DOCTOR OF ECONOMICS AND BUSINESS MANAGEMENT

Essays on Multivariate Volatility Forecasting

Violante, Francesco

Award date:
2010

Awarding institution:
University of Namur

Académie Louvain - Université de Namur
Faculté de sciences économiques, sociales et de gestion
Département des sciences économiques

Essays on Multivariate Volatility Forecasting

Francesco Violante

Thèse présentée en vue de l'obtention du grade
de docteur en sciences économiques et de gestion

December 2010

Composition of the Jury:

Chairman:	Pierre Giot (CeReFiM - FUNDP, University of Namur)
Supervisor:	Sébastien Laurent (Maastricht University)
Members:	Luc Bauwens (CORE-Université catholique de Louvain)
	Christian M. Hafner (STAT-Université catholique de Louvain)
	Franz C. Palm (Maastricht University)
	Jean-Michel Zakoian (CREST & University Lille 3)

Contents

Acknowledgments

I would first, and foremost, like to express my gratitude to my supervisor, Prof. Sébastien Laurent, who gave me the opportunity to pursue my PhD and whose expertise, support and patience added considerably to my graduate experience. I am sincerely grateful both professionally and personally.

I would like to thank the other members of the committee, Prof. Luc Bauwens, Prof. Christian Hafner, Prof. Pierre Giot and the two external members, Prof. Franz Palm and Prof. Jean-Michel Zakoian, for taking time from their busy schedule to serve as my readers and whose comments and suggestions helped to improve this thesis. I would like to particularly acknowledge Prof. Christian Hafner, whose input and contribution greatly improved Chapter 4.

I owe my deepest gratitude to Prof. Jeroen Rombouts for his invaluable contribution to this thesis, his support, advice and contagious enthusiasm towards research. My gratitude extends to his family, for kindly hosting me during long days of work and for the friendship that they offered.

I also acknowledge Raouf Boucekkine, Andrew Patton, Lars Stentoft, the participants to the 2nd International Workshop of the ERCIM Working Group on Computing and Statistics, the 64th European Meeting of the Econometric Society and the 26th Canadian Econometric Study Group, the IFM2 Mathematical Finance Days, the CIREQ 4th Times Series Conference, the (EC)2 Conference at CREATES, the 9th Journes du CIRPEE, the Royal Economic Society 2009 Conference, the 2nd MIFN Meeting, the 2nd ICEEE Conference, the (EC)2 Conference at Erasmus University and the seminar participants at Maastricht University and CREST-Paris for their helpful comments.

I owe a special thank to Maria Eugenia Sanin, invaluable companion in this adventure, for her constant and unconditional support, encouragement and understanding, at work and in life, during the last six years.

I thank my colleagues and friends at FUNDP and UCL with whom I shared five exceptional years, and in particular Jean Yves Gnabo, Jérôme Lahaye, Dennis Tuerk and Nicolas Debarsy with whom I shared more than a common interests for econometrics. I also acknowledge the members and the administrative staff of the Faculté des Sciences économiques, sociales et de gestion at FUNDP and UCL. I am particularly grateful to Pierrette Noel for her kindness and the precious assistance she provides to the PhD students.

Financial support from the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State Prime Ministers Office, science policy programming, CREF (HEC Montreal), CIRPEE, CeReFiM (FUNDP) and CORE (UCL) is gratefully acknowledged.

I dedicate this thesis to my family.

Chapter 1

Volatility Forecast Evaluation and Comparison: A Survey

1.1 Introduction

In this chapter we review recent developments on volatility forecasts evaluation and comparison based on inference of moments of functions of predictions and predictions errors. Since this chapter defines the theoretical background of this thesis, in the closing section we summarize and discuss the main contributions and results of this thesis.

In this chapter we consider both univariate and multivariate volatility models. The evaluation of forecasts performances as considered here is based on a sufficiently long sequence of point forecasts or forecast errors using a retrospective approach. Hence, it may be useful to stress that the techniques discussed here are not suited when the evaluation of a model forecast accuracy is based on a single or a small number of out of sample observations.

Depending on the purpose of the analysis, the forecaster may be interested in evaluating a single model, two or several models. When the object of the analysis is the forecasting accuracy of a single model, the quality of the model can be measured by the correlation between predictions and realizations. A common method that falls in this category is the Mincer-Zarnowitz regression (Mincer and Zarnowitz, 1969), which involves regressing the realization of a variable on a constant and its forecast. Alternatively, the forecaster may aim to compare two or more models. Tests for pairwise comparison have been proposed by Diebold and Mariano (1995), West (1996) and later generalized by Giacomini and White (2006). The multiple comparison problem can be tackled in different ways. We distinguish between two different approaches, the multiple comparison with control (e.g., the reality check for data snooping of White, 2000 and the superior predictive ability test of Hansen

and Lunde, 2005) where a benchmark forecast is chosen ex-ante and compared to all others to assess whether any of the competing forecasts outperforms the benchmark, and the multiple comparison without control (e.g., the model confidence set test of Hansen, Lunde, and Nason, 2010a) where all forecasts are compared against each other and poor performing models excluded.

A common problem in the evaluation of point forecasts is the comparison of nested models. In fact, depending on the forecasting scheme used to produce the sequence of out of sample observations, most of the tests for predictive ability discussed here may not apply, in the sense that the distribution under the null turns out to be degenerate in some cases or the test may suffer from serious size distortions. In this chapter we review the three most commonly used forecasting schemes.

Another problem, which characterizes the comparison of volatility forecasts, is the fact that the target variable is latent. Thus, the evaluation of forecasts or forecast errors has to be done with respect to some ex-post estimator based on observed outcomes as they become available. Typically, this problem is solved by using a conditionally unbiased (and possibly consistent) estimator as, for example, the squared innovations, the realized volatility or kernels, see Andersen and Bollerslev (1998) and the further developments by Barndorff-Nielsen and Shephard (2002), Zhang, Mykland, and Ait-Sahalia (2004), Zhou (1996), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) among the others, and their multivariate extension, see Andersen, Bollerslev, Diebold, and Labys (2003), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b) and Hansen and Lunde (2006b), or yet range based variance estimators, see Parkinson (1980), Garman and Klass (1980) and Brandt and Diebold (2006). In the reminder of the chapter we refer to the ex-post volatility estimator as the volatility proxy.

However, it is not always true that using a conditionally unbiased proxy will lead, asymptotically, to the same outcome that would be obtained if the true volatility was observed. Hansen and Lunde (2006a) show that when the evaluation is based on a target observed with error, the choice of the evaluation criteria becomes critical in order to avoid a perverse outcome. The problem of consistency, sometimes referred to as robustness, of the ordering between two or more volatility forecasts has been further developed in Patton (2009) and Patton and Sheppard (2009). In Chapter 2 we provide a generalization to the comparison of multivariate volatility forecasts.

Finally, since in most of the methods discussed here, the evaluation of volatility forecasts relies, more or less explicitly, on the use of a statistical loss function and on the choice of a volatility proxy, we discuss the properties of a number of admissible loss functions and elaborate on the value of high precision proxies.

The rest of the chapter is organized as follows. In Section 1.2, we discuss three forecasting schemes widely used in the forecasting literature. In Section 1.3, we introduce the basic notation used throughout the chapter. In Section 1.4, we discuss the evaluation of the predictive accuracy of single forecasts. In Section 1.5, we introduce the problem of forecast evaluation under imperfect volatility proxies and provide a number of admissible loss functions. In Sections 1.6 and 1.7, we discuss methods for pairwise and multiple forecast comparison respectively. In Section 1.8 we illustrate the problem of consistency of the ordering in a Monte Carlo simulation. In Section 1.9, we conclude and provide an overview of the thesis.

1.2 Forecasting schemes

In this section we review three forecasting schemes, namely recursive, rolling and fixed, that are most often encountered in the forecasting literature. The choice of the forecasting scheme has direct and important implications for the inference on the forecast performances. In fact, in most applications the forecasts under comparison depend on estimated parameters whose properties, in turn, depend on the forecasting scheme adopted.

Let us denote the total sample size as $\mathcal{T}+T$, where $\mathcal{T}$ represents the estimation sample and T the out of sample evaluation sample. In the reminder of the chapter we will consider one-step ahead forecasts, although the results can be easily extended to the multi-step ahead case.

In the recursive scheme the sample used to estimate the parameters of the model grows as the forecaster makes predictions for successive observations, i.e. at each step the forecast is based on all available information. For example, at $\mathcal{T}+1$ one evaluates the first forecast using parameter estimates which include information up to $\mathcal{T}$. Then, the second forecast, at $\mathcal{T}+2$, is obtained using parameter estimates which include data up to $\mathcal{T}+1$ and so forth, with the last forecast generated using parameter estimates computed using all but the last observations $(\mathcal{T}+T-1)$.

In the rolling scheme the sequence of T parameters is generated using a rolling sample of fixed size $\mathcal{T}$, i.e., the most recent information. For example, the first estimate is obtained using data from 1 to $\mathcal{T}$ and forecast at $\mathcal{T}+1$ computed. Then, the second estimate is obtained using data from 2 to $\mathcal{T}+1$ to obtain the forecast at $\mathcal{T}+2$, etc.

In the fixed scheme the parameters of the model are estimated only once using data from 1 to $\mathcal{T}$. Then, the estimates are used to generate all one-step ahead forecasts, i.e., at each step, $\mathcal{T}+1, \mathcal{T}+2, ..., \mathcal{T}+T-1$, only the data are updated with the new information.

In this chapter we will mainly focus on the rolling and fixed forecasting schemes. In

fact, as will be discussed in the following sections, the recursive scheme can be problematic when evaluating forecasts generated by nested models. Furthermore, the rolling and fixed scheme, other than allowing for the comparison of nested models, also present some other advantages. The rolling scheme is, in fact, rather appealing in case of heterogeneneity of the data or parameter drifts that cannot be easily modeled explicitly, whereas the fixed scheme can be useful when it is difficult to carry out parameter estimation. This is often the case, for instance, when comparing multivariate volatility models, where the large number of parameters makes the rolling scheme computationally challenging and time consuming, see Chapter 3 for an example where a combination of rolling and fixed schemes is used. A number of examples of applications using each of the three schemes can be found in West (2006).

1.3 Notation

We now introduce the basic notation and definitions used throughout the chapter. Let us define $t = 1, ..., T$ the time index of the forecast sample of size T. Let r_t be the a random variable whose conditional variance, $\mathrm{E}[r_t^2|\Im_{t-1}] = \mathrm{E}_{t-1}[r_t^2] = \sigma_t$, is of interest (to simplify the exposition, we assume that $\mathrm{E}[r_t|\Im_{t-1}] = \mathrm{E}_{t-1}[r_t] = 0$). The set $\Im_{t-1}$ denotes the information set at time $t-1$ and contains past realizations of r_t, but also other variables and variables measured at a higher frequency. In financial applications, r_t typically represents a sequence of returns, i.e. first difference of logarithmic prices, of some financial asset. We also assume that $r_t|\Im_{t-1} \sim F(0, \sigma_t)$, where F is some distribution with zero mean, finite variance and possibly constant higher order moments.

The (set of) variance forecast(s) (sometimes referred to as models) is denoted by h_t ($h_{k,t} \in \mathcal{H}$, $k = 1, ..., m$ when there is more than one model).

When the forecast accuracy is evaluated by means of a loss function, we denote it as $L : R_{++} \times \mathcal{H} \to R^+$ where R_+ and R_{++} denote the non-negative and positive portions of the real line respectively and $\mathcal{H}$ is a compact subset of R_{++} and identifies the set of volatility forecasts.

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

The loss function L is a scalar valued function defined as $L : R_{++}^{N \times N} \times \mathcal{H}^{N \times N} \to R^+$. Note

that, both in the univariate and multivariate case, the first argument of the loss function is the true variance or some proxy of it, whereas the second is a volatility forecast.

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

A simple variance proxy commonly used in the financial literature is the squared return, outer product of the return vector in the multivariate case. However, we discourage the use of such estimator for two reasons. First, although such proxy is conditionally unbiased for the latent variance, it is extremely noisy, which makes it unsuited in many situations, because the scarce informative content of the volatility proxy makes difficult to assess the statistical relevance of the forecast performances, see Section 1.8 for an example. Second, even for the smallest multivariate dimension, $N = 2$, this proxy violates the positive definiteness requirement for the volatility proxy which suggests the use of other variance proxies, see Section 1.1 for some examples.

1.4 Single forecast evaluation

A simple method for evaluating the accuracy of a volatility forecast is the well known Mincer-Zarnowitz (MZ) regression, see Mincer and Zarnowitz (1969). In its simplest specification, this approach requires the estimation of the coefficients of a regression of the target on a constant and the forecast under evaluation, i.e.,

$$\sigma_t = \alpha + \beta h_t + \epsilon_t. \tag{1.1}$$

The null hypothesis of optimality of the forecast can be written as $H^0 : h_t = \sigma_t \ a.s. \ \forall t$ against the alternative hypothesis $H^1 : h_t \neq \sigma_t$ for some t. As underlined in the previous sections, the regression in (1.1) is unfeasible due to the latent nature of the target variable. If we substitute the true variance by some conditionally unbiased proxy, $\hat{\sigma}_t = \sigma_t + \eta_t$ with

$\mathrm{E}_{t-1}[\eta_t] = 0$, then we can rewrite (1.1) as

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

where the innovations are $e_t = \eta_t + \epsilon_t$. Since $\hat{\sigma_t}$ is a conditionally unbiased estimator of the true variance then (1.2) yields unbiased estimates of α and β. The MZ regression allows to evaluate two different aspects of the volatility forecast. First, an obvious property of a good sequence of forecasts is that there are no systematic over- or under-predictions, that is the forecast is unbiased. The MZ regression allows to test such property by testing the joint hypothesis $H^0 : \alpha = 0 \cup \beta = 1$. Second, since the R^2 of (1.2) represents an indicator of the correlation between the realization and the forecast, it can be used as evaluation criterion of the accuracy of the forecast.

Clearly the variance of the innovation term e_t in (1.2) depends on the accuracy of the volatility proxy. Therefore, when a high quality proxy is available, the regression parameters are estimated more accurately. Furthermore, the R^2 of the regression in (1.2), $\mathrm{Cov}(\hat{\sigma_t}, h_t)^2/(\mathrm{Var}(\hat{\sigma_t})\mathrm{Var}(h_t))$, is penalized as the quality of the proxy deteriorates, see Andersen and Bollerslev (1998) for an analytical example. This results validates the use of high precision proxies in order to offset the latent variable problem. However, it is worth noting that while the use of a less accurate proxy affects the precision of the regression parameters it does not affect the validity of the test.

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

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

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

and therefore

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

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

To respond to the concern that few extreme observations can drive the forecast evaluation, many authors have argued in favor of MZ regressions on transformations of σ_t (and consequently of $\hat{\sigma}_t$ and h_t), for instance $\log(\hat{\sigma}_t)$ on $\log(h_t)$ or $|r_t|$ on $\sqrt{h_t}$, see Pagan and Schwert (1990), Jorion (1995), Bollerslev and Wright (2001) among others for some examples.

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

The R^2 of the MZ regression has been often used as a criterion for ordering over a set of volatility forecasts, see Andersen and Bollerslev (1998) and Andersen, Bollerslev, Diebold, and Labys (2003) for several examples. Hansen and Lunde (2006a) show that, due to the latent variable problem, this criterion is not always adequate to the scope and may lead to a perverse outcome. They derive sufficient conditions under which the ordering of volatility forecast is unaffected when the true variance is substituted by a proxy. They establish that the R^2 is a valid criterion if $\mathrm{E}_{t-1}[\sigma_t - \hat{\sigma}_t](\partial^i \phi(\sigma_t)/\partial \sigma_t^i) = c_i$ for some constant c_i, $\forall t$ and $i \in \mathbb{N}$ and where $\phi(.)$ represents the transformation of the dependent variable and the regressor, e.g., log, square, square root, etc. This condition validates the use of the MZ regression in level but also, for example, of the quadratic transformation, i.e., $\phi(x) = x^2$,[1] but rejects, for example, the log-regression.

The results outlined above can be directly extended, with few exceptions to the multivariate case. A simple approach is to consider the unique elements of the true variance

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

matrix (or the proxy) and of the covariance forecast. The feasible MZ regression can be written as

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

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(N(N+1)/2 \times 1)$ vectors of parameters, vech(.) is the half-vector operator and diag(.) is the operator that transforms a vector into a square diagonal matrix with the vector along the diagonal. Equation (1.5) can be estimated as seemingly unrelated regression. Then a joint test that $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{1}$ can be performed. As pointed out by Patton and Sheppard (2009) the large dimension of the system may adversely affect the finite sample properties of the joint test. The solution proposed to reduce the parameter space is to impose in (1.5) the parameter constraints $\alpha_i = \alpha$ and $\beta_i = \beta \; \forall i = 1, ..., N(N+1)/2$.

1.5 Loss functions and the latent variable problem

A common approach to the evaluation of forecast performances is the comparison of expected loss evaluated with respect to the true variance. However, as noted in Section 1.4, the latent nature of the conditional variance makes it difficult to evaluate the performances of volatility forecasts. The latent variable problem can be solved, at least partly, by substituting the true conditional variance by some ex post estimator. Examples of volatility proxies have been provided in Section 1.1.

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

Let us define the precision, measured in terms of expected loss, of some generic volatility forecast, $h_{k,t}$, with respect to the true variance as $\text{E}[L(\sigma_t, h_{k,t})]$, where $L(.)$ is some known loss function as defined in Section 1.3. The aim is to seek conditions that ensure consistency of the ranking (equivalence of the ordering) between any two forecasts k and j when a conditionally unbiased proxy is substituted to the true variance, that is

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

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

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

A loss function satisfying (1.7) is referred to as robust. It follows immediately that (1.7) rejects evaluation criteria commonly used in applied work such as absolute deviation, root of squared error, proportional error loss functions, or mean squared error of the log-transform, whereas it validates the use of mean squared error. Numerous examples of loss functions violating (1.7) are discussed by Hansen and Lunde (2006a) and Patton (2009). Section 1.8, provides an illustration based on artificial data where mean squared error (robust) and mean squared error of the log-transform (non-robust) are compared.

Focussing on the univariate dimension, Patton (2009) provides analytical results for the undesirable outcome that arises when using a number of loss function that violates (1.7), under different distributional assumption for the returns and considering different volatility proxies and a number of commonly used loss functions. Furthermore, building upon Hansen and Lunde (2006a), he provides necessary and sufficient conditions on the functional form for the loss function (defined within the class of homogeneous statistical loss functions that can be expressed as means of each period loss) ensuring consistency of the ordering when using a proxy. The following family of functions

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

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

A first attempt of extension to the multivariate case has been proposed by Patton and Sheppard (2009). In Chapter 2 we generalize this setting and provide a general framework

for the evaluation of variance matrices. We identify a number of robust vector and matrix loss functions and provide insight on their properties, interpretation and geometrical representation. In the multivariate case, the sufficient condition in (1.7) becomes

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

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

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

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

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

where $C(.)$ and $C'(.)$ are the gradient and the hessian of $\tilde{C}(.)$ with respect to the $K = N(N+1)/2$ unique elements of H_t, denoted $\mathbf{h}_t = \text{vech}(H_t)$ and $C'(H_t)$ is negative definite.

