

COMPARING AND RANKING COVARIANCE STRUCTURES OF M-GARCH VOLATILITY MODELS

Sébastien Laurent¹, Jeroen V.K. Rombouts², Annastiina Silvennoinen³ and Francesco Violante⁴

October 4, 2006

Abstract

A large number of different parameterizations have been introduced to model conditional variance dynamics in a multivariate framework. A major problem with these models is the large number of parameters that have to be estimated and the many constraints, often difficult to make explicit, that have to be imposed to ensure semi-positive definiteness of the covariance matrices. This paper compares and ranks different multivariate GARCH type models in terms of their ability to estimate the in-sample conditional variance-covariance structure and the accuracy of their out-of-sample variance forecasts. The models are compared using Hansens Superior Predictive Ability test redefined in a multivariate framework by providing three alternative metrics to evaluate the distance between variance matrices. In this paper we also test the preliminary version of our MGARCH Package for Ox, extension of Laurent and Peterss G@RCH Package. Up to date, the package includes: BEKK models (Full, Diagonal and Scalar), CCC-DCC models and RiskMetrics.

Keywords: Volatility, Multivariate GARCH, Covariance Models, SPA test, Financial econometrics.

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

¹CeReFim, Université de Namur, and CORE, Université Catholique de Louvain.

²HEC Montréal, CIRANO, CIRPEE and CREF.

³Stockholm School of Economics.

⁴FUNDP Namur and CORE, Université Catholique de Louvain.

The authors would like to thank Kilian Mie for useful suggestions and comments.

Correspondence to Sébastien Laurent, FUNDP, Rempart de la Viérge, 8, B-5000 Namur, Belgium. Telephone: +32 81 72 48 69. Fax: +32 81 72 48 40. E-mail: Sebastien.Laurent@fundp.ac.be

or to Francesco Violante, FUNDP, Rempart de la Viérge, 8, B-5000 Namur, Belgium. Telephone: +32 81 72 48 10. Fax: +32 81 72 48 40. E-mail: violante@core.ucl.ac.be

Contents

1	Introduction	1
2	Multivariate GARCH: an Overview	2
2.1	BEKK-GARCH(1,1)	2
2.2	RiskMetrics	3
2.3	CCC-GARCH(1,1)	3
2.4	DCC(1,1)-GARCH(1,1)	3
3	Model Comparison: distance metrics	4
3.1	Frobenius Metric	4
3.2	Eigenvalue Metric	4
3.3	Foerstner and Moonen Metric	5
3.4	Cosinus Mass Metric	5
4	Realized Volatility	7
5	Model Comparison: Hansen SPA Test	8
6	Data and Empirical Results	10
7	Conclusions	13

1 Introduction

Since the first seminal paper of Kraft and Engle (1982) an increasing interest in modelling conditional covariances has developed a large number of multivariate volatility models, see Bawens, Laurent and Rombouts (2006) for an extensive survey.

In principle all these models suffer from the increase of the cross-sectional dimension due to the large number of parameters to be simultaneously estimated which increases at least quadratically¹, even when parameter restrictions, typically parameter pooling, are imposed. Yet a further complication rises from the necessity to impose non-linear parameter constraints to ensure semi-positive definite conditional covariance matrices.

The aim of this paper is to compare and rank, in terms of ability to estimate the sample covariance structure and forecast accuracy, some of the most well known and widely used multivariate GARCH models. The specifications considered include the BEKK (Engle and Kroner 1995), CCC (Bollerslev 1990), DCC (Engle 2001), RiskMetrics (JPMorgan 1996).

After checking the consistency of the procedure - i.e. estimation, forecast, comparison - using *ad hoc* simulated data we compare different model specifications using daily returns for two stocks traded on the NYSE for the period 1980 - 1999, while the out-of-sample realized volatility has been computed using 5-minute data from 1995 to 1999.

Models are evaluated using a loss function, which is applied both to the sample conditional covariance estimation and to the sequence of one step ahead covariance forecasts. The latent conditional covariance in the loss function is substituted, depending on the context, either by simulated data or by realized volatility.

The sequence of conditional covariance forecasts has been obtained by means of a moving window of dimension k within which at each iteration the one step ahead forecast has been computed replacing the expectations for the squared innovation with its observed value. The coefficients are updated every k observations by reestimating the model including the last k out of sample observations into the dataset.

Since the objective of this paper is to compare covariance matrices the choice of an appropriate loss function has to be compatible to the multivariate framework. In this paper we investigate the problem providing three different notions of distance between matrices. The goodness of fit of each model to the latent conditional covariance is then tested using Hansen (2001) Superior Predictive Ability Test (SPA). This test has the advantage that it properly accounts for the full set of models and is therefore likely to detect a superior model when such exist. In fact, SPA test evaluates whether a particular model (benchmark) when compared to the true data generating process (DGP) - simulated data or realized volatility - is significantly outperformed by other models, while taking into account the whole set of models that are being compared.

