					
				GO-Garch					
				GO-Egarch					
				O-Gjr					

Notes. The initial set contains 24 models. Test statistics T_D (deviation from average between models). Significance level $\alpha = 0.1$. Sample size=650 obs. Standard errors based on 10,000 bootstrap resamples.

models identified using a high precision proxy is always included in the set obtained using a less accurate proxy. The loss of efficiency of the proxy translates into a higher variability

of the models sample evaluation making it more difficult to discriminate between them. For instance, the MCS obtained using the proxy based on 8-hour returns contains one half (1-day horizon) and two thirds (10-day horizon) of the 24 candidate models. When $\delta = 5$ and 20 minutes, we obtain the most accurate sets which is in line with the well known result that volatility proxies based on data sampled between 5 and 20 minutes strike the best compromise between loss of accuracy and noise due to microstructure frictions, see Andersen, Bollerslev, Diebold, and Labys (1999) and Russell and Bandi (2004). These results clearly demonstrate the value of high precision proxies. Although consistency of the ordering is ensured by an appropriate choice of the loss function independently of the quality of the proxy, a high precision proxy allows to efficiently discriminate between models.

Results based on the inconsistent L_{1M} reflect the presence of a distorted outcome. For both forecast horizons, as the quality of the proxy deteriorates, the MCS changes in composition and reduces in size. The sets obtained using proxies based on 5-minute and 8-hour returns do not share common elements.

With respect to the overall composition of the MCS, we find that simple models like the CCC-GARCH seem to be sufficiently accurate, relative to more sophisticated specifications, in capturing the volatility dynamics of the data. There are several reasons for this. First, the period considered in this paper, i.e. December 2001 to August 2004, is characterized by a relatively small and slow-moving volatility. Therefore, more heavily parameterized sophisticated models suffer from additional parameter uncertainty. In fact, Giacomini and White (2006) suggest that, when the process is characterized by simple dynamics, more parsimonious models (even if misspecified) may outperform more flexible specifications especially in presence of high estimation uncertainty. Second, there is no evidence of asymmetric response of volatility to shocks within the set of models under consideration. Indeed, for all models with parameters that allow for asymmetry, the null of no symmetry cannot be rejected at standard significance levels. This result is in line with most applications on exchange rate returns, see Diebold and Nerlove (1989), Andersen, Bollerslev, Diebold, and Labys (2001) and Hansen and Lunde (2005) for some examples. Third, while DCC models often produce more realistic time-varying conditional correlation forecasts, these forecasts are often severely biased which results in relatively larger contributions to the loss functions considered in this application.

Finally, we find that models imposing non-stationarity are systematically lower ranked. The same holds for models that decompose the conditional variance matrix in terms of unobserved factors, i.e., orthogonal models. In a similar context, Laurent, Rombouts, and Violante (2010) suggest that, although DCC and orthogonal models with leverage effect and long memory may outperform less sophisticated models during periods of market instability, simple assumptions like constant correlation and symmetric response of volatility to shocks cannot be rejected in periods of calm and upward trending markets.

5 Conclusion

In this paper, we cast to the multivariate dimension the sufficient conditions that a loss function has to satisfy to deliver the same ordering whether the evaluation is based on the true conditional variance matrix or an unbiased proxy of it. We show that when the proxy is sufficiently accurate with respect to the degree of similarity between models' performances, inconsistent loss functions can still deliver an unbiased ranking. We propose a generalized necessary and sufficient functional form for a class of non-metric distance measures suited to vector and matrix spaces which ensure consistency of the ordering. The general functional form represents a particular instance of the class of Bregman divergences. We introduce a range of suitable vector and matrix Bregman-type functions, such as the Kullback-Leibler divergence, the Itakura-Saito distance, the Frobenius distance, the von Neumann divergence and the Stein distance. We also identify the entire subset of loss functions based on forecast errors, i.e., the difference between forecasts and observations. The application to three foreign exchange rates illustrates, in an out-of-sample forecast comparison among 24 multivariate GARCH models, the robustness of the ordering under a consistent loss function and the importance of high precision proxies for model selection. We also study to what extent the ranking and the MCS test are affected when we combine a noisy proxy with an inconsistent loss function.