Unlike the univariate case where an analytical expression is available for the entire class of robust loss functions, in the multivariate case such generalization is unfeasible because there are many functions $\tilde{C}(.)$ that can be used to weight forecasts and forecasts errors. However, given (1.10) application-specific loss functions can be easily derived. A number of examples can be found in Chapter 2.

Despite this limitation, from (1.10), we identify the entire subset of homogeneous ($\xi = 2$) loss functions based on forecast errors, i.e., $\Sigma_t - H_t$, which can be expressed as

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

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

Finally, we also show that, under the higher level assumption of consistency of the volatility proxy, the potential distortion introduced in the ordering when using a non-robust loss function tends to disappear as the quality of the proxy improves. Since often

non robust loss functions have other desirable properties which are useful in applications, e.g. down-weight extreme forecast errors, they may still be safely used provided that the volatility proxy can be assumed to be nearly perfect.

In the following sections, we review a number of tests for forecast evaluation where performances are evaluated by means of a statistical loss function. Although most of the methodologies discussed are valid under a general loss function, we remind that, in empirical applications, when the true variance is substituted by a proxy, the loss function should be chosen, depending on the setting, according to (1.8) and (1.10) respectively.

1.6 Pairwise comparison

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

$$L_k^{MSE} - L_j^{MSE} = (\widehat{\mathrm{Var}}(e_{k,t}) - \widehat{\mathrm{Var}}(e_{j,t})) + (\bar{e_k}^2 - \bar{e_j}^2), \tag{1.13}$$

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

$$L_k^{MSE} - L_j^{MSE} = \widehat{\mathrm{Cov}}(D_t, S_t) + \bar{D}\bar{S}. \tag{1.14}$$

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

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

where $\mathrm{Cov}(D_t, S_t)$ and $\mathrm{E}D_t$ denote the population covariance between D_t and S_t and the population mean of D_t. Note that (1.15) implies that the forecasts can be biased. In fact, $\bar{D} = 0$ does not require $\bar{e}_k = \bar{e}_j = 0$ but only that the biases are equal in size and sign. The null hypothesis in (1.15) is equivalent to testing the null hypothesis $H^0 : \alpha = 0 \cup \beta = 0$ in the following regression

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

the error term ϵ_t has mean zero and it can be treated as independent of S_t. In the case independence is violated Ashley, Granger, and Schmelensee (1980) show that α is estimated without bias, while the bias in β becomes negligible for moderate samples.[2] If the forecast errors have zero-mean, i.e., they are both unbiased, and under the additional assumption that they are normally distributed and uncorrelated, the test of equal MSE is equivalent to the test proposed by Granger and Newbold (1977), henceforth GN, that is

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

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

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

A more general approach is the well known test of equal predictive ability proposed by Diebold and Mariano (1995), henceforth DM, and further refined by West (1996), McCracken (2000), Clark and McCracken (2001), Corradi, Swanson, and Olivetti (2001), Clark and West (2006), Clark and West (2007) among others. The DM test is a very general procedure[3] designed to compare two rival forecasts in terms of their forecasting accuracy using a general loss function. The measure of predictive accuracy, i.e. the loss function, can be specified according to the definition of optimality adopted by the forecaster.

Consider a loss function as defined in Section 1.3 and define the loss differential between model k and j as

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

for the univariate case, whereas

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

[2] Although (1.15) is sufficient to test the null hypothesis of equal predictive ability, a null of the type $H^0 : \text{Cov}(D_t, S_t) = 0 \cup \text{E}S_t = 0$ would achieve the same result, the difference between the two being that in the latter case the forecasts if biased have biases equal in size but of opposite sign. Testing for this null is equivalent to testing the null hypothesis $H^0 : \alpha = 0 \cup \beta = 0$ in the regression $S_t = \alpha + \beta(D_t - \bar{D}) + \epsilon_t$.

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

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

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

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

As underlined in Section 1.2, when comparing parametric forecast models, an adequate choice of the forecast scheme is crucial to the validity of the procedure. In fact, suppose the forecaster aims to compare two volatility forecasts, $h_{k,t}(\theta_k)$, which we assume to be the true model, and a competing forecast $h_{j,t}(\theta_j)$ from an unrestricted model j which nests model k, i.e., $\theta_j = (\theta_k, \beta)$.

The original formulation the DM approach relies on asymptotics based on an estimation and forecast samples which increase at the same rate, i.e. it tests the hypothesis $H^0 : \mathrm{E}[L(\sigma_t, h_{k,t}(\theta_k^*)) - L(\sigma_t, h_{j,t}(\theta_j^*))] = 0$, where θ_k^* and θ_j^* are population values, and it is therefore designed for a recursive forecast scheme.[4] However, the two models being nested, the population value $\beta^* = 0$, $L(\sigma_t, h_{k,t}(\theta_k^*)) - L(\sigma_t, h_{j,t}(\theta_j^*))$ is a degenerate random variable identically zero, which invalidates the test. Consider now that the forecasts are based on the recursive scheme. Then, as $t \to \infty$, $\hat{\beta} \xrightarrow{p} 0$ which implies that $L(\sigma_t, h_{k,t}(\hat{\theta}_k)) - L(\sigma_t, h_{j,t}(\hat{\theta}_j))$ is degenerate for large t. A solution to this problem has been suggested by Giacomini and White (2006) which hold the size of estimation sample fixed as the sample size grows. Thus, the random variable of interest, $L(\sigma_t, h_{k,t}(\hat{\theta}_k)) - L(\sigma_t, h_{j,t}(\hat{\theta}_j))$ is non-degenerate for all t since $\hat{\beta} \neq 0$ *a.s.* for fixed estimation sample size, i.e., rolling and fixed schemes.

To allow for a unified treatment of nested and non-nested models, Giacomini and White (2006) (henceforth GW) construct a conditional rather than unconditional test. They suggest to view the problem of forecast evaluation as a problem of inference about conditional expectations of forecasts and forecast errors rather than unconditional expectations. Their approach is applicable in many situations where the standard tests of Diebold and Mariano (1995) and West (2006) are not valid, e.g., when the data are heterogeneous. The assumption of heterogeneity motivates a different approach to estimation. Their asymp-

[4] To avoid size distortions, West (1996) introduces a correction which accommodates for the fact that the actual forecasts that appear in the statistic depend on estimated parameters.

totics are based on an estimation sample of fixed size, while the forecast sample tends to infinity. Hence, coefficients always include parameter estimation error. In this context, instead of considering a recursive forecasting scheme, where the estimation window expands over time, it seems to be more reasonable to use a rolling window forecast procedure where the forecasts are based on a moving window of the data where old observations are discarded. Furthermore, the size of the estimation window can itself be time-varying, as in the procedure suggested by Pesaran and Timmerman (2007). The proposed methodology is also compatible with a fixed estimation sample forecasting scheme, where the parameters are estimated only once and in general any forecasts procedure based on limited memory estimators. For instance, it is also compatible with the recursive scheme, i.e., expanding window, when coupled with a weighted estimator with weights heavily discounting observations far in the past. Rather than the comparison of 'forecast models' stricto sensu, as in Diebold and Mariano (1995) and West (2006), Giacomini and White (2006) emphasize the role of what they refer to as the 'forecasting method', which includes not only the forecast model but also any other choice that the forecaster has to make at the time of the prediction, such as the estimation procedure, which data to use for estimation, which forecast scheme to adopt, etc. Their technique also allows to compare forecasts generated using different combinations of these choices.

The GW approach tests the null hypothesis of equal predictive ability

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

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

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

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

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

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

1.7 Multiple comparison

When multiple alternative forecasts are available, it is of interest to test whether a specific forecast (hereafter the benchmark), selected independently from the data, produces systematically superior (or at least equivalent) performances with respect to the rival models. In this case, we aim to a test the null hypothesis that the benchmark is not inferior to any of the alternatives. This approach, called multiple comparison with control, differs from the techniques discussed in Section 1.6 for two reasons. First, the multiple comparison allows to recognize the multiplicity effect, i.e., statistical relevance of all comparisons between the benchmark and each of the alternative models, which calls for a test of multiple hypotheses to control for the size of the overall testing procedure. Second, while Section 1.6 involves tests of equal predictive ability, the choice of a control requires a test of superior predictive ability. The distinction is crucial because, while the former lead to simple null hypotheses, i.e., testing equalities, the latter leads to composite hypotheses, i.e. testing (weak) inequalities. The main complications in composite hypotheses testing is that (asymptotic) distributions typically depend on nuisance parameters, hence the distribution under the null is not unique.

To simplify the exposition, the notation used in this section only refers the univariate dimension. Since all the techniques discussed hereafter are based on comparisons of forecast performances measured by some statistical loss function, the extension to the multivariate case, as noted in Section 1.6, is straightforward and only involves an appropriate redefinition of the loss function, namely $L : R_{++}^{N\times N} \times \mathcal{H}^{N\times N} \to R^{+}$. Issues related to the choice of the loss function, the latent variable problem have been discussed in Section 1.5.

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

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

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

the null hypothesis of interest takes the form

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

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

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

where $\bar{d}_k = T^{-1}\sum_{t=1}^{T} d_{k,t}$. Note that, as in Diebold and Mariano (1995), White's (2000) original version of the RC test sets the parameters of the model based forecasts to their population values, thus not allowing for the comparison of nested models. Using the same arguments of Giacomini and White (2006), Hansen (2005), and in a related setting Hansen, Lunde, and Nason (2010a), extends the procedure to the comparison of nested models. Hansen's (2005) and Hansen, Lunde, and Nason's (2010a) framework is well suited when parameters are estimated once, i.e., fixed scheme, or using a moving window (of fixed or time dependent size or yet of different size for each model), i.e., rolling schemes, whereas the comparison of models with parameters that are estimated recursively is not accommodated by their framework.

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

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

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

Using a similar approach, Hansen (2005) proposes a new test for superior predictive ability (henceforth SPA). His framework differs from White (2000) in two ways. First, he

proposes a different statistic based on studentized quantities to alleviate the substancial loss of power that the RC can suffer due to the inclusion of poor and irrelevant forecasts. Second, he employs a sample dependent distribution under the null. The latter is based on a procedure that incorporates additional sample information in order to identify the relevant alternatives. In fact, while the procedure based on the LFC suggested in White (2000) implicitly relies on an asymptotic distribution under the null hypothesis that assumes $\mathrm{E}[d_{k,t}] = 0$ for all k, Hansen (2005) points out that all negative values of $\mathrm{E}[d_{k,t}]$ have also to be considered since they conform with the null hypothesis.

The new statistic takes the form

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

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

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

where $1_{\{.\}}$ denotes the indicator function and $\hat{\omega}_k$ is a consistent estimator on the kth diagonal element of Ω. The threshold $(2 \log\log T)^{1/2}$ in (1.28) represents the slowest rate that captures all alternatives with $\mu_k = 0$. More generally, any threshold rates in the interval $\left[(2 \log\log T)^{1/2},\ T^{1/2-\epsilon}\right]$ for any $\epsilon > 0$, also produce a valid test and guarantee that all poor models are discarded asymptotically. For instance, the value $0.25\,T^{0.25}$ is used in an empirical exercise. Furthermore, since different threshold rates lead to different p-values in finite samples, Hansen (2005) also provides a lower and upper bound for the SPA p-values. These p-values can be obtained by using the estimates $\hat{\mu}_k^l = \min(\bar{d}_k, 0)$ and $\hat{\mu}_k^u = 0$ respectively, where the latter yields a distribution under the null based on the LFC principle.[5] Hansen (2005) also provide a detailed description of the bootstrap scheme used to obtain the distribution under the null hypothesis.

Clearly, in many applications the choice of a benchmark may not be obvious or an objective benchmark may not exist. Other applications will not generally yield a single model that is significantly superior to all the competitors because, especially when the set of competing models is large, the data may not be sufficiently informative to give a

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

univocal answer. In these cases, the forecaster may aim to reduce the set of competing models to a smaller set that is guaranteed to contain the best forecasting model at a given confidence level by comparing all models with each other. This approach is known as multiple comparison without control and differs from the techniques discussed above for two reasons. First, the procedure does not require a benchmark to be specified. Second, the testing procedures generally rely on simple hypotheses, i.e., equalities.

Hansen, Lunde, and Nason (2010a) construct a sequential test of equal predictive ability, dubbed model confidence set (MCS), which given an initial set of forecasts allows to: *i)* test the null that no forecast is distinguishable from any other, *ii)* discard any inferior forecasts if they exist, *iii)* characterize the set of models that are (equivalent to each other and) superior to all the discarded models. The set of surviving model is called model confidence set and can be interpreted as a confidence interval for the forecasts in that it is the set containing the best forecast at some confidence level.

Designed around the testing principle of Pantula (1989) to ensure that sequential testing does not affect the overall size of the test, the MCS test involves a sequence of tests for equal predictive ability. Given an initial set of forecasts, M^0, the starting hypothesis is that all models in M^0 have equal forecasting performances. The relative performance of each pair of forecasts is measured by $d_{k,j,t} = L(\sigma_t, h_{k,t}) - L(\sigma_t, h_{j,t})$, for all $k, j \in M^0,\ k \neq j$. Under the assumption that $d_{k,j,t}$ is stationary, the null hypothesis of equal predictive ability takes the form

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

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

Let $\mathbf{L}_t$ be the $(m \times 1)$ vector of sample performances $L(\sigma_t, h_{k,t})$, $k \in M$ and $\iota_\perp$ the $(m \times (m-1))$ orthogonal complement of a m-dimensional vector of ones, where m is the dimension of M. Then, the vector $\iota_\perp' \mathbf{L}_t$ can be viewed as $m-1$ relevant contrasts as each element can be obtained as a linear combination of $d_{k,j,t},\ k, j \in M,\ k \neq j$ which has mean zero under the null (1.29). Hence, (1.29) is equivalent to $\mathrm{E}[\iota_\perp' \mathbf{L}_t] = 0$ and, under strict stationarity of $d_{k,j,t}$, it holds that $T^{-1/2} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t$ is asymptotically Normal with mean 0 and covariance matrix $\Omega = \lim_{t\to\infty} \mathrm{Var}\left(T^{-1/2} \sum_{t=1}^{T} \iota_\perp' \mathbf{L}_t\right)$. Thus, it seems natural to

employ traditional quadratic-form type tests as

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

and

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

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

In empirical applications, when m is large, it might be challenging to obtain a sensible estimate of Ω. As an alternative, Hansen, Lunde, and Nason (2010a) also propose three simpler statistics which only require the estimation of the diagonal elements of Ω. The cost is that, depending on nuisance parameters, the distribution under the null becomes non-standard. However, Hansen, Lunde, and Nason (2010a) provide a detailed description of the bootstrap scheme employed to solve the nuisance parameter problem and to obtain the distribution under the null hypothesis. The three statistics are expressed as functions of studentized quantities.

The first statistic is expressed as a sum of deviations from the common average (hence the subscript D). Under the null hypothesis $H^0 = \text{E}[\bar{d}_k] = 0 \; \forall k \in M$ the statistic takes the form[6]

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

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

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

respectively, where $t_{k,j} = \sqrt{T}\bar{d}_{k,j} / \sqrt{\hat{\omega}_s^R}$, $k, j = 1, ..., m$ $k \neq j$ and $s = 1, ..., m(m-1)$ and

[6] Note that the null hypothesis is equivalent to (1.29).

the variances $\hat{\omega}_s^R$ are consistent estimators of $\omega_s^R = lim_{t\to\infty}\text{Var}(\sqrt{T}\bar{d}_{k,j})$.

If the null hypothesis is rejected, then Hansen, Lunde, and Nason (2010a) suggest the use of the following elimination rule $\mathcal{E}_M = \arg\max_{k\in M} t_k$ which excludes the model with the largest standardized excess loss relative to the average across models. The iterative testing procedure ends as soon as there is the first non rejection, or obviously if all forecast but one have been recursively eliminated. Finally, the MCS p-value is equal to $p_i = \max(p_{i-1}, p(i))$, $i = 1, ..., m$, where p_i is the p-value of the test under the null hypothesis $H^0_{M^i}$, i.e., at the ith step of the iteration process. By convention the p-value equals one when there is only one surviving model.

1.8 Consistency of the ordering and inference on forecast performances

In this section we illustrate, using a Monte Carlo simulation, to what extent the latent variable problem induces distortions in the ranking and affects the inference on forecast accuracy. We focus on univariate volatility models, whereas a similar exercise based on the comparison of multivariate models is presented in Chapter 2.

The forecast performances are measured by the following two loss functions

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

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

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

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

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

expectation of the second order Taylor expansion of L_{LMSE} around the true value σ_t

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

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

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

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

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

Since, apart from coincidental cancelation, $L^{''}_{LMSE}(\sigma_t, h_{k,t}) \neq L^{''}_{LMSE}(\sigma_t, h_{j,t})$, then the order implied by the proxy is likely to differ from the one implied by the true variance and the bias in the ranking is more likely to appear as the quality of the proxy deteriorates. On the other hand, the true ordering is likely to be preserved as the proxy becomes nearly perfect, i.e., $\mathrm{E}[\eta_t^2 \mid \Im_{t-1}] \to 0$.