In this paper we also introduce the first version of a new OxMetrix application to estimate and forecast multivariate GARCH models. The objective is to extend the work done by Laurent

¹Engle and Sheppard (2005)

and Peters (2006) with their G@RCH package dedicated to univariate ARCH-type models. Other than including procedures to estimate and forecast BEKK, CCC, Engle's DCC and RiskMetrics specifications for the conditional covariance, it provides also some features to evaluate distance metrics, such as Frobenius norm, Eigenvalue metric (spectral norm), Foerstner and Moonen metric and Cosinus Mass metric.

The paper is organized as follows. Section 2 provide a brief overview of several GARCH specification considered in this paper, while the loss functions are described in Section 3. Section 4 and 5, contain some details on realized volatility and of the Hansen (2001) SPA test. Empirical results are presented in Section 6. Finally, Section 7 contains some concluding remarks.

2 Multivariate GARCH: an Overview

Consider a vector price process, p_t of dimension $(N \times 1)$, let define the compounded daily return by $r_t = \log(p_t) - \log(p_{t-1})$. The process r_t is a vector stochastic process conditioned on the sigma field, \mathfrak{S}_{t-1} generated by the information available at $t - 1$. We denote $\mu_t \equiv E(r_t | \mathfrak{S}_{t-1})$ the conditional mean vector and $H_t \equiv E(r_t' r_t | \mathfrak{S}_{t-1})$ the conditional variance matrix. The specification used for the mean equation is:

$$r_t = \mu_t + \epsilon_t \tag{1}$$

$$\epsilon_t = H_t^{1/2} z_t \tag{2}$$

where $H_t^{1/2}$ is a $(N \times N)$ positive definite matrix and z_t is i.i.d random vector with $E(z_t) = 0$ and $Var(z_t) = I_N$.

In this paper we consider six specifications for H_t , namely the BEKK by Engle and Kroner, 1995, together with the diagonal and the scalar variants, the RiskMetrics by JPMorgan, 1996, the Constant Conditional Correlation (CCC) by Bollerslev, 1990 and the Dynamic Conditional Correlation (DCC) by Engle, 2001. In the following Section we will briefly describe the models formulation².

2.1 BEKK-GARCH(1,1)

The BEKK-GARCH(1,1) is defined as follows:

$$H_t = C^* C^{*'} + A^{*'} \epsilon_{t-1} \epsilon_{t-1}' A^* + G^{*'} H_{t-1} G^*, \tag{3}$$

where C^* , A^* and G^* are $(N \times N)$ matrices and C^* is upper triangular. The number of parameters to be estimated is $N(5N + 1)/2$.

²See Bawens, Laurent and Rombouts (2006) for further detail

To reduce this number we can impose the diagonality of the parameter matrices A^* and G^* - i.e. Diagonal BEKK specification - or impose that A^* and G^* are equal to a scalar times the identity matrix - i.e. Scalar BEKK.

2.2 RiskMetrics

The RiskMetrics is an integrated Scalar-BEKK specification - i.e. $A^*A' + G^*G' = I_k$; where I is the identity matrix and k is the number of series - where the variance equation parameters are fixed and chosen ex-ante (does not require numerical optimization). The value assigned to the parameters is $a_{ij}^2 = 0.06$ and $g_{ij}^2 = 0.94 \forall i = j$ and 0 elsewhere.

2.3 CCC-GARCH(1,1)

The CCC is defined as follows:

$$H_t = D_t R D_t = \left(\rho_{ij} \sqrt{h_{iit} h_{jjt}} \right) \quad (4)$$

where

$$D_t = \text{diag} (h_{11t}^{1/2} \dots h_{NNt}^{1/2}) \quad (5)$$

h_{iit} can be defined as any univariate GARCH model, and

$$R = (\rho_{ij}) \quad (6)$$

is a symmetric positive definite matrix with $\rho_{ii} = 1, \forall i$.

R is the matrix containing the constant conditional correlations ρ_{ij} . In this paper we specify the univariate process as a GARCH(1,1) specification for each conditional variance in D_t :

$$h_{iit} = \omega_i + \alpha_i \epsilon_{i,t-1}^2 + \beta_i h_{iit-1} \quad (7)$$

2.4 DCC(1,1)-GARCH(1,1)

The DCC(1,1) model by Engle (2002) can be specified as:

$$R_t = \text{diag} (q_{11,t}^{-1/2} \dots q_{NN,t}^{-1/2}) Q_t \text{diag} (q_{11,t}^{-1/2} \dots q_{NN,t}^{-1/2}) \quad (8)$$

where the $(N \times N)$ symmetric positive definite matrix $Q_t = (q_{ij,t})$ is given by:

$$Q_t = (1 - \theta_p - \theta_q) \bar{Q} + \theta_p u_{t-1} u_{t-1}' + \theta_q Q_{t-1} \quad (9)$$

with $u_t = \epsilon_{i,t} / \sqrt{h_{iit}}$. \bar{Q} is the $(N \times N)$ unconditional variance matrix of u_t , and θ_p and θ_q are nonnegative scalar parameters with $\theta_p + \theta_q < 1$.

3 Model Comparison: distance metrics

It is not obvious which loss function is more appropriate to evaluate the distance between two matrices as required by the problem we are facing of comparing covariance matrices - i.e. the latent (Σ_t) and the estimated (H_t) covariance matrices sequences. In this paper we use the following four loss functions which, for their properties, seems to be appropriate for our purposes.