There are several extensions for future research. First, this paper ranks multivariate volatility models based on statistical loss functions and focuses on conditions for consistent ranking from a theoretical viewpoint. At some point an economic loss function might be introduced when the forecasted volatility matrices are actually used in financial applications such as portfolio management and option pricing. It is clear that the model with the smallest

statistical loss is always preferred but it may happen that other models with small statistical losses become indistinguishable in terms of economic loss. This issue has not been addressed in this paper. Second, from an applied viewpoint, the behavior of the ranking when using proxies other than realized covariance should be further investigated.

Appendix: Proofs

Proof of Proposition 1. Under Assumptions A2.1 to A2.3, the first order conditions of the minimization problem in (3), recalling the expansion in (4), are

$$\begin{aligned}
\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t^{(s)}, H_t)\right]}{\partial h_{k,t}} - \frac{\partial L(\Sigma_t, H_t)}{\partial h_{k,t}} &\cong \frac{1}{2}\left[\frac{\partial}{\partial h_{k,t}}\mathrm{E}_{t-1}\left(\xi_t^{(s)\prime}\Psi(\sigma_t^2, h_t)\xi_t^{(s)}\right)\right] \\
&\cong \frac{1}{2}\frac{\partial}{\partial h_{k,t}}\mathrm{E}_{t-1}\left[\sum_{l,m}\xi_{l,t}^{(s)}\xi_{m,t}^{(s)}\Psi(\sigma_t^2, h_t)_{l,m}\right] \\
&\cong \frac{1}{2}\sum_{l,m}\frac{\partial \Psi(\sigma_t^2, h_t)_{lm}}{\partial h_{k,t}}\mathrm{E}_{t-1}[\xi_{l,t}^{(s)}\xi_{m,t}^{(s)}] \\
&\cong \frac{1}{2}\sum_{l,m}\frac{\partial \Psi(\sigma_t^2, h_t)_{lm}}{\partial h_{k,t}}V_{l,m,t}^{(s)}
\end{aligned}$$

for all s, with $l, m = 1, ..., N(N+1)/2$, $k = 1, ..., N(N+1)/2$ and where $V_{l,m,t}^{(s)} = \mathrm{E}_{t-1}[\xi_{l,t}^{(s)}\xi_{m,t}^{(s)}]$ and $\Psi(\sigma_t^2, h_t)_{l,m}$ represent respectively the element $[l, m]$ of the variance matrix of the proxy $V_t^{(s)} = \mathrm{E}_{t-1}[\xi_t^{(s)}\xi_t^{(s)\prime}]$ and of $\Psi(\sigma_t^2, h_t)$, the matrix of second derivatives of $L(.,.)$ with respect to σ_t^2.

The first order conditions imply that $H_t^{*(s)}$ is the solution of

$$\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t^{(s)}, H_t^{*(s)})\right]}{\partial h_{k,t}} = 0 \ \forall k$$

and A1.1 ensures that second order conditions are satisfied. Then, we have that

$$-\frac{\partial L(\Sigma_t, H_t^{*(s)})}{\partial h_{k,t}} \cong \frac{1}{2}\sum_{l,m}\frac{\partial \Psi(\sigma_t^2, .)_{lm}}{\partial h_{k,t}}V_{l,m,t}^{(s)}. \tag{11}$$

Under i), i.e., $\frac{\partial \Psi(\sigma_t^2, .)_{lm}}{\partial h_{k,t}} = 0 \ \forall k$, the first order conditions of the loss function based on the proxy lead to the same optimal forecast as if the true variance matrix was observable, even in presence of a noisy volatility proxy. From A1.2 it follows that

$$\frac{\partial L(\Sigma_t, H_t^{*(s)})}{\partial h_{k,t}} = 0 \ \ \forall k \Leftrightarrow H_t^{*(s)} = \Sigma_t \ \ \forall s,$$

that is the identification of the optimal forecast is not affected by the presence of noise in the proxy. Since the optimal forecast equals the conditional variance, by Assumption A1.2, A2.1 and A2.3a, we also have that $H_t^{*(s)} = H_t^* = \Sigma_t = \mathrm{E}_{t-1}(\hat{\Sigma}_t)$.