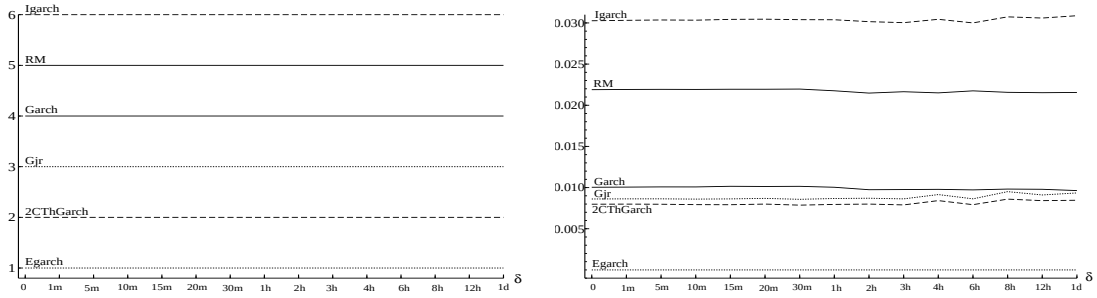
We generate artificial data from an Exponential GARCH(0,1) diffusion, see Nelson (1990) for details, that is

$$\begin{aligned}\begin{bmatrix} \mathrm{d}p(t) \\ \mathrm{d}\log(\sigma(t)) \end{bmatrix} &= \begin{bmatrix} 0 \\ -0.1 - 0.05\log(\sigma_(t)) \end{bmatrix}\mathrm{d}t \\ &\quad + \begin{bmatrix} \sigma_(t) & -0.1\sqrt{\sigma(t)} \\ -0.1\sqrt{\sigma(t)} & 0.01 + 0.04(1 - 2/\pi) \end{bmatrix}^{1/2} \begin{bmatrix} \mathrm{d}W_1(t) \\ \mathrm{d}W_2(t) \end{bmatrix}, \quad (1.34)\end{aligned}$$

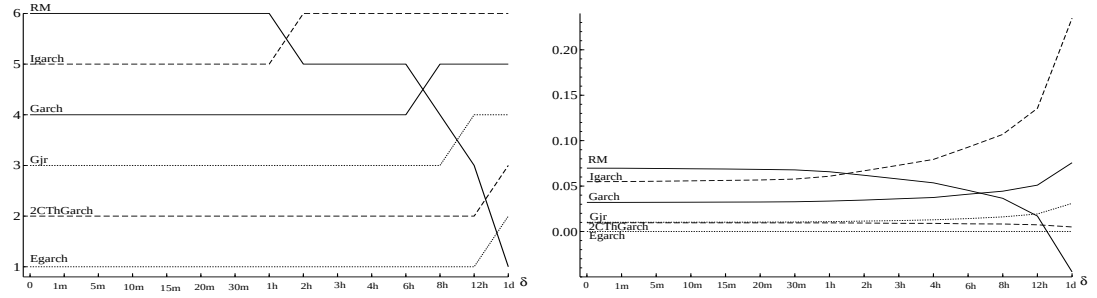
where $\mathrm{d}W_i(t)$, $i = 1, 2$ are two independent Brownian motions. The simulation is based on 500 replications. Using an Euler discretization scheme of (1.34), we approximate the continuous time process by generating 7200 observation per day. All the competing models are estimated by QMLE using data aggregated at daily frequency and one step ahead forecast computed. The estimation sample size amounts to 1500 daily observation while 1000 daily observation are used for the one-step ahead forecasts evaluation. The set of

competing models includes, Exponential (Egarch) (Nelson, 1991), Garch (Bollerslev, 1986), Gjr (Glosten, Jagannathan, and Runkle, 1992), Integrated (Igarch) (Engle and Bollerslev, 1986), RM (J.P.Morgan, 1996) and 2-Components Threshold Garch (2CThGarch) (**?**, **?**) models. The latent variance is computed as $\sigma_t = \int_{t-1}^{t} \sigma(u)\mathrm{d}u$, $t \in \mathbb{N}$. The proxy is the realized variance of Andersen and Bollerslev (1998), i.e., the sum of intraday squared returns, and is computed using returns sampled at 14 different frequencies ranging from 1-minute to daily. The proxy is denoted $\hat{\sigma}_{t,\delta}$, where δ =1m,5m,...,1h,...,1d represents the sampling frequency. In this setting the realized variance estimator is conditionally unbiased, allows to control for the accuracy of the proxy (through the level of aggregation of the data δ) and it is also consistent, i.e., $\hat{\sigma}_{t,\delta} \underset{p}{\to} \sigma_t$ as $\delta \to 0$. The underlying ordering implied by a given loss function, wether it is robust or not, is identified by ranking forecasts with respect to the true variance, σ_t (denoted as $\delta = 0$ in Figure 1.1).

Figure 1.1(a) represents the ranking based on the average sample performances (over the 500 replications) implied by the robust loss function, L_{MSE}, for the true variance ($\delta = 0$) and various levels of precision for the proxy (δ =1m to δ =1d). The ranking appears stable



a) L_{MSE} loss function

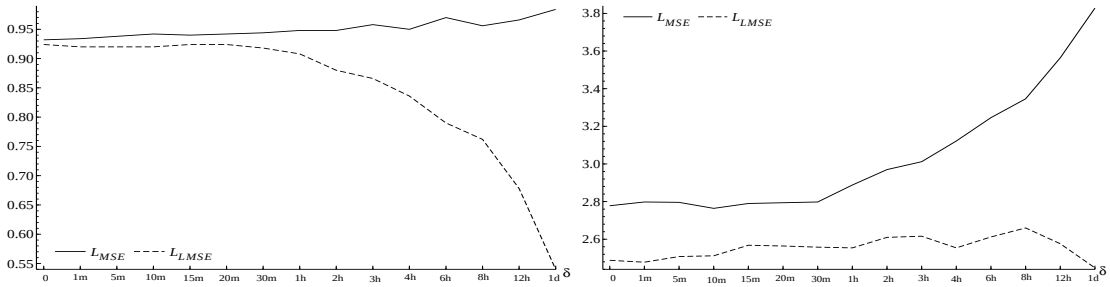


b) L_{LMSE} loss function

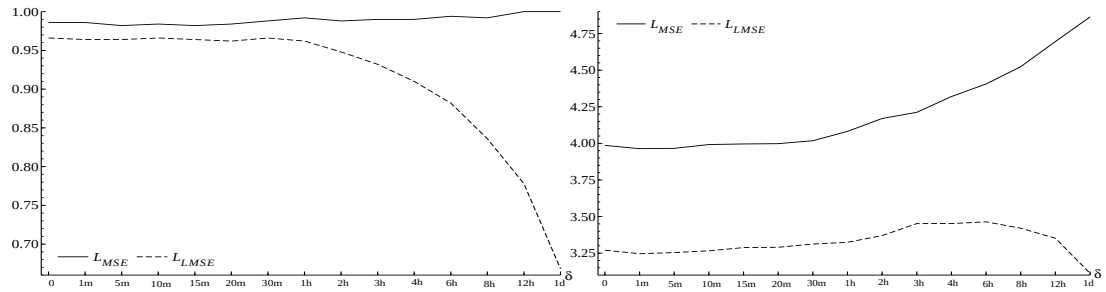
Figure 1.1: Ranking implied by L_{MSE} and L_{LMSE}. Ranking based on avg. performances (left) and avg. loss differentials from Egarch (right).

and loss differentials between models remain constant independently of the level of accuracy

of the proxy. Thus, the ranking obtained under $\hat{\sigma}_{t,\delta}$ is consistent for the one under the true, latent, conditional variance σ_t, for all values of δ.



a) $\alpha = 0.25$



b) $\alpha = 0.10$

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

When considering the non-robust loss function, L_{LMSE}, the appearance of the objective bias becomes striking. In fact, although the consistency of the proxy ensures convergence of the proxy-based ordering to the true one as $\delta \to 0$, which is the case when the ranking is based on $\hat{\sigma}_{t,\delta}$ computed using returns sampled at frequency higher than 1-hour (Figure 1.1(b)), as the quality of the proxy deteriorates inferior models emerge. The relative performances of inferior models begin to improve rapidly and we observe major distortions at all levels of the ranking. For instance, the RM model, which ranks last when using the true variance, raises to the top of the classification when the proxy used in the evaluation is the squared return ($\delta = 1$d).

We now compare the forecast performances of our set of models using the MCS test. Ideally the MCS, i.e., the set of superior models, should be a singleton containing the true process, i.e., the Egarch. However, as the quality of the proxy deteriorates, losses and thus loss differentials, become less informative, which in turn make more difficult to effectively discriminate between models. Consequently we expect the set of superior models to grow

in size as δ increases. Figure 1.2 reports two statistics, the frequency at which the Egarch is in the MCS, which shows the size properties of the test (left) and the average number of models in the MCS, which is informative about the power properties of the test (right). As before, the results are reported as a function of the precision of the proxy, δ. The levels of confidence considered are $\alpha = [0.25, 0.10]$. The statistic considered is the $MCS - T_D$ in (1.32). The number of bootstrap samples used to obtain the distribution under the null is set to 1000.

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

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

1.9 Thesis overview and motivation

In this introductory chapter we reviewed a variety of methods for volatility forecast evaluation and comparison which provide the theoretical background of this thesis. As mentioned in Section 1.5, our main interest is in the multivariate dimension. In Chapter 2 we focus on the problems arising due to the latent nature of the conditional variance in a multivariate framework. We address these issues by investigating the properties of the ranking between multivariate volatility forecasts with respect to alternative statistical loss functions used to evaluate forecast performances. We provide conditions on the functional form of the loss function that ensure consistency between the proxy-based ranking and the true, but unobservable one. We identify a large set of loss functions that yield a consistent ranking. We illustrate our findings using artificial data sampled from a continuous time multivariate diffusion process and compare the ordering delivered by both robust and non-robust loss

functions over different forecast horizons. We further discuss the sensitivity of the ranking to the quality of the proxy and the degree of similarity between models. An application to three foreign exchange rates, where we compare the forecasting performance of 24 multivariate GARCH specifications over two forecast horizons, concludes the chapter.

In Chapter 3 we address the question of the selection of multivariate GARCH models in terms of forecast accuracy with a particular focus on relatively large scale problems. We consider 10 assets from NYSE and NASDAQ and compare 125 model based one-step-ahead conditional variance forecasts over a period of 10 years using the MCS and the SPA tests. Model performances are evaluated using four statistical loss functions which account for different types and degrees of asymmetry with respect to over/under predictions. When considering the full sample, MCS results are strongly driven by short periods of high market instability during which multivariate GARCH models appear to be rather inaccurate. Over relatively unstable periods, i.e. dot-com bubble, the set of superior models is composed of more sophisticated specifications such as orthogonal and dynamic conditional correlation (DCC), both with leverage effect in the conditional variances. However, unlike the DCC models, our results show that the orthogonal specifications tend to systematically underestimate the conditional variance. Over calm periods, simple assumptions like constant conditional correlation and symmetry in the conditional variances cannot be rejected. Finally, during the 2007-2008 financial crisis, accounting for non-stationarity in the conditional variance process generates superior forecasts. The SPA test suggests that, independently from the period, the best models do not provide significantly better forecasts than the DCC model of Engle (2002) with leverage effect in the conditional variances of the returns.

In Chapter 4 we derive a class of diffusion approximations based on conditional correlation models. To our knowledge, this chapter represents a first attempt to address the relationship between multivariate discrete and continuous time models, and in particular to conditional correlation models. We consider a modified version of the standard DCC model, the consistent DCC (cDCC) model of Aielli (2006). This model is particularly appealing because it is based on a more natural representation of the process driving the correlation which, unlike the standard DCC model, preserves the martingale difference property. For this specification, we point out the existence of a degenerate diffusion limit. The degeneracy of the cDCC-GARCH diffusion limit is due to the particular structure of the discrete time model in which the noise propagation system of the variances and that of the process driving the correlation are perfectly correlated. This structure is preserved in the diffusion limit which is characterized by a singular diffusion matrix. More precisely, the diffusion of the variances and that of the diagonal elements of the process driving the correlation are pairwise governed by the same Brownian motion. We also consider, as a special case,

the Constant Conditional Correlation (CCC) model, which can be obtained by imposing suitable parameter restrictions to the cDCC model. In this case, we are able to recover a non-degenerate diffusion. Finally, we propose different sets of conditions regarding the speed of convergence of the parameters of the cDCC-GARCH model which allow to recover other types of degenerated diffusion limits, characterized by a stochastic price process while variances and/or correlations remain time varying but deterministic. We also elaborate on the type of models can be obtained as Euler approximation of the different diffusions. Our convergence results are validated through a comprehensive Monte Carlo simulation exercise.

Chapter 2

On Loss Functions and Ranking Multivariate Volatility Forecasts[1]

2.1 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 the forecasting performance of volatility models raises a problem. Since the variable of interest (i.e., volatility) is unobservable, the evaluation of the loss function has to rely on a proxy. However, this substitution may not always lead to the same ordering between models' performances as if the true volatility was observable. The impact on the ordering of the substitution of the true volatility by a proxy has been investigated for univariate models by Hansen and Lunde (2006a). They provide conditions, for both the loss function and the volatility proxy, under which the ranking based on the proxy is consistent for the true ranking. Starting from this result, Patton (2009) and Patton and Sheppard (2009) derive necessary and sufficient conditions on the functional form of the loss function for the ranking to be robust to the presence of noise in the proxy. These results have important implications on testing procedures for superior predictive ability,

[1] *This chapter has been adapted from Laurent S., Rombouts J.V.K. and Violante F. (2009), On Loss Functions and Ranking forecasting Performances of Multivariate Volatility Models. CIRANO dp 2009-45*

see Diebold and Mariano (1995), West (1996), Clark and McCracken (2001), the reality check by White (2000) and the recent contributions of Hansen and Lunde (2005) with the superior predictive ability (SPA) test and Hansen, Lunde, and Nason (2010a) with the Model Confidence Set test, among others, because 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 forecast performances little is known about the properties of the loss functions. This is the first work that addresses this issue.

In this chapter, we unify and extend the results in the univariate framework to the evaluation of multivariate volatility models, that is the comparison and ordering of sequences of variance matrices. From a methodological viewpoint, we first extend 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 denote such loss functions as robust. Contrary, non-robust loss functions denote those loss functions which ensures consistency of the ordering under an imperfect volatility proxy. Second, we derive necessary and sufficient conditions on the functional form of the robust loss function. We focus on homogeneous statistical loss functions that can be expressed as sample means of each period loss. Although conditions established in Hansen and Lunde (2006a) guarantee that the true conditional variance will be chosen (subject to sampling variation) over any other alternative regardless of the choice units, it does not guarantee that the ranking of two imperfect forecasts will be invariant to the choice of units. Patton (2009) shows that by using a homogeneous robust loss function, the ranking of two imperfect forecasts is invariant to a re-scaling of the data. Third, we identify a number of vector and matrix loss functions, some of which often used in practice, and provide insights on their properties, interpretation and geometrical representation. Although we focus on homogeneous loss functions, unlike in the univariate case, a complete identification of the set of robust loss functions is not available. This is because in the multivariate case there is an infinite number of possible combinations of the elements of the forecasting error matrix which yield a loss function that satisfies the necessary and sufficient conditions. However, given the necessary and sufficient functional form, application specific loss functions can be easily derived.

Note that different loss functions may deliver different rankings depending on the characteristics of the data that each loss function is able to capture. We show that many commonly used loss functions do not satisfy the sufficient conditions for consistent ranking and therefore may suffer from the objective bias problem. However, these loss functions often have desirable properties (e.g., down weighting extreme forecast errors) which can be

useful in applications. We show that non-robust loss functions are not *per se* inferior. In fact, provided that the proxy is sufficiently accurate with respect to the degree of similarity between models' performances, these loss functions can still deliver a ranking that is insensitive to the noise of a proxy. With respect to terminology, consistency of the ranking does not mean invariance of the ordering with respect to the choice of the loss function, see for instance Jensen (1984), but is intended with respect to the accuracy of the proxy for a given loss function. To make our theoretical results concrete, we focus on multivariate GARCH models to forecast the conditional variance matrix of a portfolio of financial assets.

We illustrate using artificial data, through a comprehensive Monte Carlo simulation, the impact of the deterioration of the quality of the proxy on the ranking of multivariate GARCH models with respect to different choices for the loss function and forecast horizons. In line with our theoretical results, when ranking over a discrete set of volatility forecasts, robust loss functions ensure consistency of the ranking to noise in the volatility proxy. On the other hand, non-robust loss functions allow to identify the underlying ordering only when the quality of the proxy is sufficiently good relative to the degree of similarity between models' performances. Furthermore, the ordering becomes less sensitive to the noise in the proxy as the forecast horizon increases.

We illustrate our findings using three exchange rates (Euro, UK pound and Japanese yen against US dollar). We consider 24 multivariate GARCH specifications which are widely used in practice and two forecast horizons. The advantage of choosing a robust loss function to evaluate model performances is striking. The ranking based on a non-robust 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. To test statistically which set of models forecasts better multivariate exchange rate volatility, we apply the model confidence test approach of Hansen, Lunde, and Nason (2010a). This approach depends on the orderings implied by the loss function. We show that even if the testing procedure is formally valid, anon-robust loss function can result in an incorrect identification of the set of superior models. The results also clearly demonstrate the value of high precision proxies. In fact, while robustness of the loss function ensures consistency of the ordering, only a high precision proxy allows to efficiently discriminate between models.

The rest of the chapter is organized as follows. In Section 2.2 we provide conditions for consistency of the ranking and derive the admissible functional form of the loss function. We derive a number of specific parameterizations and discuss their properties, interpretation and geometrical representation. In Section 2.3 we provide a brief overview of the multivariate GARCH specifications considered in this chapter. In Section 2.4, we illustrate, using

artificial data, the practical implications of our theoretical results in a simulation based comparison of multivariate GARCH models in a realistic setting. In Section 2.5 we present an empirical application using three exchange rates. In Section 2.6 we conclude and discuss directions for further research. All proofs are provided in the Appendix A. A number of examples are discussed in Appendix B.

2.2 Consistent ranking and distance metrics

As explained in Andersen, Bollerslev, Christoffersen, and Diebold (2006), the problem when comparing and ranking forecasting performance of volatility models is that the true conditional variance is unobservable. Although the problem can be solved by replacing the latter by a proxy, the substitution may not always lead to the same ranking as if the true conditional variance was observed. Hansen and Lunde (2006a) provide a theoretical framework for the analysis of the ordering of stochastic sequences and identify conditions that a loss function has to satisfy to deliver an ordering consistent with the true one when a proxy for the conditional variance is used. Their results can be cast in the more general framework of invariant decision rules, see Ferguson (1967) for example. Patton (2009) derives necessary and sufficient conditions on the functional form of the loss function for the latter to order consistently in presence of noise in the proxy.

In this section, we extend and unify these results to the case of multivariate volatility models, which requires the comparison and ordering of sequences of variance matrices. We first set the notation, working assumptions and basic definitions and, as an example, we present a set of loss functions commonly used to measure models' performances in a multivariate volatility context. Second, we discuss sufficient conditions for consistent ranking. Third, we characterize the functional form of a robust loss function. Fourth, we illustrate how robust loss functions can be constructed in practice and we discuss some special cases.

2.2.1 Notation and definitions

We first fix the notation and make explicit what we mean by a well defined loss function and by consistent ranking. For N time series at time t we denote $\mathbb{R}_{++}^{N\times N}$ the space of $N \times N$ 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 element indexed by m, $H_{m,t}$, such that $H_{m,t} \in \dot{H}$. The matrix $\Sigma_t \in \mathbb{R}_{++}^{N\times N}$ denotes the true but unobservable conditional variance matrix and $\hat{\Sigma}_t$ a proxy. Note that $H_{m,t}$ and $\hat{\Sigma}_t$ are variance matrices and therefore are symmetric. 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, $\mathrm{E}[L(\Sigma_t, H_{m,t})]$ as the true ordering. Similarly, $L(\hat{\Sigma}_t, H_{m,t})$ is the loss with respect to the proxy $\hat{\Sigma}_t$, and $\mathrm{E}[L(\hat{\Sigma}_t, H_{m,t})]$ 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., $T^{-1}\sum_t L(\hat{\Sigma}_t, H_{m,t})$, where T is the length of the forecast sample. 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 on $\dot{H}$ and it is uniquely minimized at H_t^*. If $H_t^* \in int(\dot{H})$, $L(\cdot,\cdot)$ is convex in $\dot{H}$. H_t^* represents the optimal forecast

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

$$H_t^* = \underset{H_t \in \dot{H}}{\arg\min} L(\Sigma_t, H_t) \Leftrightarrow H_t^* = \Sigma_t. \tag{2.1}$$

Definition 1. *Under Assumptions A1.1 to A1.2, the loss function is well defined.*

Without loss of generality, throughout the chapter we normalize the loss function so that it implies zero loss when the forecast error is zero, i.e. $L(\Sigma_t, H_t) = 0 \Leftrightarrow H_t = \Sigma_t$. This allows to interpret the loss function as a distance.