3.1 Frobenius Metric

The idea behind Frobenius norm is very similar to that of the Euclidean norm on R^n . The Frobenius distance between two matrices (Σ_t) and (H_t) is defined as:

$$\begin{aligned} d_t^2 &= \sum_{i,j} |\sigma_{t,ij} - h_{t,ij}|^2 \quad \forall i, j \\ &= \text{Trace} [(\Sigma_t - H_t)(\Sigma_t - H_t)^*] \\ &= \sum_{i=1}^{\text{rows}(\Sigma_t - H_t)} SV_i(\Sigma_t - H_t) \end{aligned} \quad (10)$$

where $(.)^*$ denotes the conjugate transpose of $(\Sigma_t - H_t)$ and $SV(.)$ are the singular values of $(\Sigma_t - H_t)$ which are defined as the eigenvalues of $[(\Sigma_t - H_t)(\Sigma_t - H_t)^*]^{1/2}$. Since $(\Sigma_t - H_t)^*$ has only real entries and it is symmetric then it is defined as self-adjoint or Hermitian matrix, that is $(\Sigma_t - H_t)' = (\Sigma_t - H_t)^*$. Therefore we can compute the Frobenius norm as the square root of the sum of element-wise squared differences, that is:

$$d_t = \sqrt{\sum_{i,j} (\sigma_{t,ij} - h_{t,ij})^2} \quad \forall i, j \quad (11)$$

Given the second equality in Eq. 10 this distance metric is also called trace norm which is defined as the sum of all the singular values of $(\Sigma_t - H_t)$.

3.2 Eigenvalue Metric

The Eigenvalue metric, or spectral norm, is based on the concept of a standard principal component analysis. Assuming Σ_t and H_t to be two covariance matrices, therefore symmetric and with real entries, then the difference matrix at time t is Hermitian and the distance is given by:

$$d_t = \sqrt{\lambda_{max}^1(\Sigma_t, H_t)} \quad (12)$$

where $\lambda_{max}^1(\Sigma_t, H_t)$ represents the largest eigenvalue of the matrix $(\Sigma_t - H_t)$ times its transpose - i.e. $(\Sigma_t - H_t)(\Sigma_t - H_t)'$.

Since Σ_t and H_t are covariance matrices, therefore symmetric and semi-positive definite, standard principal component analysis implies that the largest eigenvalue is non-negative and captures the difference in form of Σ_t and H_t completely, providing a metric for the largest amount of error between these two matrices.

Alternatively, the spectral norm can be defined as the largest singular value of the semi positive definite matrix $(\Sigma_t - H_t)$, i.e.:

$$d_t = \lambda_{max}^1 \left(\sqrt{(\Sigma_t - H_t)(\Sigma_t - H_t)'} \right) \quad (13)$$

where once again $\lambda_{max}^1(\cdot)$ represents the largest eigenvalue of (\cdot) .

From here to the end of this paper and in our empirical analysis we will always refer to the definition provided in Eq. 12.

3.3 Foerstner and Moonen Metric

This metric has been introduced by Foerstner and Moonen (1999) as a generalization of the idea introduced in Section 3.2. The distance between two symmetric semi-positive definite matrices Σ_t and H_t is defined as the sum of the squared logarithms of the eigenvalues, that is:

$$d_t = \sqrt{\sum_{i=1}^n \ln^2(\lambda_i^1(\Sigma_t, H_t))} \quad (14)$$

with eigenvalues $\lambda_i^1(\Sigma_t, H_t)$ from the system $|\lambda\Sigma_t - H_t| = 0$. The square root of summing squares closely resemble the Euclidean metric. Again, since the covariance matrices are symmetric and semi-positive definite, the eigenvalues are ensured to be positive³.

3.4 Cosinus Mass Metric

This measure is quite uncommon with respect to metrics commonly used in the econometrics literature but its simplicity appeals in our framework. Since Σ_t and H_t are symmetric semi-positive definite matrices the distance can be defined as:

$$d_t = \cos^{-1} \left(\frac{(LT\Sigma_t)'(LTH_t)}{\sqrt{[(LT\Sigma_t)'(LT\Sigma_t)] \cdot [(LTH_t)'(LTH_t)]}} \right) \quad (15)$$

where $LT\Sigma_t = vech(\Sigma_t)$ and $LTH_t = vech(H_t)$ and $vech(\cdot)$ denotes the operator that stacks the lower triangular portion of a $(k \times k)$ matrix as a $(k(k+1)/2 \times 1)$ vector.

Intuitively, the distance is measured by the angle between the two vectors defined by the stacked lower triangulars of Σ_t and H_t , and hence the smaller is the angle, the more Σ_t and H_t are closer.