Under ii), i.e., $\frac{\partial \Psi(\sigma_t^2, h_t)_{lm}}{\partial h_{k,t}} \neq 0$ for some k, then as $s \to \infty$, by A2.3b and (11) we have

$$\frac{\partial L(\Sigma_t, H_t^{*(s)})}{\partial h_{k,t}} \xrightarrow{p} 0 \quad \forall k \Leftrightarrow H_t^{*(s)} \xrightarrow{p} \Sigma_t,$$

which concludes the proof. ■

Proof of Proposition 2. To prove the proposition, we proceed as in Patton (2011). We show the equivalence of the following statements:

-S1: the loss function is a Bregman divergence of the form given in the proposition;

-S2: the loss function is consistent in the sense of (2);

-S3: the optimal forecast under the loss function is the conditional variance matrix.

Step 1: S1⇒S2. The result follows directly form Proposition 1, in fact:

$$\frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t'} = \nabla^2 \tilde{C}(\Sigma_t) = \Psi(\sigma_t^2, .)$$

since $\frac{\partial^2 (C(H_t)'\sigma_t)}{\partial \sigma_t \partial \sigma_t'} = 0$, and does not depend on H_t.

Step 2: S2⇒S3. By Assumption A3.2, H_t^* in the interior of the support of $L(\hat{\Sigma}_t, H_t)$, hence *Step 1* implies that $H_t^* = \mathrm{E}_{t-1}(\hat{\Sigma}_t)$. Thus $\forall H_t \in int(\dot{H}) \setminus \{H_t^*\}$:

$$\mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t^*)\right] \quad \leq \quad \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]$$

and by the law of iterated expectations:

$$\mathrm{E}\left[L(\hat{\Sigma}_t, H_t^*)\right] \quad \leq \quad \mathrm{E}\left[L(\hat{\Sigma}_t, H_t)\right].$$

Then we can write (2) as

$$\mathrm{E}(L(\hat{\Sigma}_t, H_t^*)) \leq \mathrm{E}(L(\hat{\Sigma}_t, H_t)) \Leftrightarrow \mathrm{E}(L(\Sigma_t, H_t^*)) \leq \mathrm{E}(L(\Sigma_t, H_t)).$$

Setting $H_t = \Sigma_t$, by Assumptions A1.1 and A1.2, $\mathrm{E}(L(\Sigma_t, \Sigma_t)) = 0 \Rightarrow \mathrm{E}(L(\Sigma_t, H_t^*)) = 0$ and therefore $H_t^* = \Sigma_t$.

Step 3: S1⇔S3. The last step uses the arguments of Gourieroux and Monfort (1995), which prove sufficiency and necessity of the linear exponential functional form for the pseudo true density to prove consistency of the pseudo maximum likelihood estimator.

First, we prove sufficiency (S1⇒S3). Consider the first order conditions evaluated at the optimum ($H_t = H_t^*$), that is

$$\begin{aligned}\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]}{\partial h_t} &= C(H_t^*) + \nabla^2\tilde{C}(H_t)\mathrm{vech}(\mathrm{E}_{t-1}(\hat{\Sigma}_t) - H_t^*) - C(H_t^*) = 0 \\ &= \nabla^2\tilde{C}(H_t)\mathrm{vech}(\mathrm{E}_{t-1}(\hat{\Sigma}_t) - H_t^*) = 0 \\ &\Leftrightarrow \mathrm{E}_{t-1}(\hat{\Sigma}_t) = H_t^*.\end{aligned}$$

Second, to prove necessity (S3⇒S1), consider that at the optimum we must have $\mathrm{E}_{t-1}(\hat{\Sigma}_t) = H_t^*$, and consequently

$$\mathrm{E}_{t-1}\left(\frac{\partial L(\hat{\Sigma}_t, H_t^*)}{\partial h_t}\right) = 0,$$

for any conditional distribution $F_t \in F$.