The notion of consistency of ranking is defined as follows:

Definition 2. *Consistency between the true ranking and the ordering based on a proxy is achieved if*

$$E(L(\Sigma_t, H_{l,t})) \geq E(L(\Sigma_t, H_{m,t})) \Leftrightarrow E(L(\hat{\Sigma}_t, H_{l,t})) \geq E(L(\hat{\Sigma}_t, H_{m,t})) \tag{2.2}$$

is true for all $l \neq m$, where $L(\cdot,\cdot)$ is a well defined loss function in the sense of Definition 1 and $\hat{\Sigma}_t$ is some conditionally unbiased proxy of Σ_t. A loss function that satisfies (2.2) is denoted as "robust".

By Definition 2, the ranking between any two models indexed by l and m, is consistent if it is the same whether it is based on the true conditional variance matrix or a conditionally unbiased proxy. Note that conditional unbiasedness is the only assumption we will make about the covariance proxy.

As underlined in Patton (2009) it is common practice to use several alternative measures of forecast accuracy to respond to the concern that some particular characteristics of the data may affect the result. We discuss next a selection of loss functions commonly used to evaluate multivariate models' forecast accuracy, or, in a more general context, to measure the distance between matrices and vectors and provide their classification. Examples can be found in Ledoit and Wolf (2003), James and Stein (1961), Bauwens, Lubrano, and Richard (1999), Koch (2007) and Herdin, Czink, Ozcelik, and Bonek (2005). Although the loss function listed in Table 2.1 are in principle well suited to measure variance forecast performances, it turns out that several are inappropriate in this setting.

Table 2.1: Loss functions and their classification

Matrix loss functions			
L_F	Frobenius distance	$\sum_{1\leq i,j\leq N}(\sigma_{i,j,t}-h_{i,j,t})^2$	robust
L_S	Stein distance	$\mathrm{Tr}[H_t^{-1}\Sigma_t]-\log\left\lvert H_t^{-1}\Sigma_t\right\rvert - N$	robust
L_{1M}	Entrywise 1 - (matrix) norm	$\sum_{1\leq i,j\leq N}\lvert\sigma_{i,j,t}-h_{i,j,t}\rvert$	non-robust
L_{PF}	Proportional Frobenius dist.	$\mathrm{Tr}[(\Sigma_t H_t^{-1}-I)^2]$	non-robust
$L_{LF,1}$	Log Frobenius distance (1)	$\left(\log\left\lvert\Sigma_t H_t^{-1}\right\rvert\right)^2$	non-robust
$L_{LF,2}$	Log Frobenius distance (2)	$\left(\log\frac{\mathrm{Tr}[\Sigma_t\Sigma_t]}{\mathrm{Tr}[H_t H_t]}\right)^2$	non-robust
L_{Cor}	Correlation distance	$1-\frac{\mathrm{Tr}(\Sigma_t H_t)}{\sqrt{\mathrm{Tr}(\Sigma_t\Sigma_t)\mathrm{Tr}(H_t H_t)}}\in[0,1]$	non-robust
Vector loss functions			
L_E	Euclidean distance	$\sum_{1\leq k\leq N(N+1)/2}(\sigma_{k,t}-h_{k,t})^2$	robust
L_{WE}	Weighted Euclidean distance (with matrix of weights W)	$(\sigma_t-h_t)'W(\sigma_t-h_t)$	robust
L_{1V}	Entrywise 1 - (vector) norm	$\sum_{1\leq k\leq N(N+1)/2}\lvert\sigma_{k,t}-h_{k,t}\rvert$	non-robust

The first loss function, L_F, is the natural extension to matrix spaces of the mean squared error (MSE). The second, L_S, is the scale invariant loss function introduced by James and Stein (1961). L_{1M} represents the extension to matrix spaces of the mean absolute deviation (MAD) and is known as the entrywise 1 - (matrix) norm. L_{PF} is the extension of the

heteroskedasticity adjusted MSE and is a quadratic loss function with the same parametric form of the Frobenius distance but measuring deviations in relative terms. We refer to this loss function as proportional Frobenius distance. $L_{LF,1}$ and $L_{LF,2}$ are adaptations of the MSE logarithmic scale. In particular, the loss function in $L_{LF,2}$, alternatively defined as $\left(\log\left[\left(\sum_i \lambda_i^2(\Sigma_t)\right)\left(\sum_i \lambda_i^2(H_t)\right)^{-1}\right]\right)^2$, considers the singular values as a summary measure of a matrix. The sum of squared singular values (i.e., $\sum_i \lambda^2(A)_i = \mathrm{Tr}(AA')$) represents the Frobenius distance of Σ_t and H_t from 0. The ratio measures the discrepancy in relative terms while the logarithm ensures that deviations are measured as factors and the squaring ensures that factors are equally weighted (Moskowitz, 2003). We refer to this loss function as log Frobenius distance. L_{Cor} is also based on the Frobenius distance but it exploits the Cauchy-Schwarz inequality. In fact, by the inequality, the ratio is equal to one when $H_t = \Sigma_t$ and tends to 0 if H_t and Σ_t differ to a maximum extent. The ratio resembles a correlation coefficient between the matrices H_t and Σ_t. L_E is the Euclidean distance computed on all unique elements of the forecast error matrix, while L_{WE} is a weighted version of L_E. The last function, L_{1V}, represents an extension of the MAD, but defined on a vector space. It differs from L_{1M} for equally weighting the unique elements of the forecast error matrix.

2.2.2 Conditions for consistent ranking of multivariate volatility models

We provide sufficient conditions that a loss function has to satisfy to deliver asymptotically the same ordering whether the evaluation is based on the true conditional variance matrix or a proxy. Since Σ_t and H_t are variance matrices and hence symmetric, without loss of generality we can redefine the function $L(\cdot,\cdot)$ from the space $\mathbb{R}_{++}^{N\times N} \times \dot{H}$ to $\mathbb{R}_+$ as a function of all unique elements of Σ_t and H_t, i.e., $\mathbb{R}^{N(N+1)/2} \times \dot{\mathcal{H}} \to \mathbb{R}_+$, with $\mathrm{vech}(H_{m,t}) \in \dot{\mathcal{H}}$, $\mathrm{vech}(\Sigma_t) \in \mathbb{R}^{N(N+1)/2}$ and $\dot{\mathcal{H}} \subset \mathbb{R}^{N(N+1)/2}$. This simplification allows to ignore $N(N-1)/2$ redundant first order conditions in the minimization problem defined in (2.1). We make use of the following assumptions:

A2.1 $L(\Sigma_t, H_t)$ and $L(\hat{\Sigma}_t, H_t)$ have the same parametric form $\forall H_t \in \dot{H}$ so that uncertainty depends only on $\hat{\Sigma}_t$.

A2.2 Σ_t and H_t are $\Im_{t-1}$ measurable.

A2.3 $L(\cdot,\cdot)$ is twice continuously differentiable with respect to $\hat{\sigma}_t$ and h_t.

A2.4 $\xi_t = (\hat{\sigma}_t - \sigma_t)$ is a vector martingale difference sequence with respect to $\Im_{t-1}$ with finite conditional variance matrix $V_t = \mathrm{E}_{t-1}[\xi_t \xi_t']$.

Assumption A2.1 and A2.2 imply that Σ_t and H_t are considered as observable, though model forecasts can be biased or inaccurate in any way (e.g., due to parameter uncertainty,

misspecification, etc.). Assumption A2.4 requires conditional unbiasedness of the covariance proxy. Proposition 2.1 states a sufficient condition on the loss function to ensure consistency of the ranking in presence of noise.

Proposition 2.1. *Under Assumptions A2.1 to A2.4, a well defined loss function in the sense of Definition 1 with* $\frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_{l,t} \partial \sigma_{m,t}}$ *finite and independent of* H_t $\forall l, m = 1, ..., N(N+1)/2$ *is robust in the sense of Definition 2.*

Proposition 2.1 applies for any conditionally unbiased proxy independently of its level of accuracy. The difference between the true and the approximated ordering which is likely to occur whenever Proposition 2.1 is violated, is denoted as the objective bias. The bias must not be confused with sampling variability, that is the distortion between the approximated and the empirical ranking. In fact, while the latter tend to disappear asymptotically (i.e., $T^{-1} \sum_t L(\hat{\Sigma}_t, H_t) \xrightarrow{p} E\left[L(\hat{\Sigma}_t, H_t)\right]$ under ergodic stationarity of $L(\hat{\Sigma}_t, H_t)$), the presence of the objective bias may induce the sample evaluation to be inconsistent for the true one irrespectively of the sample size. Note that, from the set of loss functions given in Table 2.1, it is straightforward to show that only L_F, L_S, L_E and L_{WE} satisfy Proposition 2.1. It is also clear that even simple transformations of a robust loss function may cause the violation of Proposition 2.1. An example is the well known Frobenius norm which is the square root of L_F.

We can further discuss the implications of Proposition 2.1 and elaborate on the case when Proposition 2.1 is violated. We show that the bias between the true and the approximated ranking depends on the accuracy of the proxy for the variance matrix: the presence of noise in the volatility proxy introduces a distortion in the approximated ordering, which tends to disappear when the accuracy of the proxy increases. More formally, consider a sequence of volatility proxies $\hat{\Sigma}_t^{(s)}$ indexed by s and denote $H_t^{*(s)}$ such that

$$H_t^{*(s)} = \underset{H_t \in int(\dot{H})}{\arg\min}\, \mathrm{E}_{t-1}[L(\hat{\Sigma}_t^{(s)}, H_t)]. \tag{2.3}$$

In a setting like for example Andersen, Bollerslev, Diebold, and Labys (2003), the index s can be thought of as the sampling frequency used to compute the covariance proxy.

Lemma 2.1. *Under Assumptions A2.1 to A2.3 and for* $\hat{\Sigma}_t^{(s)}$ *satisfying* $E_{t-1}[\xi_t^{(s)}] = 0$ $\forall s$ *(unbiasedness) and* $V_t^{(s)} = E_{t-1}[\xi_t^{(s)} \xi_t^{(s)\prime}]$ *finite for all s, with* $V_t^{(s)} \xrightarrow{p} 0$ *as* $s \to \infty$ *(consistency), then for a well defined loss function in the sense of Definition 1 it holds:*

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

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 first point states that, under Proposition 2.1, the optimal forecast is the conditional variance, and consistency is achieved regardless of the quality of the proxy. The second point in Lemma 2.1 shows that the distortion introduced in the ordering when using an non-robust loss function tends to disappear as the quality of the proxy, controlled through s, improves. Therefore, when ordering over a discrete set of models, for a loss function that violates Proposition 2.1, the more precise the proxy the less likely is the objective bias to appear. In other words, when the variance of the proxy is small with respect to discrepancy between any two models, the distortion that is likely to be induced by the proxy becomes negligible, leaving the ordering unaffected. In the simulation study in Section 2.4, we further investigate this issue and in particular investigate the relationship between the accuracy of the proxy (i.e., the variability of the proxy) and the degree of similarity between model performances (i.e., how close performances are). However, in practice, it may be difficult to determine ex-ante the degree of accuracy of a proxy. Since the trade off accuracy vs. similarity is difficult to quantify ex-ante, model comparison and selection based on non-robust loss function becomes unreliable and may lead to undesired results. The empirical application in Section 2.5 reveals that a sufficiently accurate proxy may not be available.

2.2.3 Functional form of the consistent loss function

In the univariate framework, Patton (2009) identifies necessary and sufficient conditions on the functional form of the loss function to ensure consistency in the sense of Definition 2. The set of robust loss functions relates to the class of linear exponential densities of Gourieroux, Monfort, and Trognon (1984) and partially coincides with the subset of homogeneous loss functions associated with the most important linear exponential densities. In fact, the family of loss functions with degree of homogeneity equal to zero, one and two defined in Patton (2009), can be alternatively derived from the objective functions of the Gaussian, Poisson and Gamma densities respectively (see Gourieroux and Monfort, 1995 page 244 for details).

We propose necessary and sufficient conditions on the functional form of the loss function defined such that it is well suited to measure distances in matrix and vector spaces. Although, unlike in the univariate case, a complete identification of the set of robust loss functions is not feasible, we are able to identify a large set of parameterizations which yield robust loss functions. We show that several well known vector and matrix distance functions also belong to this set. 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 $R_{++}^{N\times N}$.

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 $H_t \in \dot{H}$, $\left|\mathrm{E}_{t-1}\left[\frac{\partial^2 L(\hat{\Sigma}_t, H_t)}{\partial h_t}\Big|_{H_t=\Sigma_t}\right]\right| < \infty$ and $\left|\mathrm{E}_{t-1}\left[\frac{\partial L(\hat{\Sigma}_t, H_t)}{\partial h_t \partial h_t'}\Big|_{H_t=\Sigma_t}\right]\right| < \infty$ for all t where the last two inequalities hold element-wise.

Note that A3.2 follows directly from A1.2 and A2.4 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 A2.4. Assumption A3.3 allows to interchange differentiation and expectation, see L'Ecuyer (1990) and L'Ecuyer (1995) for details.

Proposition 2.2. *Under Assumptions A2.1 to A2.4 and A3.1 to A3.3 a well defined loss function, in the sense of Definition 1, is robust in the sense of Definition 2 if and only if it takes the form*

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)' vech(\hat{\Sigma}_t - H_t), \tag{2.4}$$

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

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

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.

An alternative expression for the loss function defined in Proposition 2.2 is provided in the following corollary.

Corollary 2.1. *Given $\hat{\Sigma}_t$ and H_t symmetric and positive definite, then the loss function specified in (2.4) is isometric to*

$$L(\hat{\Sigma}_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + Tr[\bar{C}(H_t)(\hat{\Sigma}_t - H_t)], \tag{2.5}$$

with $\tilde{C}(\cdot)$ defined as in Proposition 2.2 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.

Unlike in the univariate framework, the multivariate dimension offers a large flexibility in the formulation of the loss function, see Table 2.1 for several parameterizations. In applied work, a careful analysis of the functional form of the loss function is a crucial preliminary step to the selection based on the specific properties of a given loss function. In this respect, it is clear that Assumption A1.2 has a central role in this setting. It is interesting to elaborate on the case when A1.2 is dropped while keeping all other assumptions in place. We can show that a badly formulated loss function, still yields an ordering that is insensitive to the accuracy of the proxy, i.e. *apparently robust* but inherently invalid. In fact, the loss function would point to an optimal forecast that differs from the true conditional variance. To illustrate this, starting from the functional form defined in Proposition 2.2, we consider the following generalization of (2.5)

$$L(\Sigma_t, H_t) = \tilde{C}(H_t) - \tilde{C}(\Sigma_t) + f[\bar{C}(H_t)(\Sigma_t - H_t)], \tag{2.6}$$

assuming that there exists a linear map $f[\cdot] : \mathbb{R}^{N\times N} \to \mathbb{R}$ such that $L(\Sigma_t, H_t)$ satisfies second order conditions. We summarize the implications of relaxing Assumption A1.2 from Proposition 2.1, 2.2 and Lemma 2.1 in the following remark. The following statements are proved in the Appendix.

Remark 2.1. *Define $\succ$ the true ordering between variance matrix forecasts, i.e., based on the true conditional variance matrix, and $\succ_a$ the approximated ordering, i.e., based on the volatility proxy. Given the functional form in (2.6), if*

i) $f[\cdot] \equiv Tr[\cdot]$ (A1.2 is satisfied): $\succ$ and $\succ_a$ are equivalent in the sense of Definition 2 and $L(\Sigma_t, H_t)$ is such that $H_t^* = E(\hat{\Sigma}_t|\Im_{t-1}) = \Sigma_t$, i.e., the loss function is well defined in the sense of Definition 1;

ii) $f[\cdot] \not\equiv Tr[\cdot]$ (A1.2 is violated): $\succ$ and $\succ_a$ are equivalent in the sense that the substitution of the true covariance by a proxy does not affect the ordering. However, the ordering implied by $L(\Sigma_t, H_t)$ is ill-defined because $H_t^* \neq E(\hat{\Sigma}_t|\Im_{t-1}) = \Sigma_t$, i.e.,

the loss function points to an optimal forecast that differs from the true conditional variance irrespectively of the quality of the proxy.

The first part of Remark 1 reaffirms sufficiency and necessity of the functional form defined in Proposition 2.2. With respect to the second part, note that, under (2.6), the general idea of consistency of the ranking, is still valid. In fact, if $f[\cdot]$ is a linear map, then $f[\bar{C}(H_t)(\Sigma_t - H_t)]$ is linear in $\sigma_{i,j,t}$ $\forall i, j = 1, ..., N$, and therefore, similarly to what stated in Proposition 2.1, it holds that $\partial^3 L(\Sigma_t, H_t)/\partial\sigma_t\partial\sigma_t'\partial h_{k,t} = 0$ $\forall k = 1, ..., N(N+1)/2$. This result ensures the ranking based on the volatility proxy to be apparently robust for the one based on the true conditional variance and insensitive to the level of accuracy of the proxy, i.e., the objective bias does not represent an issue in this case. However, in absence of Assumption A1.2 the underlying ordering is invalid even when based on the true conditional variance, since an ill-defined loss function would point to an optimal forecast different from the true conditional variance.

2.2.4 Building a class of robust loss functions

Endowed with the functional form defined in Proposition 2.2, we illustrate how to recover several robust loss functions. These loss functions can be categorized with respect to different characteristics, for instance the degree of homogeneity, the shape, the underlying family of distributions or the functional form for $\tilde{C}(\cdot)$.

We start by investigating the case of loss functions that are based only on the forecast error, that is $L(\hat{\Sigma}_t, H_t) = L(\hat{\Sigma}_t - H_t)$. Patton (2009) shows that in the univariate case the MSE loss function is the only robust loss function that depends solely on the forecast error. The multivariate setting offers more flexibility in the functional form for a robust loss function based on the forecast error. The following proposition defines the family of such loss functions.

Proposition 2.3. *Under Assumptions A2.1 to A2.4 and A3.1 to A3.3, a loss function based only on the forecast error $\hat{\Sigma}_t - H_t$, robust in the sense of Definition 2, is defined by the quadratic form*

$$L(\hat{\Sigma}_t, H_t) = L(\hat{\Sigma}_t - H_t) = vech(\hat{\Sigma}_t - H_t)'\hat{\Lambda}vech(\hat{\Sigma}_t - H_t) \tag{2.7}$$

where $\hat{\Lambda}$ is a positive definite matrix of constants.

The loss function in (2.7) is homogeneous of degree 2 and is symmetric under a 180° rotation about the origin, i.e. $L(\hat{\Sigma}_t - H_t) = L(H_t - \hat{\Sigma}_t)$. The matrix $\hat{\Lambda}$ defines the weights assigned to the elements of the forecast error matrix $\hat{\Sigma}_t - H_t$.

Proposition 2.3 defines the entire family of quadratic loss functions, i.e. MSE type, which depends on the choice of the matrix of weights $\hat{\Lambda}$. Formally, the quadratic polynomial in (2.7) 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. The advantage 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. In this sense, the quadratic form in (2.7) is always symmetric under 180° rotation about the origin, but particular choices of $\hat{\Lambda}$ can generate some other types of asymmetries. In the following, we derive and discuss the properties of some well known loss functions belonging to the family defined by Proposition 2.3.