Given these four definitions we can define the fit of a multivariate GARCH-type model by the average distance between the matrices Σ_t and H_t over t :

$$ModelFitness = \frac{1}{T} \sum_{t=1}^T d_t \quad (16)$$

To provide an example we generate two random variables with mean zero, conditional variance that follows individually an univariate $GARCH(1, 1)$ process and conditional correlation generated by the $DCC(1,1)$ process by Engle (2002). The parameters of the univariate models and of the unconditional correlation equation are:

³Further details on the properties of this metric are in Foerstner and Moonen (1999)

Table 1: DGP specification for GARCH(1,1)-DCC(1,1) process

Sample size: 20000		Number of series: 2			
GARCH(1,1)		DCC(1,1)			
Equation for σ_1^2		Equation for σ_2^2		Correlation Equation	
ω_1	0.2	ω_2	0.3	$\bar{\rho}$	0.6
α_1	0.1	α_2	0.2	θ_p	0.2
β_1	0.7	β_2	0.5	θ_q	0.7

We will now estimate the simulated process using a CCC and a DCC model respectively and then we will compute the distance between the sequences of both in-sample conditional variance (H_t) and conditional correlation (R_t) sequences and the underlying Σ_t and Ψ_t using the four distance metrics illustrated in the beginning of this Section. Table 2 and 3 contain the estimation results.

Table 2: Estimation results for GARCH(1,1)-CCC

Sample size: 20000			Number of series: 2				
GARCH(1,1)							
Equation for σ_1^2			Equation for σ_2^2				
	Coefficient	Std. Error	t-value		Coefficient	Std. Error	t-value
ω_1	0.178802	0.024298	7.359	ω_2	0.289979	0.019586	14.81
α_1	0.089794	0.0080861	11.10	α_2	0.187751	0.0098843	18.99
β_1	0.733449	0.029769	24.64	β_2	0.525422	0.024627	21.33
CCC							
	Coefficient	Std. Error	t-value				
ρ	0.46227	0.0050469	91.596				

The results in Table 4 confirm that DCC model better fits the true DGP outperforming the CCC whether we consider the Frobenius norm, Eigenvalue metric, Forstner and Moonen or Cosinus Mass metric. The difference in the scale is obviously due to the different intuition behind each metric.

If on one side this provides a simple way to rank different models, based on sample performances, whether we compare in sample estimation or out of sample forecasts, in the next sections we will describe a tool to test whether the differences between these loss functions - i.e. $X_t = (\Sigma_t - H_{t,0}) - (\Sigma_t - H_{t,i}) \quad \forall i$ - significantly differs from zero and therefore if one model statistically outperforms the all the others.

Table 3: Estimation results for GARCH(1,1)-DCC(1,1)

Sample size: 20000				Number of series: 2			
GARCH(1,1)							
Equation for σ_1^2				Equation for σ_2^2			
	Coefficient	Std. Error	t-value		Coefficient	Std. Error	t-value
ω_1	0.178802	0.024298	7.359	ω_2	0.289979	0.019586	14.81
α_1	0.089794	0.0080861	11.10	α_2	0.187751	0.0098843	18.99
β_1	0.733449	0.029769	24.64	β_2	0.525422	0.024627	21.33
DCC(1,1)							
	Coefficient	Std. Error	t-value				
ρ	0.51891	0.011481	45.197				
θ_p	0.17323	0.0048449	35.755				
θ_q	0.74330	0.0076578	97.064				

Table 4: Distance Metrics

Sample size: 20000 - Number of series: 2

Frobenius Metric	$\Sigma_t - H_{t,DCC}$	0.13564	$\Psi_t - R_{t,DCC}$	0.13953
	$\Sigma_t - H_{t,CCC}$	0.37177	$\Psi_t - R_{t,CCC}$	0.36814
Eigenvalue Metric	$\Sigma_t - H_{t,DCC}$	0.10710	$\Psi_t - R_{t,DCC}$	0.098662
	$\Sigma_t - H_{t,CCC}$	0.27548	$\Psi_t - R_{t,CCC}$	0.26032
Forstner Metric	$\Sigma_t - H_{t,DCC}$	0.23337	$\Psi_t - R_{t,DCC}$	0.22773
	$\Sigma_t - H_{t,CCC}$	0.56888	$\Psi_t - R_{t,CCC}$	0.56565
Cos Mass Metric	$\Sigma_t - H_{t,DCC}$	0.062785	$\Psi_t - R_{t,DCC}$	0.062947
	$\Sigma_t - H_{t,CCC}$	0.16320	$\Psi_t - R_{t,CCC}$	0.16641

4 Realized Volatility

In our empirical analysis we substitute the realized variance matrix to the latent conditional covariance when evaluating the goodness of the endogeneously generated volatility implied by the different M-GARCH-type models. The daily realized variance is computed from the intraday returns and is defined as:

$$r_{t+\Delta,\Delta} \equiv p_{t+\Delta} - p_t \quad (17)$$

for $t = \Delta, 2\Delta, \dots, T$ and $\Delta = 1/m$, where m is the number of daily intervals.