Applying Lemma 8.1 in Gourieroux and Monfort (1995) page 240, there exists a square matrix Λ of size $k = N(N+1)/2$ which is only function of H_t^* such that

$$\frac{\partial L(\hat{\Sigma}_t, H_t^*)}{\partial h_t} = \Lambda(H_t^*)\mathrm{vech}(\hat{\Sigma}_t - H_t^*). \tag{12}$$

Since we want to ensure that H_t^* is the minimizer of $L(\hat{\Sigma}_t, H_t^*)$ then we must have $\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]}{\partial h_t \partial h_t'}$ satisfying second order necessary or sufficient conditions. Using Assumption A3.3 we can interchange differentiation and expectation (see L'Ecuyer (1990) and L'Ecuyer (1995) for details) to obtain

$$\begin{aligned}\mathrm{E}_{t-1}\left(\frac{\partial L(\hat{\Sigma}_t, H_t^*)}{\partial h_t \partial h_t'}\right) &= \mathrm{E}_{t-1}\left(\frac{\partial \Lambda(H_t^*)\mathrm{vech}(\hat{\Sigma}_t - H_t^*)}{\partial h_t}\right) \\ &= \mathrm{E}_{t-1}\left(\begin{bmatrix} \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{1i}}{\partial h_1}(\sigma_i - h_i^*) & \dots & \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{1i}}{\partial h_k}(\sigma_i - h_i^*) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{ki}}{\partial h_1}(\sigma_i - h_i^*) & \dots & \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{ki}}{\partial h_k}(\sigma_i - h_i^*) \end{bmatrix}\right) - \Lambda(H_t^*) \\ &= -\Lambda(H_t^*),\end{aligned}$$

with $K = N(N+1)/2$. $\Lambda(H_t^*)$ is positive definite which ensures that the necessary condition for the minimum is satisfied.

Now, it suffices to integrate (12) (up to a constant and/or a term that solely depends on $\hat{\Sigma}_t$) to recover the loss function of the form stated in the proposition. In fact, if we define

$$\Lambda(H_t) = \nabla^2\tilde{C}(H_t) = C'(H_t),$$

and rewrite (12) as

$$C'(H_t)\text{vech}(\hat{\Sigma}_t) - C'(H_t)\text{vech}(H_t),$$

we have that

$$\begin{aligned} C'(H_t)\text{vech}(\hat{\Sigma}_t) &= \frac{\partial C(H_t)'\text{vech}(\hat{\Sigma}_t)}{\partial h_t} \\ C'(H_t)\text{vech}(H_t) &= \frac{\partial C(H_t)'\text{vech}(H_t)}{\partial h_t} - C(H_t) \\ &= \frac{\partial C(H_t)'\text{vech}(H_t)}{\partial h_t} - \frac{\partial \tilde{C}(H_t)}{\partial h_t}. \end{aligned}$$

Therefore (12) admits as primitive

$$C(H_t)'\text{vech}(\hat{\Sigma}_t) - C(H_t)'\text{vech}(H_t) + \tilde{C}(H_t).$$

Rearranging and allowing for a term that depends on $\hat{\Sigma}_t$, we obtain

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

where $\frac{\partial \tilde{C}(\hat{\Sigma}_t)}{\partial h_t} = 0$, which concludes the proof. ■