We provide next six examples of well known vector and matrix loss functions which satisfy Proposition 2.1 and 2.2. The first four examples below belong to the family defined by (2.7) and are introduced in increasing order of generality. The simplest parameterizations of $\hat{\Lambda}$ yield loss functions based on the vech() transformation of the forecast error matrix, i.e., based on the notion of distance on a vector space rather than a matrix space. These loss functions are typically the squares of norms, and therefore are homogeneous of degree 2. The last two examples belong to the more general form in (2.4) and are loss functions based on the notion of distance on a matrix space. In the Appendix B, we give interpretation, geometrical representation and numerical examples for each of these loss functions.

Example 1: Euclidean distance

From (2.7), by setting $\hat{\Lambda} = I_K$ we obtain a loss function of the form

$$L_E = (\hat{\sigma}_t - h_t)' I_K (\hat{\sigma}_t - h_t) = \sum_{1 \leq k \leq K} (\hat{\sigma}_{k,t} - h_{k,t})^2. \tag{2.8}$$

The loss function defined in (2.8) is the square of the Euclidean norm on the vech() transformation of the forecast error matrix $(\hat{\Sigma}_t - H_t)$. The matrix $\hat{\Lambda}$ is such that variances and covariances forecast errors are equally weighted.

Example 2: Weighted Euclidean distance

A more flexible version of (2.8) is the weighted Euclidean distance, where $\hat{\Lambda}$ is defined as $\hat{\lambda}_{i,i} > 0$ and $\hat{\lambda}_{i,j} = 0$, $i, j = 1, ..., K$, that is

$$L_{WE} = (\hat{\sigma}_t - h_t)' \hat{\Lambda} (\hat{\sigma}_t - h_t) = \sum_{1 \leq k \leq K} \hat{\lambda}_{k,k} (\hat{\sigma}_{k,t} - h_{k,t})^2. \tag{2.9}$$

This loss function allows to differently weight each variance and covariance forecast error.

Example 3: Pseudo Mahalanobis distance

This loss function represents a generalization of (2.9). It is obtained by setting $\hat{\lambda}_{i,j} \in \mathbb{R}$, $i, j = 1, ..., K$ and such that $\hat{\Lambda}$ is positive definite, that is

$$L_M = (\hat{\sigma}_t - h_t)'\hat{\Lambda}(\hat{\sigma}_t - h_t) = \sum_{1 \leq k,l \leq K} \hat{\lambda}_{k,l}(\hat{\sigma}_{k,t} - h_{k,t})(\hat{\sigma}_{l,t} - h_{l,t}), \tag{2.10}$$

with $\hat{\Lambda}$ chosen according to Proposition 2.3. Though sharing the same parametric form as the Mahalanobis distance, in this loss function the matrix of weights $\hat{\Lambda}$ is deterministic and does not depend on $(\hat{\sigma}_t - h_t)$. Since $\hat{\Lambda}$ is non diagonal, L_M also includes the cross product of variances and covariances forecast errors. The matrix $\hat{\Lambda}$ here plays a similar role to the correlation in a multivariate symmetric distribution: positive (negative) weights $\hat{\lambda}_{k,l}, k \neq l$ imply that systematic over/under predictions are penalized less (more).

Example 4: Frobenius distance

From (2.10), if we set $\hat{\Lambda}$ diagonal with $diag(\hat{\Lambda}) = \text{vech}(V)$ where V is symmetric with typical element, indexed by $i, j = 1, ..., K$, $v_{ij} = 1$ if $i = j$, $v_{ij} = 2$ if $i \neq j$, then the resulting loss function is

$$L_F = \text{Tr}[(\hat{\Sigma}_t - H_t)'(\hat{\Sigma}_t - H_t)]. \tag{2.11}$$

The loss function in (2.11) is the square of the Frobenius norm and represents the matrix equivalent of MSE loss function. Although it can be cast into Proposition 2.3, this loss function is based on the notion of distance on a matrix space. Alternatively, (2.11) can be written as $L_F = \sum_{1 \leq i,j \leq N} (\hat{\sigma}_{i,j,t} - h_{i,j,t})^2 = \sum_{1 \leq i \leq N} \varsigma_i(\hat{\Sigma}_t - H_t)$, where $\varsigma_i(\hat{\Sigma}_t - H_t), i = 1, \ldots, N$ are the singular values of the forecast error matrix $(\hat{\Sigma}_t - H_t)$. The loss function can be cast into Corollary 2.1 with the following parameterization: $\tilde{C}(H_t) = -\text{Tr}(H_t H_t)$ and $\bar{C}(H_t) = -2H_t$. The particular structure of the weights $\hat{\Lambda}$ implies that the covariance forecast error are double weighted. The Frobenius distance can also be obtained as the objective function associated to the matrix normal density, see Gourieroux and Monfort (1995) for details.

Alternatively, if we consider the Wishart distribution we identify a loss function that is characterized by a degree of homogeneity equal to zero and depending only on the standardized (in matrix sense) forecast error.

Example 5: Stein loss function

The objective function associated with the Wishart distribution is

$$L_S = \text{Tr}[H_t^{-1}\hat{\Sigma}_t] - \log \left| H_t^{-1}\hat{\Sigma}_t \right| - N. \tag{2.12}$$

L_S belongs to the family defined by (2.5) with $\tilde{C}(H_t) = \log |H_t|$ and $\bar{C}(H_t) = H_t^{-1}$. It

corresponds to the scale invariant loss function introduced by James and Stein (1961). L_S is asymmetric with respect to over/under predictions (in matrix sense), where underpredictions are heavily penalized.

Example 6: Patton and Sheppard (2009)
Consider $\tilde{C}(\cdot) = \text{Tr}(A^d)$ for some $d > 2$ and where A is symmetric and positive definite. Since the trace is a linear operator, the resulting loss function takes the form

$$L(\hat{\Sigma}_t, H_t) = \text{Tr}(\hat{\Sigma}_t^d) - \text{Tr}(H_t^d) - d\text{Tr}[H_t^{d-1}(\hat{\Sigma}_t - H_t)].$$

This loss function, introduced by Patton and Sheppard (2009), is homogeneous of degree d and asymmetric with respect to over/under predictions (in matrix sense), where overpredictions are heavily penalized.

We have seen that the multivariate dimensional case allows to construct a large variety of robust loss functions. However, unlike the univariate case where an analytical expression is available for the entire class of robust loss functions, in the multivariate case such generalization is unfeasible because there are many functions $\tilde{C}(.)$ that can be used to weight forecasts and forecast errors. However, given (2.4) or (2.5), application specific loss function can be easily derived by choosing ex-ante some functional form for $\tilde{C}(\cdot)$ and verifying on a case by case basis whether the resulting loss function satisfies Proposition 2.2. Note that, although robust, the resulting loss function can be rather difficult to interpret.

2.3 Competing multivariate GARCH models

The multivariate volatility models that we consider in this chapter belong to the multivariate GARCH (MGARCH) class. Consider an N-dimensional discrete time vector stochastic process r_t. Let $\mu_t = \text{E}(r_t|\Im_{t-1})$ be the conditional mean vector and $H_{m,t} = \text{E}(r_t r_t'|\Im_{t-1})$ the conditional variance matrix for model m so that we can write the model of interest as:

$$r_t = \mu_t + H_{m,t}^{1/2} z_t,$$

where $H_{m,t}^{1/2}$ is a $(N \times N)$ positive definite matrix and z_t is an independent and identically distributed random innovation vector with $\text{E}(z_t) = 0$ and $\text{Var}(z_t) = I_N$.

The MGARCH specifications considered in this chapter are: diagonal BEKK (D-Bekk) model (Engle and Kroner, 1995), multivariate RiskMetrics model (J.P.Morgan, 1996), Constant Correlation (CCC) model (Bollerslev, 1990), Dynamic Conditional Correlation (DCC) model (Engle, 2002a), the Orthogonal model (O) of Alexander (2000) and its generalized version (GO) (van der Weide, 2002). The univariate models used for the conditional vari-

ances of the marginal processes in the DCC, CCC, O and GO are: Garch (Bollerslev, 1986), Gjr (Glosten, Jagannathan, and Runkle, 1992), Exponential (Egarch) (Nelson, 1991), Asymmetric Power (Aparch) (Ding, Granger, and Engle, 1993), Integrated (Igarch) (Engle and Bollerslev, 1986), RiskMetrics (Rm) (J.P.Morgan, 1996) and Hyperbolic (Hgarch) (Davidson, 2004). The functional forms for H_t are briefly defined in Table 2.2. See Bauwens, Laurent, and Rombouts (2006) for further details.

Table 2.2: Multivariate GARCH specifications

Model	Multivariate GARCH models for H_t	# par.
D-Bekk	$H_t = C^{'}C + A^{'}\epsilon_{t-1}\epsilon_{t-1}^{\prime}A + G^{'}H_{t-1}G$	$\frac{(N+5)N}{2}$
RiskMetrics	$H_t = 0.04\epsilon_{t-1}\epsilon_{t-1}^{\prime} + 0.96H_{t-1}$	0
(G)O	$V^{-1/2}\epsilon_t = Lf_t$ $H_t = V^{1/2}LZ_tLV^{1/2}$ $Z_t = \mathrm{diag}(\sigma^2_{f_{1,t}}, \ldots, \sigma^2_{f_{m,t}})$ $L = P\Lambda^{1/2}U$ $U = I_N$ (O); $\quad U = \prod_{i<j} R_{i,j}(\delta_{i,j})$, $-\pi \leq \delta_{i,j} \leq \pi$ (GO)	0 (O) $\frac{N(N-1)}{2}$ (GO)
CCC	$H_t = D_tRD_t$ $D_t = \mathrm{diag}(h_{1,1,t}^{1/2} \ldots h_{N,N,t}^{1/2})$	$\frac{N(N-1)}{2}$
DCC	$H_t = D_tR_tD_t$ $R_t = \mathrm{diag}(q_{1,1,t}^{-1/2} \ldots q_{N,N,t}^{-1/2})Q_t\mathrm{diag}(q_{1,1,t}^{-1/2} \ldots q_{N,N,t}^{-1/2})$ $u_t = D_t^{-1}\epsilon_t$ $Q_t = (1-\alpha-\beta)\bar{Q} + \alpha u_{t-1}u_{t-1}^{\prime} + \beta Q_{t-1}$	$\frac{N(N-1)}{2} + 2$
Univariate GARCH models in Z_t and D_t		
Garch	$h_{l,t} = \omega_l + \alpha_l\epsilon^2_{l,t-1} + \beta_l h_{l,t-1}$	3
Egarch	$\log(h_{l,t}) = \omega_l + g(z_{l,t-1}) + \beta_l \log(h_{l,t-1})$ $g(z_{l,t-1}) = \theta_{l,1}z_{l,t-1} + \theta_{l,2}(\lvert z_{l,t}\rvert - \mathrm{E}(\lvert z_{l,t}\rvert))$	4
Gjr	$h_{l,t} = \omega_l + \alpha_l\epsilon^2_{l,t-1} + \gamma_l S^-_{l,t-1}\epsilon^2_{l,t-1} + \beta_l h_{l,t-1}$ $S^-_{l,t} = 1$ if $\epsilon_{l,t} < 0$; $S^-_{l,t} = 0$ if $\epsilon_{l,t} \geq 0$	4
Aparch	$h^{\delta_l}_{l,t} = \omega_l + \alpha_l[\lvert\epsilon_{l,t-1}\rvert - \gamma_l\epsilon_{l,t-1}]^{\delta_l} + \beta_l h^{\delta_l}_{l,t-1}$	5
Hgarch	$h_{l,t} = \omega_l[1-\beta_l]^{-1} + \lambda(L)\epsilon^2_{l,t}$ $\lambda(L) = \left\{1 - [1-\beta_l]^{-1}\alpha_l[1+\gamma_l(1-L)^d]\right\}$	5

All the specifications are characterized by a constant conditional mean and the models are estimated by quasi maximum likelihood. The sample log-likelihood is given (up to a

constant) by

$$-\frac{1}{2}\sum_{t=1}^{\mathcal{T}}\log \mid H_{m,t} \mid -\frac{1}{2}\sum_{t=1}^{\mathcal{T}}(r_t-\mu)^{'}H_{m,t}^{-1}(r_t-\mu), \tag{2.13}$$

where $\mathcal{T}$ is the size of the estimation sample. We maximize numerically for μ and the parameters in $H_{m,t}$. All calculations and results reported in this chapter are based on programs written by the authors using Ox version 6.0 (Doornik, 2009) and G@RCH version 6.0 (Laurent, 2009).

2.4 Simulation study

Using artificial data generated from a continuous time model we investigate the ranking of the MGARCH models with respect to three dimensions: the quality of the volatility proxy, the choice of the loss function and the forecast horizon. Continuous time models are a convenient framework to illustrate our theoretical results in a realistic setting. The Brownian semi-martingale assumption for the infinitesimal returns ensures that simple proxies based on intra-day data are unbiased and consistent estimators of true variance matrix. Since the proxy for the underlying volatility requires discretization and aggregation over different sampling frequencies, a discrete time counterpart of the continuous time process must exist to ensure identification of the optimal forecast.

2.4.1 MGARCH diffusion approximation and realized covariance

Let us assume the observed return vector to be generated by an N-dimensional log-price diffusion $\mathrm{d}p(t)$, $t \in \mathbb{R}_+$, and an $N(N+1)/2$-dimensional covariance diffusion, $\mathrm{d}\sigma(t)$, with $\sigma(t) = \mathrm{vech}(\Sigma(t)) = [\sigma_{ij}(t)]$ for $i,j = 1,...,N$, $i \geq j$. The diffusion process of the system admits the following Brownian semi-martingale representation

$$\begin{bmatrix} \mathrm{d}p(t) \\ \mathrm{d}\sigma(t) \end{bmatrix} = b(t)\mathrm{d}t + s(t)\mathrm{d}W(t), \tag{2.14}$$

with drift $b(t)$ locally bounded and measurable, diffusion matrix $a(t) = s(t)s(t)^{'}$ which can be partitioned as

$$\begin{bmatrix} \Sigma_{p,p}(t) & \Sigma_{p,\sigma}(t) \\ \Sigma_{p,\sigma}(t) & \Sigma_{\sigma,\sigma}(t) \end{bmatrix}, \tag{2.15}$$

and driven by a $N(N+3)/2$ vector of independent standard Brownian motions $W(t)$.

We consider the diffusion limit of the bivariate CCC-Egarch(0,1) model (see Table 2.2), which is derived following Nelson (1990), Strook and Varadhan (1979), Ethier and Kurtz (1986a) and Kushner (1984), see Chapter 4 for details. The model for variance of the marginal processes is expressed as a function of the log of $\sigma(t)$, while the constant correlation allows to compute the infinitesimal conditional covariance, at each point in time, as $\rho\sqrt{\sigma_1(t)\sigma_2(t)}$. The CCC-Egarch model admits a limit diffusion for the continuous time vector stochastic process $X_t = [p_1(t) \quad p_2(t) \quad \log\ \sigma_1(t) \quad \log\ \sigma_2(t)]'$ of the form introduced by (2.14) with drift and scale given respectively by

$$b(t) \;=\; \left[\begin{array}{cccc} \mu_1 & \mu_2 & (\omega_1 - \theta_1 \log\sigma_1(t)) & (\omega_2 - \theta_2 \log\sigma_2(t)) \end{array}\right]' \tag{2.16}$$

and

$$a(t) = \left[\begin{array}{cccc} \sigma_1(t) & \rho\sqrt{\sigma_1(t)\sigma_2(t)} & \alpha_1\sqrt{\sigma_1(t)} & \rho\alpha_2\sqrt{\sigma_1(t)} \\ \rho\sqrt{\sigma_1(t)\sigma_2(t)} & \sigma_2(t) & \rho\alpha_1\sqrt{\sigma_2(t)} & \alpha_2\sqrt{\sigma_2(t)} \\ \alpha_1\sqrt{\sigma_1(t)} & \rho\alpha_1\sqrt{\sigma_2(t)} & \alpha_1^2 + \gamma_1^2(1-2/\pi) & \rho\alpha_1\alpha_2 + \gamma_1\gamma_2 C \\ \rho\alpha_2\sqrt{\sigma_1(t)} & \alpha_2\sqrt{\sigma_2(t)} & \rho\alpha_1\alpha_2 + \gamma_1\gamma_2 C & \alpha_2^2 + \gamma_2^2(1-2/\pi) \end{array}\right], \tag{2.17}$$

where $C = \frac{2}{\pi}\left[\sqrt{1-\rho^2} + \rho\arcsin(\rho) - 1\right]$.

The process defined by (2.14), (2.16) and (2.17) ensures realistic dynamics for the return and the variance process and can be calibrated to real data parameters, see Nelson (1990) and Drost and Nijman (1993) for theoretical details, and Barndorff-Nielsen and Shephard (2004a) and Andersen and Bollerslev (1998) for examples. Furthermore, it allows to control for the nature and the size of the leverage effect and to preserve the correlation structure of the vector stochastic process X_t ensuring internal consistency of the model. Alternatively, (2.14) can be specified as any other continuous time stochastic volatility, e.g. models with a factor representation, as long as its discrete version is in the forecasting model set. See Dovonon, Meddahi, and Goncalves (2009), Barndorff-Nielsen and Shephard (2004a) and Voev and Lunde (2006) for related simulation designs.

In this simulation setup, the true variance is defined as the daily integrated covariance, $\Sigma_t = \int_{t-1}^{t} \Sigma_{p,p}(u)\mathrm{d}u$, $t \in \mathbb{N}$, where $\Sigma_{p,p}(u)$ is the infinitesimal volatility, see Barndorff-Nielsen and Shephard (2004c). A conditionally unbiased and consistent proxy is the realized covariance estimator proposed by Andersen, Bollerslev, Diebold, and Labys (2003) defined as $\hat{\Sigma}_{t,\delta} = \sum_{i=1}^{1/\delta} r_{t+i\delta-1} r'_{t+i\delta-1}$, where $r_{t+i\delta-1} = p_{t+i\delta-1} - p_{t+(i-1)\delta-1}$ and δ defines the intraday sampling frequency (i.e., the time span between two consecutive observations). The quality of the proxy is controlled through the level of aggregation of the data (i.e., the accuracy deteriorates as δ increases). Note that $\hat{\Sigma}_{t,\delta}$ satisfies the requirements in Lemma

2.1 with $s = 1/\delta$. See also Barndorff-Nielsen and Shephard (2004a), Hansen and Lunde (2006b) and Andersen, Bollerslev, Frederiksen, and Nielsen (2010) for details.

For the simulation study, we set, for $i = 1, 2$, $\mu_i = 0$ $\omega_i = -0.02$, $\theta_i = 1 - \beta_i = 0.03$, $\alpha_i = -0.09$, $\gamma_i = 0.4$ and $\rho = 0.9$ which ensures realistic dynamics for the return process. Our simulation results are based on 500 replications with an estimation sample of 2000 daily observations and a forecasting sample of 500 daily observations. We use an Euler discretization of (2.14) with $\Delta = 1/7200$, corresponding to 12-second returns. The integrated covariance is then computed as $\Delta \sum_{i=1}^{1/\Delta} \Sigma_i$, while the proxy $\hat{\Sigma}_{t,\delta}$ is computed using equally spaced intraday returns sampled at 14 different frequencies, ranging from 1 minute ($1/\delta = 1440$ intervals per day) to 24 hours ($1/\delta = 1$). The set of MGARCH models is estimated on daily returns and then recursive 1, 5, 10 and 20-step ahead forecasts are computed. Apart from the CCC-Egarch, the set of models includes the diagonal Bekk, RiskMetrics, CCC-Garch, CCC-Igarch, CCC-RiskMetrics, GO-Garch, GO-Egarch, GO-Igarch and GO-Hgarch (see Table 2.2). The models that we consider span a large variety of degrees of similarity between models.