The corresponding daily return, for $t = 1, 2, \dots, T$ over the time interval of length $\Delta = 1/m$, is defined by:

$$r_{t+1} \equiv \sum_{i=1}^m r_{t+i\Delta,\Delta} \equiv r_{t+\Delta,\Delta} + r_{t+2\Delta,\Delta} + \dots + r_{t+1,\Delta} \quad (18)$$

Then, starting from the $(Tm \times n)$ matrix of intraday returns, we define the *daily realized*

volatility as:

$$V_{t+1} \equiv \sum_{i=1}^{1/\Delta} r_{t+i\Delta,\Delta} r'_{t+i\Delta,\Delta} = R'_{t+1} R_{t+1} \quad (19)$$

where the $(m \times n)$ matrix R_{t+1} is defined by $R'_{t+1} = (r_{t+\Delta,\Delta}, r_{t+2\Delta,\Delta}, \dots, r_{t+1,\Delta})$. The matrix V_{t+1} represents the empirical counterpart to the daily quadratic return variation. In fact as the sampling frequency of the intraday returns increases, $\Delta \rightarrow 0$ - i.e. $m \rightarrow \infty$, V_{t+1} converges almost surely to the quadratic variation (QV):

$$\Delta \rightarrow 0, \quad V_{t+1}(\Delta) \xrightarrow{as} \int_t^{t+1} \Sigma_u du \equiv QV_t \quad (20)$$

Therefore the realized volatility V_t becomes a less noisy measure for the latent volatility as the intraday frequency Δ reduces⁴. As shown in Andersen *et al.* (2002), to ensure positive semi definiteness of the realized covariance matrices it suffices that the columns in the matrix of returns R_t are linearly independent. However this condition, which could eventually be preserved even for high dimensional problems, will be violated when the cross-sectional dimension increases with respect to sampling frequency of the intraday return. The condition $n < m$ constitutes the dimensional upper bound to ensure semi positive definiteness of the realized covariance matrices since if the $rank(R_t) < n$, $V_t = R'_t R_t$ will not be full rank and will fail to be semi positive definite.

5 Model Comparison: Hansen SPA Test

A way to test model differences across models is the asymptotic test by Diebold and Mariano (1995) or the non parametric test by Pesaran and Timmermann (1992). However the limit of these tests is that they only implement pairwise comparisons across models. Since our aim is to test whether a model, when compared to the true DGP, is outperformed by *all the other models*, we need to test their performances together. An alternative method (Andersen *et al.*, 2002), in the tradition of Mincer-Zarnowitz (1969), is to evaluate volatility forecasts, and thus their relative performances, by projecting the realized volatility on a constant and the various model forecasts. The Mincer-Zarnowitz regression takes the form:

$$vech(V_{t+1}) = b_0 + b_1 vech(H_{t+1,i}) + b_2 vech(H_{t+1,j}) + u_{t+1} \quad (21)$$

where b_0 is expected to be equal to $(0, \dots, 0)$ and $b_1 = (1, \dots, 1)$ and $b_2 = (0, \dots, 0)$ - or alternatively $b_1 = (0, \dots, 0)$ and $b_2 = (1, \dots, 1)$ - defines the best model forecast for the realized volatility. The R^2 of the regression represents the goodness of fit and therefore the term of comparison between different models. However, the R^2 of a Mincer-Zarnowitz regression is not an ideal criterion for comparing volatility models, because it does not penalize a biased forecast⁵.

⁴Note that the optimal choice for the intraday frequency needs to take into account the distortion due to microstructure effects

⁵See Hansen and Lunde (2005)

In this paper, we use the Test for Superior Predictive Ability by Hansen (2001) which is based on the Reality Check Test by White (2000)⁶ and the work of Diebold and Mariano (1995) and West (1996). This test controls for the whole set of models to be compared.

Since our aim is to compare the performances of two or more multivariate models, while Hansen's SPA Test is originally designed for the univariate framework⁷, first we have to introduce new notions of distance metrics - i.e. the distance metrics defined in Section 3 - to fit the test to the multivariate setting⁸. However, since the test in itself is based on the average amount of distance between pairs of points in time, which is a scalar regardless of the dimension, it remains substantially the same but for some algebraical adaptation to the higher dimensional setting.

Let $n = 0, 1, \dots, N$ denote the number of estimated models for which we compute one-step-ahead covariance forecasts and let denote $H_{n,1}, \dots, H_{n,T}$ the sequence of covariance matrices forecasts for model n which will be compared to the latent covariance matrix Σ_t by mean of a predefined loss function $L_{n,t} \equiv L(\Sigma_t, H_{n,t})$, $t = 1, \dots, T$. Yet, let denote with $n = 0$ the model that is chosen as the benchmark and that is compared to the set of competing models $n = 1, \dots, N$. By using SPA Test we can analyse whether any of the competing models outperforms the benchmark model in terms of predictive ability.