Proof of Corollary 1. Since $\hat{\Sigma}_t$ and H_t are symmetric, then

$$\begin{aligned} \text{Tr}[\bar{C}(H_t)(\hat{\Sigma}_t - H_t)] &= \sum_i \bar{c}_{i,i}(H_t)(\hat{\sigma}_{i,i,t} - h_{i,i,t}) + 2\sum_{i<j} \bar{c}_{i,j}(H_t)(\hat{\sigma}_{i,j,t} - h_{i,j,t}) \quad i,j = 1,...,N \\ &= \sum_i \frac{\partial \tilde{C}(H_t)}{\partial h_{i,i,t}}(\hat{\sigma}_{i,i,t} - h_{i,i,t}) + 2\sum_{i<j} \frac{1}{2}\frac{\partial \tilde{C}(H_t)}{\partial h_{i,j,t}}(\hat{\sigma}_{i,j,t} - h_{i,j,t}) \\ &= C(H_t)'\text{vech}(\hat{\Sigma}_t - H_t), \end{aligned}$$

with $C(H_t)'$ as defined in Proposition 2. ■

Proof of Proposition 3. By Proposition 2, a consistent loss functions based on the forecast error must have the form

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)'\text{vech}(\hat{\Sigma}_t - H_t). \tag{13}$$

Consider

$$\begin{aligned} \frac{\partial L(\hat{\Sigma}_t, H_t)}{\partial h_t} &= \nabla^2 \tilde{C}(H_t)\text{vech}(\hat{\Sigma}_t - H_t) \\ \frac{\partial L(\hat{\Sigma}_t, H_t)}{\partial \sigma_t} &= C(H_t) - C(\hat{\Sigma}_t). \end{aligned}$$

Note that since the loss function is only based on the forecast error then it is symmetric under 180° rotation around the origin and, which implies

$$-\frac{\partial L(\hat{\Sigma}_t, H_t)}{\partial h_t} = \frac{\partial L(\hat{\Sigma}_t, H_t)}{\partial \sigma_t}, \tag{14}$$

and therefore

$$\nabla^2 \tilde{C}(H_t)\text{vech}(\hat{\Sigma}_t - H_t) = C(H_t) - C(\hat{\Sigma}_t),$$

for all $\hat{\Sigma}_t$ and H_t. Differentiating both sides of (14) with respect to σ_t we obtain

$$\nabla^2 \tilde{C}(H_t) = \nabla^2 \tilde{C}(\hat{\Sigma}_t),$$

which implies

$$\nabla^2 \tilde{C}(H_t) = \Lambda, \tag{15}$$

where Λ is a matrix of constants.

Equation (15) implies that $C(H_t) = \nabla^2 \tilde{C}(H_t)\text{vech}(H_t)$ is homogeneous of degree 1, and hence $\tilde{C}(\cdot)$ is homogeneous of degree 2 then so is $L(\hat{\Sigma}_t, H_t)$. Applying Euler theorem for homogeneous functions we have that $2\tilde{C}(H_t) = C(H_t)'\text{vech}(H_t)$. The loss function in (13) can be rewritten as

$$L(\hat{\Sigma}_t, H_t) = -\tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)'\text{vech}(\hat{\Sigma}_t). \tag{16}$$

In order to satisfy second order conditions Λ must be negative definite, according to Proposition 2. Since $L(\hat{\Sigma}_t, H_t)$ is homogeneous of degree 2, starting from (15), we can apply Euler theorem for homogeneous functions and obtain

$$\begin{aligned} C(H_t) &= \Lambda\text{vech}(H_t) \\ \tilde{C}(H_t) &= \frac{1}{2}\text{vech}(H_t)'\Lambda\text{vech}(H_t). \end{aligned}$$

Substituting the expression for $\tilde{C}(.)$ in (16) and rearranging we obtain the quadratic loss

$$\begin{aligned} L(\hat{\Sigma}_t, H_t) &= -\frac{1}{2}\text{vech}(\hat{\Sigma}_t - H_t)'\Lambda\text{vech}(\hat{\Sigma}_t - H_t) \\ &= \text{vech}(\hat{\Sigma}_t - H_t)'\hat{\Lambda}\text{vech}(\hat{\Sigma}_t - H_t), \end{aligned}$$

with $\hat{\Lambda} = -\frac{1}{2}\Lambda$. ■

References

Alexander, C. (2000): "Orthogonal Methods for Generating Large Positive Semi-Definite Covariance Matrices," Henley Business School, Reading University.