2.4.2 Sample performance ranking and objective bias

Since we are ranking over a set of estimated volatility models, the true ranking implied by a given loss function, except for the optimal forecast, is not known ex-ante. The ranking of two imperfect volatility forecasts may differ between loss functions and it depends on how each specific loss function penalizes deviations from the target. The underlying ordering implied by a given loss function, either robust or non-robust, is identified by ranking forecasts with respect to the true covariance, Σ_t (denoted as $\delta = 0$ in Figures 2.1 and 2.2).

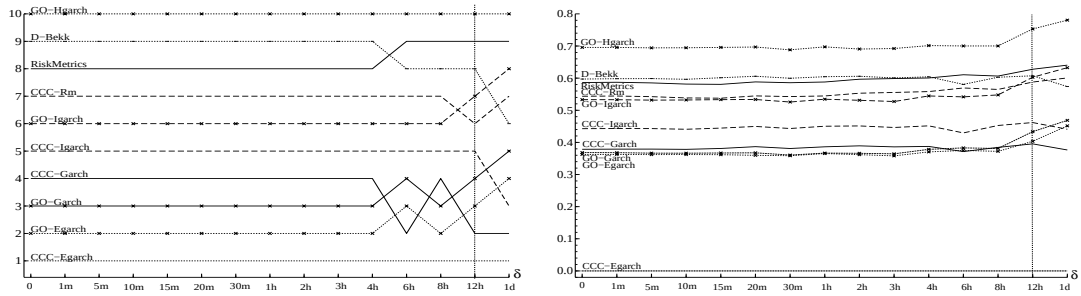
Without loss of generality, we consider only one robust (Frobenius distance, L_F) and one non-robust (entrywise 1 - matrix norm, L_{1M}) loss function. It is worth noting that, being non-differentiable L_{1M} does not directly fit our setting. However, in the univariate framework, Patton (2009) quantifies analytically the discrepancy between the optimal forecast obtained by minimizing the conditional expectation of mean absolute error (MAE) loss function, under different distributional assumption for the returns and considering different volatility proxies, and the perfect forecast (i.e., the one that is exactly equal to the true conditional variance). Since the loss function L_{1M} is a linear combination of MAE, this result extends directly. The choice of L_{1M} is not coincidental. In fact, as L_F it is a function of the forecast errors and it is somewhat comparable to L_F in terms of symmetry (i.e., over/under-prediction are equally weighted). Other loss functions in Table 2.1 have been investigated and give qualitatively the same results.

The vertical line in Figure 2.1 and 2.2 denotes the lowest sampling frequency that

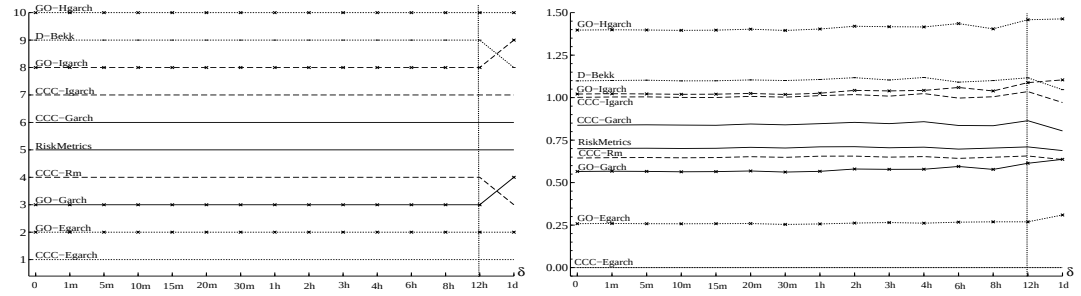
ensures positive definiteness of $\hat{\Sigma}_{t,\delta}$. For reference, we also report the evaluation based on $\hat{\Sigma}_{t,1d}$, which is singular by construction for $N = 2$. Figure 2.1(a-left) shows the ranking based on the average sample performances (over the 500 replications) implied by the robust loss function, L_F, for various levels of precision for the proxy (controlled through δ), when considering 1-step ahead forecasts. In line with Proposition 2.1, Figure 2.1(a-right) shows that loss differentials between models remain constant independently of the level of accuracy of the proxy. The CCC-Egarch model is correctly ranked first and minor shifts in position occur when the average sample performances are extremely close, with differences in the order of 10^{-2}.

A different picture emerges when considering the non-robust loss function, L_{1M}. As shown in Lemma 2.1, under the non-robust loss function, consistency of the proxy ensures convergence of the approximated ordering to the true one. This is the case when the ranking is based on $\hat{\Sigma}_{t,\delta}$ computed using returns sampled at frequency higher than 30-minutes (Figure 2.2(a-left)). As the quality of the proxy deteriorates inferior models seem to emerge. The relative performances of inferior models begin to improve rapidly starting from the 30-minute frequency. When the proxy is computed using lower frequency data, the objective bias starts to appear and we observe major distortions at most levels of the classification. The impact of the objective bias is amplified by the fact that except for the first two positions, i.e., CCC-Egarch and GO-Egarch respectively, all the other models exhibit extremely close average sample performances (Figure 2.2(a-right)). Although the objective bias does not become an issue when ordering between these two models, Figure 2.2(b-right) shows that, as the sampling frequency used to compute $\hat{\Sigma}_{t,\delta}$ lowers, the loss differential between these two models reduces. Since the variability of the loss increases with the variability of the proxy, the probability of ranking the GO-Egarch first in each replication increases at low frequencies. On the other hand, poorly performing models like CCC-Rm and RiskMetrics, 9th and 10th according to Σ_t, improve up to the 4th and 5th position respectively.

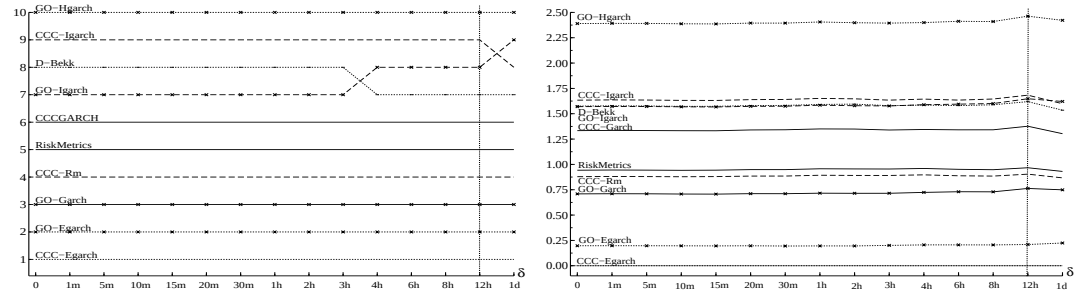
With respect to longer forecast horizons (5, 10 and 20-step ahead), we find that the ranking becomes more stable as the horizon increases. The lower variability of sample performances is due to the fact that longer horizon volatility forecasts are smoother. For the robust loss function, Figure 2.1(b), (c) and (d) show that as the forecasting horizon increases models' loss differentials become larger. The broad difference between integrated models, which exhibit diverging paths for long horizon forecasts (except the RiskMetrics and CCC-Rm for which the multistep ahead forecast coincides with the 1-step ahead for every horizon), and stationary models, which in turn converge towards the long run variance matrix, is particularly noticeable. For the non-robust loss function, Figure 2.2(d), (e) and



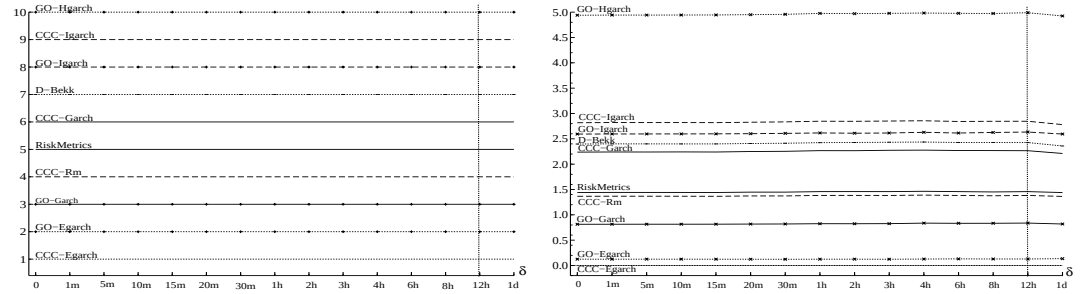
a) 1-step



b) 5-step

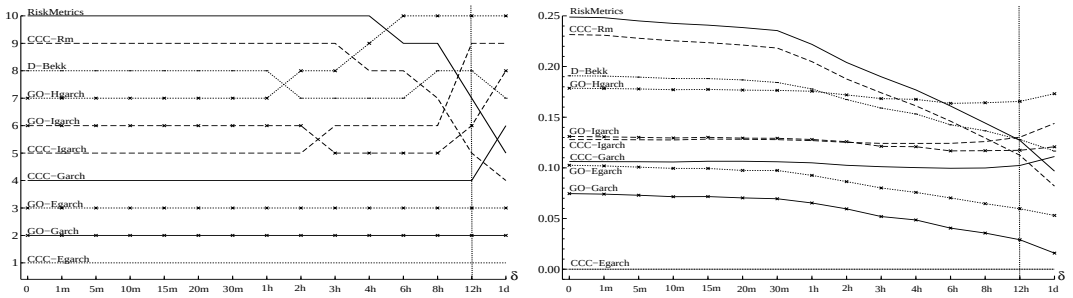


c) 10-step

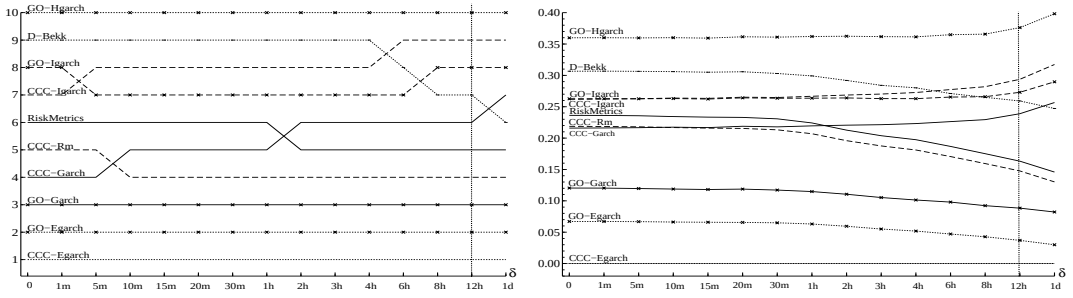


d) 20-step

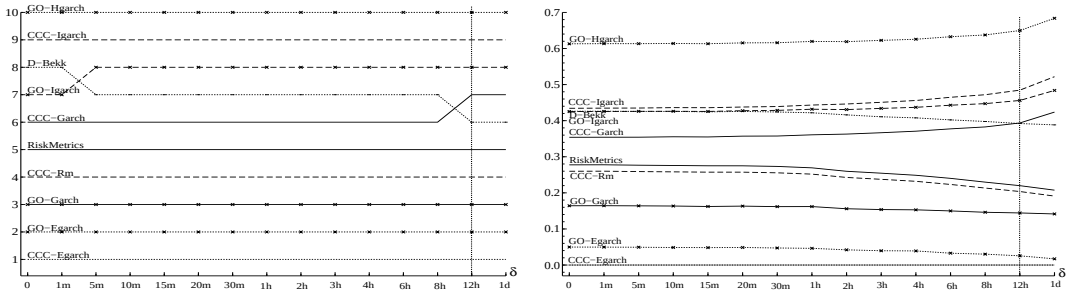
Figure 2.1: Ranking implied by L_F (robust) for different forecast horizons (1-step to 20-step). Ranking based on avg. performances (left) and avg. loss differentials from CCC-Egarch (right).



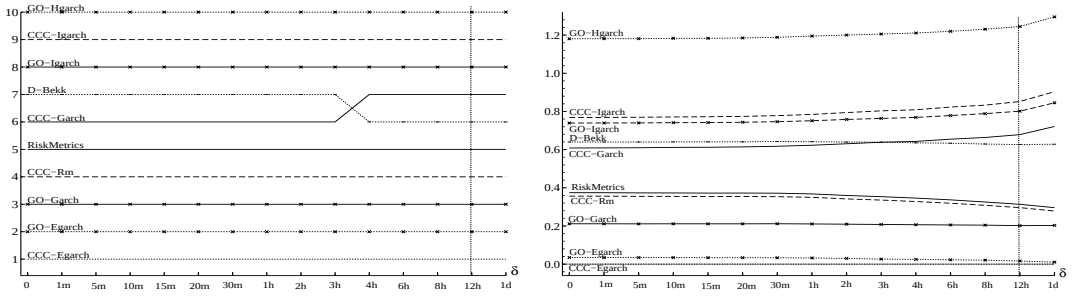
a) 1-step



b) 5-step



c) 10-step



d) 20-step

Figure 2.2: Ranking implied by L_{1M} (non-robust) for different forecast horizons (1-step to 20-step). Ranking based on avg. performances (left) and avg. loss differentials from CCC-Egarch (right).

(f), we find that the impact of the bias is less striking and tends to become marginal as the forecasting horizon increases. The break even point after which the loss differentials cease to be reasonably stable and identified for the 1-step ahead case in the 30-minute frequency, seems to remain the same regardless the forecast horizon. However, since the loss differentials get larger with the forecast horizon as expected, the objective bias is less likely to appear, and it induces only marginal distortions in the ranking. Finally, for both robust and non-robust loss functions, we find that generalized orthogonal models perform better when the comparison is based on longer forecast horizons.

In conclusion, for a robust loss function, even when the relative performances are extremely close, the ordering remains unaffected by the proxy quality and we are always able to correctly discriminate between models. For an non-robust loss function, we find that if the volatility proxy is sufficiently accurate relative to the degree of similarity between model performances is it still possible to recover the underlying ranking. As the quality of the proxy deteriorates the relative performances of some models appear to improve with respect to others. We identify a threshold such that orderings based on lower precision proxies appear strongly biased. The results are qualitatively similar for multistep-ahead forecasting where the problem of the bias seems to affect the ordering to a lower extent as the horizon increases. Finally, we also investigate the impact of the estimation sample size on the rankings. Increasing the sample size to 3000 observations gives qualitatively similar results.

2.5 Empirical application

2.5.1 Data description

The empirical application is based on the Euro, British Pound and the Japanese Yen exchange rates expressed in US dollars (EUR, GBP and JPY). The data has been provided by Olsen Financial Technologies. The estimation sample ranges from January 6, 1987 to December 28, 2001 (3666 trading days). 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), as defined in Section 2.4.1, computed using intra-day returns sampled at 17 different frequencies ranging from $\delta = 5$ minutes (288 intervals/day) to $\delta = 1$ day (1 interval/day). Note that the realized variance matrix is positive definite until the 8-hour sampling frequency (3 intervals/day).

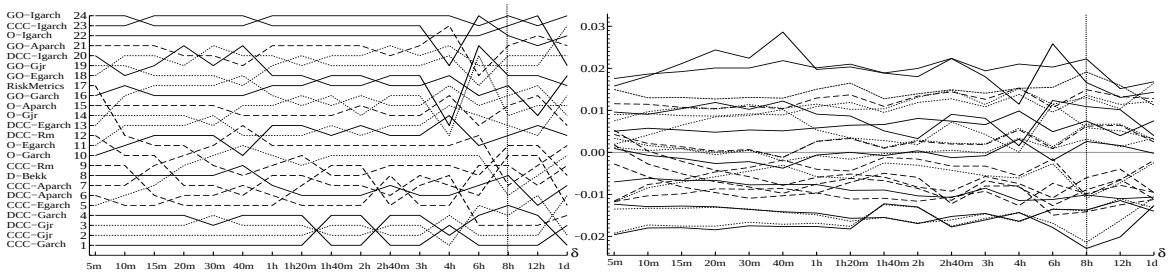
The forecasting models set includes 24 specifications: D-Bekk, RiskMetrics, CCC-Garch, CCC-Igarch, CCC-Egarch, CCC-Aparch, CCC-Gjr, CCC-Rm, DCC-Garch, DCC-Igarch,

DCC-Egarch, DCC-Aparch, DCC-Gjr, DCC-Rm, O-Garch, O-Igarch, O-Egarch, O-Aparch, O-Gjr, GO-Garch, GO-Igarch, GO-Egarch, GO-Aparch and GO-Gjr. One and ten-step ahead forecasts are compared to the proxy $\hat{\Sigma}_{t,\delta}$ using one robust (L_F) and one non-robust (L_{1M}) loss function. 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).

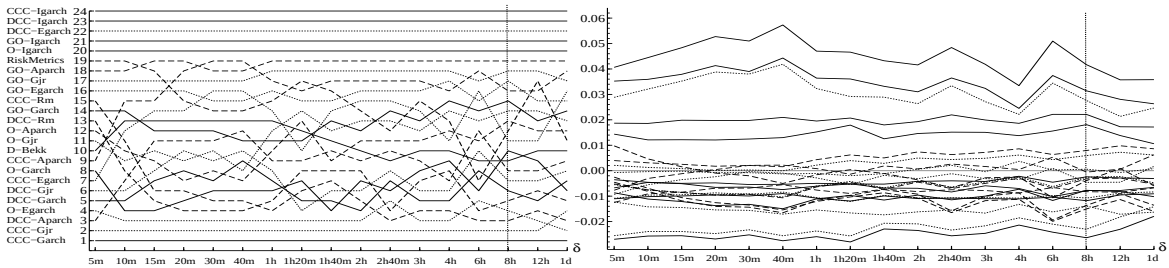
2.5.2 Model comparison

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 2.3 and 2.4. The vertical line at δ =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 robust loss function, see Figure 2.3(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, performs clearly better than all the other models. These models exhibit particularly stable and relatively close sample performances (Figure 2.3(a-right)). The fact that, CCC models show smaller losses than the equivalent DCC (the same holds also for O and GO models) is not surprising. In fact, when the process is characterized by simple dynamics, as in the case of exchange rates, simple, even misspecified models can outperform more flexible specifications especially in presence of high estimation uncertainty, as pointed out by Giacomini and White (2006). 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 sensibly more volatile. As pointed out by Hansen and Lunde (2006a), we can observe discrepancies between the empirical and the approximated ranking in finite samples (i.e., sampling variability). These differences must not be confused with the objective bias. As the accuracy of the proxy deteriorates, the loss function becomes less informative, and as a result it is more difficult to effectively order models' performances. The effect becomes more severe when there is a high degree of similarity between models under evaluation.