The relative performance of model n ($n = 1, \dots, N$), with respect to the benchmark model ($n = 0$) at time t , is defined as:

$$X_{t,n} = L_{0,t} - L_{n,t}; \quad n = 1, \dots, N; \quad t = 1, \dots, T \quad (22)$$

Hansen (2001) showed that under reasonable assumptions the moment $\lambda_n \equiv E[X_{t,n}]$ is well-defined for $n = 1, \dots, N$. Therefore, a $\lambda_n > 0$ implies that model n outperforms the benchmark and allows for a formulation of the null hypothesis of the SPA test. More clearly, under the null hypothesis we will verify if $\lambda_n \leq 0$ for $n = 1, \dots, N$. If one of the competing models outperforms the benchmark, the null hypothesis will be rejected. Focusing just on the model which yields the highest λ_n gives us the following null hypothesis:

$$H_0 : \quad \lambda_{max}^s \equiv \max_n \frac{\lambda_n}{\omega_{nn}} \leq 0 \quad (23)$$

where λ_{max}^s denotes the best standardized relative performance and ω_{nn} the asymptotic variance of λ_n . Since the sample average $\bar{X}_{t,n} = t^{-1} \sum_{i=1}^T X_{i,n}$ is a consistent estimate for λ_n the corresponding test statistic is:

$$T_t^{sm} = \sqrt{t} \bar{X}_{t,max}^s \quad (24)$$

where $\bar{X}_{t,max}^s = \max_n \frac{\bar{X}_{t,n}}{\hat{\omega}_{nn}^2}$ and $\hat{\omega}_{nn}^2$ is a consistent estimator of $\omega_{nn}^2 \equiv \lim_{n \rightarrow \infty} Var(\sqrt{n} \bar{X}_{t,n})$ and is

⁶Hansen's Spa Test demonstrates better power properties. See Hansen (2001) for further details

⁷In the multivariate framework arrays of covariance matrices rather than vectors of variances are compared

⁸The choice of an appropriate loss function in the univariate case usually boils down to some measure of squared errors. The test in fact provides two predefined loss functions: Mean Squared Error (MSE) and Mean Absolute Deviation (MAD)

estimated using the bootstrap method⁹. Hence, T_t^{sm} represents the largest t-statistic (of relative performance) where the superscript sm stands for 'standardized maximum'. Under regularity conditions, stated above:

$$T_t^{sm} = \max_n \frac{\bar{X}_{t,n}}{\hat{\omega}_{nn}} \xrightarrow{p} \max_n \frac{\lambda_n}{\omega_{nn}} \quad (25)$$

which is lower than 0 if and only if $\lambda_n \leq 0$ for some n . More clearly, the question is whether $\bar{X}_{t,max}^s$ is large enough in order to reject the null hypothesis of $\lambda_{max}^s \leq 0$. Therefore, the SPA test, using the bootstrap technique in order to estimate the distribution of $\bar{X}_{t,max}^s$ under the null hypothesis, allows to determine critical values as thresholds where $\bar{X}_{t,max}^s$ becomes too large to be consistent with the null hypothesis. Rather than giving a detailed description of the bootstrapping¹⁰ we will focus on the general idea behind this method. Given the previous assumptions - i.e. $\lambda_n \equiv E[X_{t,n}]$ is well-defined for $n = 1, \dots, N$ - it holds that:

$$\sqrt{t}(\bar{X} - \lambda) \xrightarrow{d} N(0, \Omega) \quad (26)$$

where $\lambda = (\lambda_1, \dots, \lambda_N)'$, $\bar{X} = (\bar{X}_1, \dots, \bar{X}_N)'$ and $\Omega = E[n(\bar{X} - \lambda)(\bar{X} - \lambda)']$. The covariance matrix Ω will be estimated by bootstrapping methods. More precisely, the test uses the observed sample in order to generate k random resamples - i.e. *iid* bootstrap method - from the distribution of \bar{X} . For each of these resamples an estimate of ω_{nn}^2 is computed and therefore the distribution of $\bar{X}_{t,max}^s$ approximated. Finally, critical values and p-values can be calculated from the estimated distribution of $\bar{X}_{t,max}^s$.

In other words, the SPA test is a method which yields the probability distribution against which one compares the best performance from those of some competing models. It generates the probability distribution of the model which performs best relative to the benchmark. This distribution is certainly not a standard one which we can look up in a text book but rather strongly depends on the models which enter our model comparison. As a result, the distribution is different in every case and one has to find it by means of statistical methods, such as the bootstrap.

6 Data and Empirical Results

Our empirical analysis is based on General Electric (GE) and IBM stock returns for the period January 1, 1981 - December 31, 1999. Returns and realized variances are computed from equally spaced five-minute prices from the Trade and Quote (TAQ) database - i.e. 78 intraday observations. The choice of 5-min. intervals strikes a satisfactory trade off between the accuracy of the continuous time underlying process and the noise due to market microstructure effects.

Missing values are obtained by linearly interpolating the closest previous and the first following 5-min. price. The five-minute returns are computed as the first difference of the logarithmic prices premultiplied by 100. The dataset has been cleaned from weekends and holidays and similarly,

⁹See Hansen (2001) and Hansen, Kim and Lunde (2003) for further details

¹⁰See Hansen (2001) for technical details

we do not consider early closing days. The period from January 1, 1995 to the end of the sample will be used for the out of sample evaluation to compare model forecasts. We finally end up with 3652 in sample and 1187 out of sample observations. Out of sample forecasts have been computed by means of a moving window of dimension $k = 100$ within which, at each iteration k , the one step ahead forecast has been computed replacing the expectations for the $\epsilon'_{t+1}\epsilon_{t+1}$ in the variance equation forecast by its observed value. The coefficients are updated every 100 observations reestimating the model by including the last 100 out of sample observations into the dataset.