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, P. Christoffersen, and F. Diebold (2006): *Volatility and Correlation Forecasting* chap. 15, pp. 777–877, Handbook of Economic Forecasting. North Holland.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (1999): "(Understanding, Optimizing, Using and Forecasting) Realized Volatility and Correlation," Discussion paper, New York University, Leonard N. Stern School of Business.

——— (2001): "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.

——— (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.

Banerjee, A., S. Merugu, I. Dhillon, and J. Ghosh (2005): "Clustering with Bregman divergences," *Journal of Machine Learning Research*, 6, 17051749.

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.

——— (2008b): "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading," *DP, Oxford University.*

Bauwens, L., S. Laurent, and J. Rombouts (2006): "Multivariate GARCH Models: A Survey," *Journal of Applied Econometrics*, 21, 79–109.

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

——— (1990): "Modeling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH model," *Review of Economics and Statistics*, 72, 498–505.

——— (2010): "Glossary to ARCH (GARCH)," in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, ed. by T. Bollerslev, J. Russell, and M. Watson, pp. 137–163. Oxford University Press.

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.

Bregman, L. (1967): "The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming," *URSS Computational Mathematics and Physics*, 7, 200–217.

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

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

Diebold, F., and M. Nerlove (1989): "The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model," *Journal of Applied Econometrics*, pp. 1–21.

Ding, Z., C. W. J. Granger, and R. F. Engle (1993): "A Long Memory Property of Stock Market Returns and a New Model," *Journal of Empirical Finance*, 1, 83–106.

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

Elliott, G., and A. Timmermann (2008): "Economic Forecasting," *Journal of Economic Literature*, 46, 3–56.

Engle, R. (2002): "Dynamic Conditional Correlation - a Simple Class of Multivariate GARCH Models," *Journal of Business and Economic Statistics*, 20, 339–350.

Engle, R., and T. Bollerslev (1986): "Modelling the Persistence of Conditional Variances," *Econometric Reviews*, 5, 1–50.

Engle, R., and F. Kroner (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150.

Ferguson, T. (1967): *Mathematical Statistics - A Decision Theoretic Approach*. Academic Press.

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.

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.

——— (2006a): "Consistent Ranking of Volatility Models," *Journal of Econometrics*, 131, 97–121.

——— (2006b): "Realized Variance and Market Microstructure Noise," *Journal of Business and Economic Statistics*, 24, 127–218.

——— (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.

James, W., and C. Stein (1961): "Estimation with Quadratic Loss," *Proc. Fourth Berkley Symp. on Math. Statist. and Prob.*, 1, 361–379.

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

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

Laurent, S., J. Rombouts, and F. Violante (2010): "On the Forecasting Accuracy of Multivariate GARCH Models," CORE discussion paper 2010-25.

L'Ecuyer, P. (1990): "A Unified View of the IPA, SF and LR Gradient Estimation Techniques," *Management Science*, 36.

——— (1995): "On the Interchange of Derivative and Expectation for Likelihood Ratio Derivative Estimators," *Management Science*, 41.

Nelson, D. (1991): "Conditional Heteroskedasticity in Asset Returns: a New Approach," *Econometrica*, 59, 349–370.

Patton, A. (2011): "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 160.

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.

Russell, J., and F. Bandi (2004): "Microstructure Noise, Realized Volatility, and Optimal Sampling," Econometric Society 2004 Latin American Meetings 220, Econometric Society.

Silvennoinen, A., and T. Terasvirta (2009): "Multivariate GARCH Models," in *Handbook of Financial Time Series*, ed. by T. Andersen, R. Davis, J. Kreiss, and T. Mikosch. Springer.

van der Weide, R. (2002): "GO-GARCH: A Multivariate Generalized Orthogonal GARCH Model," *Journal of Applied Econometrics*, 17, 549–564.

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

White, H. (2000): "Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Zhou, B. (1996): "High-frequency Data and Volatility in Foreign Exchange Rates," *Journal of Business and Economic Statistics*, 14.