Results for the 10-step ahead forecast horizon, see Figure 2.3(b), are in line with our simulation results. Model performances tend to cluster and loss differentials between clusters broaden. The CCC-Garch and CCC-Gjr perform largely better than all the other models and, in particular, they leave behind the corresponding DCC specifications. Although the



a) 1-step

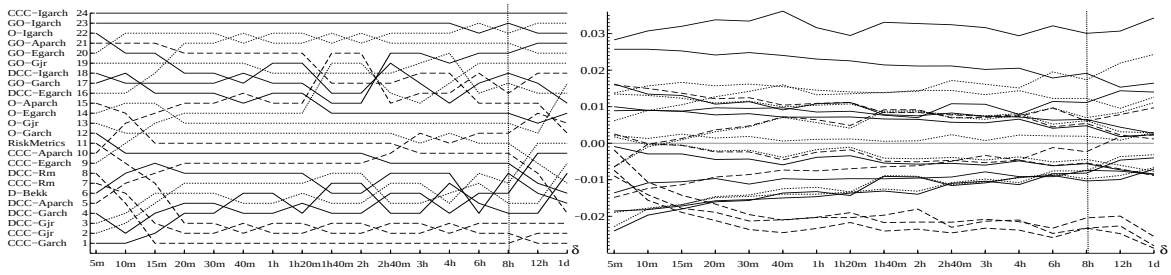


b) 10-step

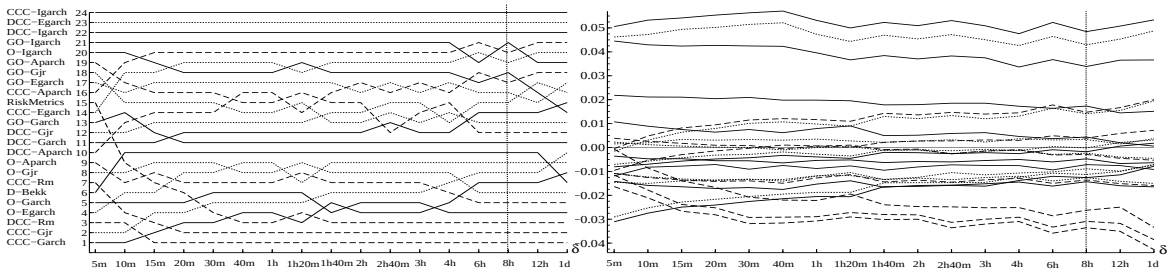
Figure 2.3: Ranking implied by L_F (robust). Ranking based on sample performances (left) and loss differentials from common average (right).

models' performances are more stable than in the 1-step ahead forecast horizon (Figure 2.3(b-right)), the convergence of the sample performances of a number of models induces an extremely large variability of the ranking in the middle of the classification (Figure 2.3(b-left)).

Figure 2.4(a-left) illustrates to what extent the presence of the objective bias can affect the ranking when using a non-robust loss function. In this case, the distortion in the ordering is striking. The CCC and the DCC models with Rm conditional variances rank 7th and 8th respectively at $\delta = 5$ minutes, but they rapidly climb towards the top of the classification as the frequency for $\hat{\Sigma}_{t,\delta}$ lowers. Starting from $\delta = 15$ minutes they reach the top of the classification, ranking first and third. Similarly, the RiskMetrics model, ranking 11th when $\delta = 5$ minutes, joins the top of the ranking at a relatively high frequency. When $\hat{\Sigma}_{t,\delta}$ is computed using data sampled at a frequency equal or lower than 20 minutes, the RiskMetrics model ranks 3rd, behind the CCC-Rm and DCC-Rm models. Given that these models are characterized by a dynamic in the variance structure imposed ex-ante and independent from the data (with the only exception the DCC-Rm for which the parameters of the dynamic correlation are estimated), it is unlikely that such models are the best forecasting models. Similarly to what we observe in the simulation study, a distorted ordering is not the only



a) 1-step



b) 10-step

Figure 2.4: Ranking implied by L_{1M} (non-robust). Ranking based on sample performances (left) and loss differentials from common average (right).

evidence of a distorted outcome. In fact, from Figure 2.4(a-right) we remark also that, as the sampling frequency used to compute $\hat{\Sigma}_{t,\delta}$ lowers, the performance of RiskMetrics-type models steadily improve with respect to the other models. This result is fallacious because it 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 2.3(b)).

2.5.3 Model confidence set

To illustrate how critical an adequate choice of the loss function is for model selection based on forecasting ability, we apply the Model Confidence Set (MCS) test of Hansen, Lunde, and Nason (2010a) to the set of 24 MGARCH models considered in the previous section. As explained in Chapter 1, the MCS test 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. The selection of the superior models by the MCS approach obviously depends on the orderings implied by a loss function (e.g., the ranking given in Figures 2.3(a) and 2.4(a) - left panels). Thus, an unfortunate choice of the loss function can result in an incorrect identification of the set of superior models even if the testing procedure is formally valid.

Table 2.3 reports the MCS obtained under L_F and L_{1M}, with respect to five volatility proxies $\hat{\Sigma}_{t,\delta}$ (δ=5m, 20m, 1h20m, 2h40m, 8h), and for a 1-step and 10-step forecast horizon.

Table 2.3: Model Confidence Set

1-step ahead forecast horizon

Frobenius distance (robust)					Entrywise-1 norm (non-robust)				
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 (robust)					Entrywise-1 norm (non-robust)				
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 common average). Significance level $\alpha = 0.1$. Sample size=650 obs. Standard errors based on 10,000 bootstrap resamples.

The MCSs with a confidence level $\alpha = 0.1$ are reported in Table 2.3. Under the robust loss function, L_F, the sets of superior models appear to be consistent across sampling frequencies (δ). In terms of MCS, consistency of the ranking implies that the set of superior models identified using a high precision proxy is always included in the set obtained using a less accurate proxy. The fact that the set of superior models increases in size as the sampling frequency lowers is expected. This result is due to the loss of accuracy of the proxy which

translates into a higher variability of the sample evaluation of each model. Since the loss function becomes less informative it is more difficult to discriminate between models and, for a given confidence level, the set of superior models increases. For instance, the MCS obtained using a proxy based on 8-hour returns contains one half for 1-day horizon and two thirds for 10-day horizon of the 24 candidate models.

Relying on volatility signature plots and signal to noise ratios, Andersen, Bollerslev, Diebold, and Labys (1999) and Russell and Bandi (2004) show that a volatility proxy based on a sampling frequency between 5 and 20 minutes strikes a good compromise between the loss of accuracy (lower frequencies) and the presence of noise due to microstructure frictions (higher frequencies). Indeed, when $\delta = 5$ and 20 minutes, we obtain the most accurate sets, showing that the MCS is able to separate efficiently superior from inferior models under the robust loss function.

Our 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 non-robust Entrywise-1 norm loss function confirms the presence of the objective bias. For both forecast horizons, the MCS changes in composition and reduces in size as the quality of the proxy deteriorates. This is the opposite effect that we find with the robust loss function. The sets obtained using a proxy based on 5-minute and 8-hour returns do not share common elements. At δ=8 hours, the set includes only RiskMetrics-type models, which corroborates the findings in the previous subsection.

2.6 Conclusion

Ranking the forecasting performances of multivariate volatility models raises two important issues. First, there is the choice of the loss function (the criterion used to measure the accuracy of the predicted covariance matrices) and second, the choice of a proxy of the unobservable volatility measure used to evaluate models forecasts. However, the evaluation of 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.

In this chapter, we first define sufficient conditions for a loss function to ensure consistency between the true, unobservable, ranking - based on the true conditional variance matrix - and the approximated one - based on a conditionally unbiased proxy. Second, we identify a necessary and sufficient functional form for the loss function to ensure consistent

ordering, under the use of a proxy, in matrix and vector spaces. Finally, we provide a number of parameterizations, some of which are often used in practice and discuss their properties.

In the simulation study, we illustrate using artificial data the practical implications of our theoretical results in a simulation based comparison of multivariate GARCH models in a realistic setting. We sample from a continuous time multivariate diffusion process and estimate discrete time multivariate GARCH models to illustrate the sensitivity of the ranking to different choices of the loss function and to the quality of the proxy for different forecast horizons. We observe that, if the quality of the proxy is sufficiently good, both robust and non-robust loss functions rank properly. However, when the quality of the proxy is poor, only the robust loss functions rank properly while the ranking implied by the non-robust loss functions appears heavily biased. Our findings also hold when the sample size in the estimation period increases.

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 robust loss function and the importance of high precision proxy for model selection. We also study to what extent the ranking and the MCS test are affected when we combine an uninformative proxy with a robust and a non-robust loss function. Coupling a robust loss function with a relatively uninformative proxy is likely result in an uninformative MCS, i.e., in the impossibility to efficiently discriminate between forecast performances. When the evaluation is based on an non-robust loss function and an uninformative proxy the MCS obtained under more accurate proxies is likely to differ (in full or in part) from the MCS obtained under less accurate proxies reflecting the inconsistency of the ordering. Hence, an adequate choice of the combination loss function/quality of the proxy, is crucial to the evaluation, because even if the testing procedure is formally valid (independently from the choice of the loss function or the proxy), it may result in a perverse or at best an uninformative outcome. This is an important message for the applied econometrician.

There are several extensions for future research. First, this thesis ranks multivariate volatility models based on statistical loss functions and focuses on conditions for consistent ranking from a more theoretical viewpoint. At some point an economic loss function has to 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 thesis. Second, from an applied viewpoint, the behavior of the ranking when using proxies other than realized covariance should be further investigated.

2.7 Appendix A: Proofs

Proof of Proposition 2.1. To illustrate the validity of Proposition 2.1, consider the second order Taylor expansion of $L(\hat{\Sigma}_t, H_t)$ around the true value Σ_t:

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

Taking conditional expectations with respect to $\Im_{t-1}$ we get

$$\mathrm{E}_{t-1}[L(\hat{\Sigma}_t, H_t)] \cong L(\Sigma_t, H_t) + \frac{1}{2}\left[\mathrm{E}_{t-1}\left(\xi_t' \frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t'} \xi_t\right)\right], \tag{2.18}$$

because, under A2.2 and A2.4 and when Proposition 2.1 is satisfied, we have:

(a) $\mathrm{E}_{t-1}\left[\left(\frac{\partial L(\Sigma_t, H_t)}{\partial \sigma_t}\right)' \xi_t\right] = \left(\frac{\partial L(\Sigma_t, H_t)}{\partial \sigma_t}\right)' \mathrm{E}_{t-1}(\xi_t) = 0$, i.e., $\hat{\sigma}_t$ is conditionally unbiased with respect to σ_t;

(b) $\frac{\partial^2 L(\Sigma_t, H_{m,t})}{\partial \sigma_t \partial \sigma_t'} = \Psi(\sigma_t, .)\ \forall m$, i.e., the last term in (2.18) does not depend on model m,

hence $\mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]$ and $L(\Sigma_t, H_t)$ induce the same ordering over m.

To conclude, (2.18) implies that in order to achieve consistency in the sense of Definition 2, the equivalence between $\mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]$ and $L(\Sigma_t, H_t)$ is not required, but it suffices that the discrepancy between the two terms, $\frac{1}{2}\mathrm{E}_{t-1}\left(\xi_t' \Psi(\sigma_t, .)\xi_t\right)$, is constant across models, thus not affecting the ordering. ■

Proof of Lemma 2.1. Under Assumptions A2.1 to A2.4, the first order conditions of the minimization problem in (2.3), considering the expansion in (2.18), are

$$\begin{aligned}
\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t^{(s)}, H_t)\right]}{\partial h_{k,t}} - \frac{\partial L(\Sigma_t, H_t)}{\partial h_{k,t}} &\cong \frac{1}{2}\left[\frac{\partial}{\partial h_{k,t}} \mathrm{E}_{t-1}\left(\xi_t^{(s)'} \Psi(\sigma_t, h_t) \xi_t^{(s)}\right)\right] \\
&\cong \frac{1}{2} \frac{\partial}{\partial h_{k,t}} \mathrm{E}_{t-1}\left[\sum_{l,m} \xi_{l,t}^{(s)} \xi_{m,t}^{(s)} \Psi(\sigma_t, h_t)_{l,m}\right] \\
&\cong \frac{1}{2} \sum_{l,m} \frac{\partial \Psi(\sigma_t, h_t)_{lm}}{\partial h_{k,t}} \mathrm{E}_{t-1}[\xi_{l,t}^{(s)} \xi_{m,t}^{(s)}] \\
&\cong \frac{1}{2} \sum_{l,m} \frac{\partial \Psi(\sigma_t, h_t)_{lm}}{\partial h_{k,t}} V_{l,m,t}^{(s)}
\end{aligned}$$

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 $\Psi(\sigma_t, h_t)_{l,m}$ 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 $\Psi(\sigma_t, h_t)$, the matrix of second derivatives of $L(.,.)$ with respect to σ_t.

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, under A2.3, A1.1 ensures that second order conditions are satisfied. Then, we have that

$$-\frac{\partial L(\Sigma_t, H_t^{*(s)})}{\partial h_{k,t}} \cong \frac{1}{2}\sum_{l,m}\frac{\partial \Psi(\sigma_t, .)_{lm}}{\partial h_{k,t}} V_{l,m,t}^{(s)}. \tag{2.19}$$

Under i), i.e., $\frac{\partial \Psi(\sigma_t,.)_{lm}}{\partial h_{k,t}} = 0 \ \forall k$, 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, A2.2 and A2.4, 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 A2.5 and (2.19) we have

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

which concludes the proof. ■

Proof of Proposition 2.2. To prove the proposition, we proceed as in Patton (2009). We show the equivalence of the following statements:

-S1: the loss function takes the form in the proposition;

-S2: the loss function is robust in the sense of Definition 2;

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

The proof exploits the results of Lemma 8.1 in Gourieroux and Monfort (1995) page 240 which can be formalized as follows. Let $g_1, ..., g_k, h$ be some functions from $\mathbb{R}^G$ into $\mathbb{R}$, satisfying

i. for any k, there exist some probability law P_{1k} and P_{2k} such that

$$\int_y g_k(y)\mathrm{d}P_{1k}(y) > 0, \quad \int_y g_k(y)\mathrm{d}P_{2k}(y) < 0,$$

$$\int_y g_j(y)\mathrm{d}P_{1k}(y) = 0, \quad \int_y g_j(y)\mathrm{d}P_{2k}(y) = 0, \ \forall j \neq k,$$

ii. for any probability law P such that

$$\int_y g_k(y)\mathrm{d}P(y) = 0, \ \forall k = 1, ..., K, \ \text{ then } \int_{\mathrm{y}} \mathrm{h(y)dP(y)} = 0,$$

then there exist some real numbers λ_k, $k = 1, ..., K$, such that

$$h(y) = \sum_{k=1}^{K} \lambda_k g_k(y), \quad \forall y \in \mathbb{R}^G.$$

Step 1: S1⇒S2. The result follows directly form Proposition 2.1, in fact:

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

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, there exists an H_t^* in the support of $L(\hat{\Sigma}_t, H_t)$ such that $H_t^* = \mathrm{E}_{t-1}(\hat{\Sigma}_t)$. This implies that $\forall H_t \in int(\dot{H}) \setminus \{H_t^*\}$:

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

and therefore 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 by Definition 2, under S2, we can write

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

if we set $H_t = \Sigma_t$, then by Assumptions A1.1 to A1.3, $\mathrm{E}(L(\Sigma_t, \Sigma_t)) = 0 \Rightarrow \mathrm{E}(L(\Sigma_t, H_t^*)) = 0$ 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, and considering $y = \mathrm{vech}(\hat{\Sigma}_t)$ and as functions $g_k(y) = (\mathrm{vech}(\hat{\Sigma}_t - H_t^*)_k$ and as function $h(y)$ the partial derivatives $\frac{\partial L(\hat{\Sigma}_t, H_t^*)}{\partial h_{k,t}}$, $k = 1, ..., K = N(N+1)/2$, then there exists a square matrix Λ of size K which is only function of H_t^* such that

$$\frac{\partial L(\hat{\Sigma}_t, H_t^*)}{\partial h_t} = \Lambda(H_t^*)\mathrm{vech}(\hat{\Sigma}_t - H_t^*). \tag{2.20}$$

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

$$\begin{aligned}\mathrm{E}_{t-1}\left(\frac{\partial^2 L(\hat{\Sigma}_t, H_t^*)}{\partial h_t \partial h_t'}\right) &= \mathrm{E}_{t-1}\left(\frac{\partial \Lambda(H_t^*)\mathrm{vech}(\hat{\Sigma}_t - H_t^*)}{\partial h_t}\right) \\ &= \mathrm{E}_{t-1}\left(\begin{bmatrix} \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{1i}}{\partial h_1}(\sigma_i - h_i^*) & \dots & \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{1i}}{\partial h_k}(\sigma_i - h_i^*) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{ki}}{\partial h_1}(\sigma_i - h_i^*) & \dots & \sum_{i=1}^{K} \frac{\partial \Lambda(H_t^*)_{ki}}{\partial h_k}(\sigma_i - h_i^*) \end{bmatrix}\right) - \Lambda(H_t^*) \\ &= -\Lambda(H_t^*),\end{aligned}$$

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 (2.20) (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 (2.20) 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 (2.20) 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 2.1. 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 2.2. ■

Proof of Remark 1. The proof of part *i)* of the Remark follows from Proposition 2.2.

For the second part, notice that

$$\frac{\partial^2 L(\Sigma_t, H_t)}{\partial \sigma_t \partial \sigma_t'} = -\tilde{C}''_{\sigma_t}(\Sigma_t),$$

since if $f[\cdot]$ is a linear map, then $f[\bar{C}(H_t)(\Sigma_t - H_t)]$ is linear in $\sigma_{i,j,t}$ $\forall i, j = 1, ..., N$. Hence, the general conclusion of Proposition 2.1 holds even under violation of A1.2: the ordering implied by $\mathrm{E}_{t-1}[L(\hat{\Sigma}_t, H_t)]$ is apparently robust for the one based on $L(\Sigma_t, H_t)$ in the sense that is insensitive to the substitution of the true variance matrix by a proxy (by the same reasoning provided in the proof of Proposition 2.1), i.e., $\arg\min_{H_t \in \dot{H}} L(\Sigma_t, H_t) = \arg\min_{H_t \in \dot{H}} \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]$.

We now show that, though apparently robust, the ordering obtained when $f[\cdot] \not\equiv \mathrm{Tr}[\cdot]$ is not a valid one, that is it differs from any valid or acceptable ordering and in particular it holds $H_t^* \neq \mathrm{E}_{t-1}(\hat{\Sigma}_t) = \Sigma_t$.