Estimation results for the six competing models are available on request. The dynamic behind the one step ahead covariance matrices forecast is reported in Figure 1. Table 5 contains the p-values of the out of sample one-step ahead forecasts comparisons under the null hypothesis that the benchmark model, the BEKK-GARCH(1,1), is the best forecasting model. The consistent *p-value* - i.e. *p-value_c* - is produced by the SPA test and is asymptotically valid and unbiased¹¹. This p-value informs whether the benchmark model is outperformed, in terms of predictive ability, by one or more competing models. More precisely, a high p-value informs that there is no evidence that one or more competing models outperform the benchmark. The SPA test provides also an upper - i.e. *p-value_u* - and a lower - i.e. *p-value_l* - bound for the consistent p-value. Tables 5, 6 and 7 report the *p-values* - i.e. the consistent and its bounds - the sample performance of the model under the null (benchmark), sample informations and the bootstrap parameters¹².

Table 5: SPA test p-values - A

Benchmark model: BEKK-GARCH(1,1)				
Number of models: $n = 5$			Sample size: $n = 1100$	
Test Statistic: TestStatScaledMax()				
Bootstrap parameters: $B = 1000$ (resamples)			$q = 0.5$ (dependence)	
Loss function	<i>p-value_l</i>	<i>p-value_c</i>	<i>p-value_u</i>	Sample performance
Frobenius Metric	0.48000	0.71600	0.93500	-3.48049
Eigenvalue Metric	0.47700	0.73800	0.95500	-3.35674
Forstner Metric	0.59000	0.59000	0.99800	-0.84536
Cos Mass Metric	0.48600	0.48600	1.00000	-0.27176

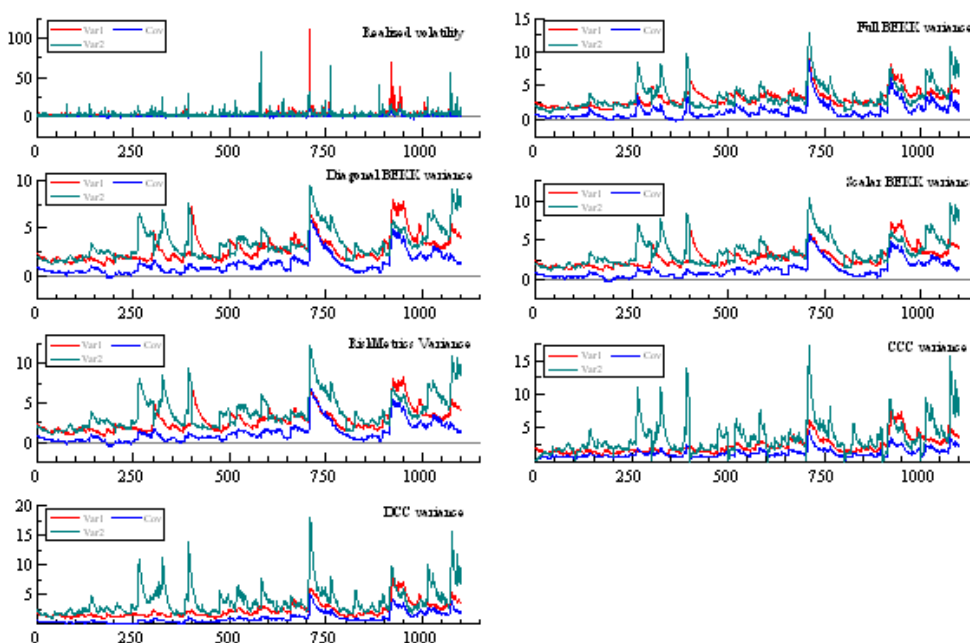
The results in Table 5 show that there is no evidence that BEKK-GARCH(1,1) is outperformed by any of the other models. Moreover BEKK-GARCH(1,1) has always the best sample performance measured as the average distance between the out of sample realized variance and the BEKK-GARCH(1,1) one step ahead variance forecasts.

Further evidence is given by the results reported in Table 6 where the SPA test is performed

¹¹See Hansen, Kim and Lunde (2003) for technical details

¹²See Hansen, Kim and Lunde (2003) for details

Figure 1: Variance matrices forecasts



under the null that the DCC(1,1)-GARCH(1,1) is the best model. Here the null is rejected at least at 5% significance level confirming the results reported above.

Table 6: SPA test p-values - B

Benchmark model: DCC(1,1)-GARCH(1,1)				
Number of models: $n = 5$		Sample size: $n = 1100$		
Test Statistic: TestStatScaledMax()				
Bootstrap parameters: $B = 1000$ (resamples)		$q = 0.5$ (dependence)		
Loss function	$p - value_l$	$p - value_c$	$p - value_u$	Sample performance
Frobenius Metric	0.00600	0.00600	0.00900	-3.57184
Eigenvalue Metric	0.00300	0.00300	0.00500	-3.44561
Forstner Metric	0.02400	0.02400	0.04100	-0.86912
Cos Mass Metric	0.00000	0.00000	0.00000	-0.31193

A partially contradicting conclusion is suggested by the results reported in Table 7 where the Diagonal-BEKK-GARCH(1,1) model has been chosen as benchmark model. This check is performed since the latter specification is a variant of the BEKK-GARCH model, which results in a lower parametrization by means of parameter pooling, and it ranks as second best model when sample performances are considered.

There is no evidence that, when using the Eigenvalue loss function, the model is outperformed

Table 7: SPA test p-values - C

Benchmark model: Diagonal BEKK-GARCH(1,1)				
Number of models: $n = 5$			Sample size: $n = 1100$	
Test Statistic: TestStatScaledMax()				
Bootstrap parameters: $B = 1000$ (resamples)			$q = 0.5$ (dependence)	
Loss function	$p - value_l$	$p - value_c$	$p - value_u$	Sample performance
Frobenius Metric	0.28600	0.28600	0.67100	-3.49623
Eigenvalue Metric	0.26300	0.26300	0.62500	-3.37531
Forstner Metric	0.00600	0.00900	0.02300	-0.86185
Cos Mass Metric	0.00000	0.00000	0.00000	-0.28482

by the others since the SPA test consistent p-value is not negligible, even though if we consider any of the other metrics the null is always rejected.

7 Conclusions

In this paper we have estimated and compared six different multivariate GARCH-family volatility models, namely BEKK (Full, Diagonal and Scalar specifications), Riskmetrics, CCC and DCC (Engle 2001) models in terms of their out of sample conditional covariance forecast ability. The dataset used consists of daily stock returns of GE and IBM. Five-min. returns are used to compute the realized variance matrix, which represents a proxy for the latent conditional covariance when evaluating out of sample forecasts model fitness. Our aim was to extend Hansen (2001) SPA test to the multivariate framework by providing four alternative matrix distance metrics, namely Frobenius, Eigenvalue, Forstner and Moonen and Cosinus Mass metric.

The main finding is that there is no evidence that the BEKK-GARCH(1,1) specification is outperformed by any of the other models, resulting to be the best forecasting model. Furthermore there is no evidence of SPA test lacking power in the multivariate setting since all the other models are rejected when the test is performed under the null that Diagonal and Scalar BEKK, Riskmetrics, CCC and DCC are considered as the benchmark¹³.

With this paper we also provide and test a first version of a new OxMetrix application to estimate and forecast multivariate GARCH-type volatility models. The package includes estimation and forecast procedure for: BEKK models (Full, Diagonal and Scalar), CCC and DCC models and RiskMetrics. It also includes simulation procedures and some functions to evaluate the distance between arrays of matrices.

¹³With the exception of Frobenius and Eigenvalue metrics under the null that Diagonal BEKK is the best forecasting model

References

- Andersen TG, Bollerslev T, Christoffersen PF, Diebold FX. 2005. Volatility forecasting. PIER Working Paper 05-011.
- Andersen TG, Bollerslev T, Diebold FX. 2002. Parametric and non parametric volatility measurement. *Handbook of Financial Econometrics*.
- Andersen TG, Bollerslev T, Diebold FX, Labys P. 2003. Modeling and forecasting realized volatility. *Econometrica* **71**: 579–625.
- Barndorff-Nielsen OE, Shephard N. 2001. Normal modified stable processes. *Theory of Probability and Mathematics Statistics* **65**: 1–19.
- Bauwens L, Laurent S, Rombouts JVK. 2003. Multivariate GARCH models: a survey. CORE DP 2003/31.
- 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.
- Brooks C, Burke SP, Persaud G. 2001. Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting* **17**: 45–56.
- Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.
- Engle R, Kroner FK. 1995. Multivariate simultaneous generalized ARCH. *Econometric Theory* **11**: 122–150.
- Engle RF. 2002. Dynamic conditional correlation - a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20**: 339–350.
- Engle RF, Sheppard K. 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Mimeo, UCSD.
- Foerstner W, Moonen B. 1999. A metric for covariance matrices. Dept. of Geodesy and Geoinformatics Technical report Stuttgart University.
- Hansen PR. 2001. A test for superior predictive ability. Brown University, Economics Working Paper. 2001-06.
- Hansen PR, Lunde A. 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1). *Journal of Applied Econometrics* **20**: 873–889.

- Hansen PR, Kim J, Lunde A. 2003. Testing For superior predictive ability using Ox. A manual for SPA for Ox. Brown University.
- Horn R, Johnson C. 1985. Matrix Analysis. *Cambridge University Press*.
- Pesaran MH, Timmerman A. 1992. A simple nonparametric test of predictive performance. *Journal of business and economic statistics* **10**: 461–465
- Riskmetrics. 1996. *Riskmetrics Technical Document, 4th ed.* New York: J.P. Morgan.
- Sheppard K. 2006. Realized Covariance and Scrambling. University of Oxford.
- Tse YK, Tsui AKC. 2002. A multivariate GARCH model with time-varying correlations. *Journal of Business and Economic Statistics* **20**: 351–362.
- West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067-1084.
- White H. 2000. reality check for data snooping. *Econometrica* **68**: 1097-1126.