Consider the first order conditions of (2.3) under the loss in (2.6) evaluated at the optimum H_t^*

$$\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]}{\partial h_t} = C(H_t^*) + f'_{h_t}[\bar{C}(H_t^*)(\mathrm{E}_{t-1}(\hat{\Sigma}_t) - H_t^*)] = 0.$$

Recall that $C(H_t) = \nabla \tilde{C}(H_t)$ and f'_{h_t} is the gradient of f with respect to h_t. Using the fact that f is a linear map, the typical element of the gradient of $\mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]$, indexed by $i, j = 1, ..., N$, $i \leq j$ is (we omit the time index to simplify notation)

$$\frac{\partial \mathrm{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]}{\partial h_{i,j}} = \tilde{C}'_{h_{i,j}}(H_t^*) + f\left[\frac{\partial \bar{C}(H_t^*)}{\partial h_{i,j}}(\mathrm{E}_{t-1}(\hat{\Sigma}_t) - H_t^*)\right] - f\left[\bar{C}(H_t^*)\frac{\partial H_t^*}{\partial h_{i,j}}\right] = 0. \tag{2.21}$$

To deliver an appropriate ordering, the loss function must be such that it is uniquely minimized at $H_t^* = \mathrm{E}_{t-1}(\hat{\Sigma}_t) = \Sigma_t$, that is optimal forecast is the true conditional variance, which implies

$$f\left[\frac{\partial \bar{C}(H_t^*)}{\partial h_{i,j}}(\mathrm{E}_{t-1}(\hat{\Sigma}_t - H_t^*)\right] = 0.$$

Therefore, in (2.21), it must hold

$$f\left[\bar{C}(H_t^*)\frac{\partial H_t^*}{\partial h_{i,j}}\right] = \tilde{C}'_{h_{i,j}}(H_t^*). \tag{2.22}$$

Since $\frac{\partial H_t^*}{\partial h_{i,j}}$, for each $i, j = 1, ..., N$ $i \leq j$, is a $N \times N$ symmetric matrix with elements indexed

by $[i,j]$ and $[j,i]$ equal to 1 and zero elsewhere, (2.22) holds if and only if $f(.) = \text{Tr}(.)$. In fact, from (2.22)

$$\begin{aligned} i &= j \Longrightarrow \text{Tr}\left[\bar{C}(H_t^*)\frac{\partial H_t^*}{\partial h_{i,i}}\right] = \bar{c}_{i,i}(H_t^*) = \tilde{C}'_{h_{i,i}}(H_t^*) \\ i &\neq j \Longrightarrow \text{Tr}\left[\bar{C}(H^*)\frac{\partial H^*}{\partial h_{i,j}}\right] = 2\bar{c}_{i,j}(H_t^*) = \tilde{C}'_{h_{i,j}}(H_t^*). \end{aligned} \tag{2.23}$$

Substituting (2.23) in (2.21), we obtain

$$\begin{aligned} \frac{\partial \text{E}_{t-1}\left[L(\hat{\Sigma}_t^*, H_t)\right]}{\partial h_{i,j}} &= \tilde{C}'_{h_{i,j}}(H_t^*) + \text{Tr}\left[\frac{\partial C(H_t^*)}{\partial h_{ij}}(\text{E}_{t-1}(\hat{\Sigma}_t) - H_t^*)\right] - \tilde{C}'_{h_{ij}}(H_t^*) \\ &= \text{Tr}\left[\frac{\partial C(H_t^*)}{\partial h_{i,j}}(\hat{\Sigma}_t - H_t^*)\right], \end{aligned}$$

and finally

$$\begin{aligned} \frac{\partial \text{E}_{t-1}\left[L(\hat{\Sigma}_t, H_t)\right]}{\partial h_t} &= \begin{bmatrix} \text{Tr}\left[\frac{\partial \bar{C}(H_t^*)}{\partial h_{1,1}}(\text{E}_{t-1}(\hat{\Sigma}_t) - H_t^*)\right] \\ \cdots \\ \text{Tr}\left[\frac{\partial \bar{C}(H_t^*)}{\partial h_{i,j}}(\text{E}_{t-1}(\hat{\Sigma}_t) - H_t^*)\right] \\ \cdots \end{bmatrix} = 0 \\ &\Leftrightarrow \quad H_t^* = \text{E}_{t-1}(\hat{\Sigma}_t), \end{aligned}$$

which concludes the proof. ■

Proof of Proposition 2.3. By Proposition 2.2, a robust loss functions based on the forecast error must have the form

$$L(\hat{\Sigma}_t - H_t) = \tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)'\text{vech}(\hat{\Sigma}_t - H_t). \tag{2.24}$$

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 $L(\hat{\Sigma}_t - H_t)$ then $L(\hat{\Sigma}_t - H_t) = L(H_t - \hat{\Sigma}_t)$, i.e., $L(.,.)$ is symmetric under 180° rotation around the origin

and, which implies

$$-\frac{\partial L(\hat{\Sigma}_t - H_t)}{\partial h_t} = \frac{\partial L(\hat{\Sigma}_t - H_t)}{\partial \sigma_t}, \tag{2.25}$$

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 (2.25) with respect to σ_t we obtain

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

which implies

$$\nabla^2 \tilde{C}(H_t) = \Lambda, \tag{2.26}$$

where Λ is a matrix of constants.

Equation (2.26) 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 (2.24) can be rewritten as

$$L(\hat{\Sigma}_t - H_t) = -\tilde{C}(H_t) - \tilde{C}(\hat{\Sigma}_t) + C(H_t)'\text{vech}(\hat{\Sigma}_t). \tag{2.27}$$

In order to satisfy second order conditions Λ must be negative definite, according to Proposition 2.2. Since $L(\hat{\Sigma}_t, H_t)$ is homogeneous of degree 2, starting from (2.26), 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 (2.27) 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 $\hat{\Lambda} = -\frac{1}{2}\Lambda$. ■

2.8 Appendix B: Examples for Section 2.2.4

In the following examples, for ease of exposition, we consider a forecast error matrix of dimension $N = 2$.

In the first three examples we investigate the properties of loss functions belonging to the family of quadratic loss functions defined in Proposition 2.3. The vector of forecast errors of interest is therefore

$$vech(\Sigma_t - H_t) = \begin{pmatrix} \sigma_{1,1,t} - h_{1,1,t} \\ \sigma_{1,2,t} - h_{1,2,t} \\ \sigma_{2,2,t} - h_{2,2,t} \end{pmatrix},$$

which allows to plot the contours of the loss function.

The first loss function that we consider is the Euclidean distance, which corresponds to a choice of $\hat{\Lambda} = I_K$ and can be expressed as

$$L_E = (\sigma_{1,1,t} - h_{1,1,t})^2 + (\sigma_{1,2,t} - h_{1,2,t})^2 + (\sigma_{2,2,t} - h_{2,2,t})^2.$$

Figure 2.5 reports the contour of $L_E = 1$.

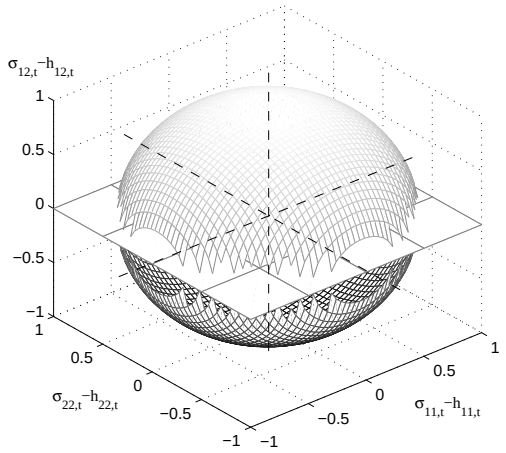


Figure 2.5: Euclidean distance - $L_E = 1$

The contours of L_E are spheres centered at the origin. The loss function has mirror symmetry about all coordinate planes, it is also symmetric under any rotation about the origin and, being a symmetric polynomial, it is symmetric about the bisector planes.

The second loss function is the weighted Euclidean distance with

$$\hat{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

which implies

$$L_{WE} = (\sigma_{1,1,t} - h_{1,1,t})^2 + 4(\sigma_{1,2,t} - h_{1,2,t})^2 + 2(\sigma_{2,2,t} - h_{2,2,t})^2,$$

which implies that $(\sigma_{2,2,t} - h_{2,2,t})$ is penalized double with respect to $(\sigma_{1,1,t} - h_{1,1,t})$, while the covariance forecast error is penalized four times more. The reason behind such particular choice of $\hat{\Lambda}$ is to emphasize the role of each weight and to show how they affect the shape of the loss function. The contour of $L_{WE} = 1$ is an ellipsoid centered at the origin (Figure 2.6).

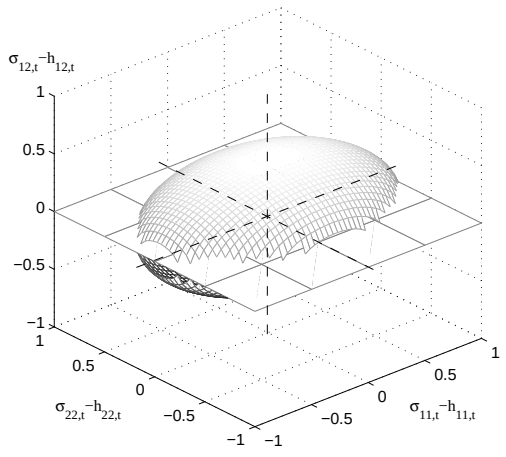


Figure 2.6: Weighted Euclidean distance - $L_{WE} = 1$

The contour is squeezed around the $(\sigma_{1,1,t} - h_{1,1,t})$ axis due to the unequal weighting. The loss function in symmetric about all coordinate planes and it is also symmetric under a 180° rotation around the origin, i.e., considering the absolute forecast error vector $|\sigma_t - h_t| =$

$(0.2, 0.4, 0.8)$, we have

$$\begin{aligned} L_{WE}(0.2, 0.4, 0.8) &= L_{WE}(-0.2, -0.4, -0.8) = 1.96 \\ L_{WE}(0.2, 0.4, -0.8) &= L_{WE}(0.2, 0.4, -0.8) = 1.96 \\ L_{WE}(0.2, -0.4, 0.8) &= L_{WE}(0.2, -0.4, 0.8) = 1.96 \\ &\ldots \end{aligned}$$

However, L_{WE} is not symmetric about the bisector planes, i.e.

$$\begin{aligned} L_{WE}(0.2, 0.4, 0.8) &= 1.96 \neq 2.92 = L_{WE}(0.2, 0.8, 0.4) \\ &\neq 1.12 = L_{WE}(0.8, 0.2, 0.4) \\ &\ldots \end{aligned}$$

The last example is the pseudo Mahalanobis distance with

$$\hat{\Lambda} = \begin{bmatrix} 1 & 0 & 0.6 \\ 0 & 4 & 0 \\ 0.6 & 0 & 2 \end{bmatrix},$$

that is

$$L_M = (\sigma_{1,1,t} - h_{1,1,t})^2 + 4(\sigma_{1,2,t} - h_{1,2,t})^2 + 2(\sigma_{2,2,t} - h_{2,2,t})^2 + 1.2(\sigma_{1,1,t} - h_{1,1,t})(\sigma_{2,2,t} - h_{2,2,t}).$$

For illustrative purposes, we set only one off diagonal element of the matrix of weights different from 0. As in the previous case, the contour of $L_M = 1$ is an ellipsoid centered at the origin (Figure 2.7). It is clear that L_M is only symmetric under a 180° around the origin. Furthermore, its axes of symmetry (dashed lines in Figure 2.7), whose directions depend on the sign of the off diagonal elements of $\hat{\Lambda}$, are rotated with respect to the coordinate axes (e.g. in Figure 2.7, $\hat{\Lambda}$ implies an horizontal rotation). In this regard, since the loss function also includes the cross product of the elements of $(\sigma_t - h_t)$ weighted by the off diagonal elements of $\hat{\Lambda}$ (which can be positive and/or negative provided $\hat{\Lambda}$ satisfies Proposition 2.3), a positive weight means that, for given absolute forecast errors $|\sigma_t - h_t|$, L_M will penalize more the outcomes where both variances are over/under predicted. In fact, consider L_M evaluated at $|\sigma_t - h_t| = (0.8, 0, 0.4)$, then

$$\begin{aligned} L_M(0.8, 0, -0.4) &= L_M(-0.8, 0, 0.4) = 0.576 \\ L_M(0.8, 0, 0.4) &= L_M(-0.8, 0, -0.4) = 1.344. \end{aligned}$$

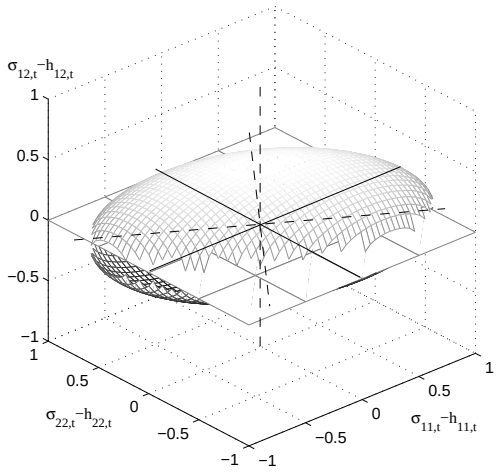


Figure 2.7: Pseudo Mahalanobis distance - $L_M = 1$

In the last example we focus on the Stein loss function. Note that providing a comprehensive illustration of the geometric properties of L_S is somehow more complex then in the previous cases. We have shown that quadratic loss functions are defined on the forecast error matrix $\Sigma_t - H_t$ which implies that $L(\Sigma_t - H_t) : R^{N\times N} \to R_+$ even if Σ_t and $H_t \in R^{N\times N}_{++}$ (the space of positive definite matrices). This allows for a graphical representation of the forecast error vector, i.e., the vector of unique elements of $\Sigma_t - H_t$, in the space $R^{N(N+1)/2}$. On the other hand, L_S is defined on the standardized (in matrix sense) forecast error $\Sigma_t H_t^{-1}$, that is positive definite. Since the domain of $L(\Sigma_t H_t^{-1})$ is $R^{N\times N}_{++} \subset R^{N\times N}$, the graphical representation of the contours in the Euclidean space is not an easy task. Furthermore, unlike the loss functions based on the forecast error matrix, L_S cannot be expressed as a combination of functions of the elementwise forecast errors, i.e., $L(\Sigma_t, H_t) = L(l(\sigma_{1,t}, h_{1,t}), ..., l(\sigma_{K,t}, h_{K,t}))$, except in the trivial case when Σ_t and H_t are diagonal. Therefore, to illustrate the properties of the Stein loss function we rely on some numerical examples and the analysis of the conditional loss.

Consider a standardized forecast error matrix of dimension $N = 2$.

$$\Sigma_t H_t^{-1} = \begin{bmatrix} \frac{\sigma_{1,2,t}h_{1,2,t} - \sigma_{1,1,t}h_{2,2,t}}{h^2_{1,2,t} - h_{1,1,t}h_{2,2,t}} & \frac{\sigma_{1,1,t}h_{1,2,t} - \sigma_{1,2,t}h_{1,1,t}}{h^2_{1,2,t} - h_{1,1,t}h_{2,2,t}} \\ \frac{\sigma_{2,2,t}h_{1,2,t} - \sigma_{1,2,t}h_{2,2,t}}{h^2_{1,2,t} - h_{1,1,t}h_{2,2,t}} & \frac{\sigma_{1,2,t}h_{1,2,t} - \sigma_{2,2,t}h_{1,1,t}}{h^2_{1,2,t} - h_{1,1,t}h_{2,2,t}} \end{bmatrix}.$$

The Stein loss function is therefore

$$L_S = \frac{\sigma_{1,1,t}h_{2,2,t} + \sigma_{2,2,t}h_{1,1,t} - 2\sigma_{1,2,t}h_{1,2,t}}{h_{1,1,t}h_{2,2,t} - h_{1,2,t}^2} - \ln(\frac{\sigma_{1,1,t}\sigma_{2,2,t} - \sigma_{1,2,t}^2}{h_{1,1,t}h_{2,2,t} - h_{1,2,t}^2}) - 2$$

For ease of exposition, we set Σ_t to some arbitrary values, say

$$\Sigma_t = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 3 \end{bmatrix}.$$

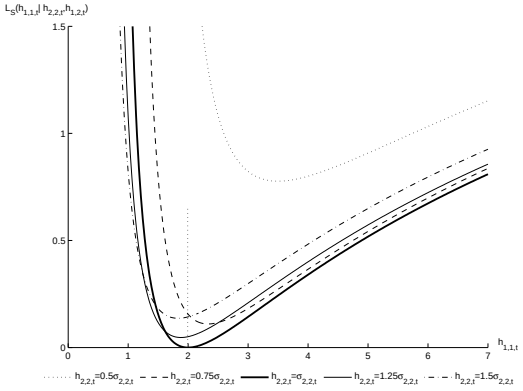
Since the loss function is expressed in terms of standardized forecast errors, we first assess the case of over/under prediction in measure of $\pm 0.5\Sigma_t$. The loss when each element of H_t over/under predicts the corresponding element of Σ_t (setting the others at their optimal values), is

	$(-)$	$(+)$
$L_S(h_{1,1,t} = 2 \pm 1)$	2.390	0.143
$L_S(h_{2,2,t} = 3 \pm 1.5)$	2.390	0.143
$L_S(h_{1,2,t} = 1.5 \pm 0.75)$	2.213	0.164
$L_S(H_t = (1 \pm 0.5)\Sigma_t)$	0.613	0.144

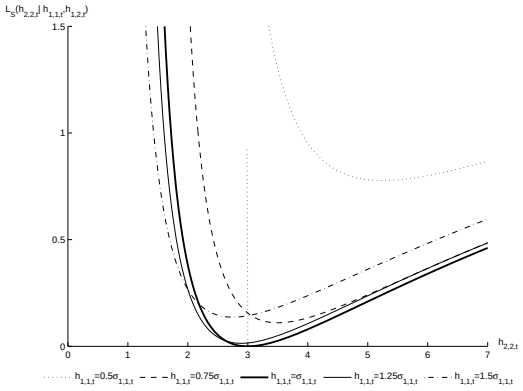
The Stein loss function is therefore asymmetric with respect to over/under predictions, and, in particular, under-predictions are heavily penalized. However, the conditional losses with respect to the variances are symmetric up to a proportionality constant. Figure 2.8(a) and 2.8(b) report L_S as a function of $h_{1,1,t}$ for several values of $h_{2,2,t}$ (with $h_{1,2,t} = \sigma_{1,2,t}$). Figure 2.8(c) reports L_S as a function of both $h_{1,1,t}$ and $h_{2,2,t}$, given $h_{1,2,t} = \sigma_{1,2,t}$, while Figure 2.8(d) reports the contours of the representation in Figure 2.8(c).

Of particular interest is the representation of L_S as a function of the covariance. Note that, if for any given $h_{1,1,t}$ and $h_{2,2,t}$, $h_{1,2,t} = \rho\sqrt{h_{1,1,t}h_{2,2,t}}$ with $\rho \in (-1.1)$. The domain of the conditional loss of $h_{1,2,t}$ is therefore centered at 0 and, using the values suggested above, its representation is given in Figure 2.9.

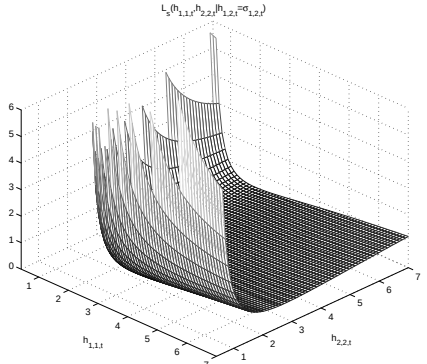
Finally, note that the conditional loss in Figure 2.9 is symmetric about the vertical axis only in the trivial case where Σ_t is diagonal.



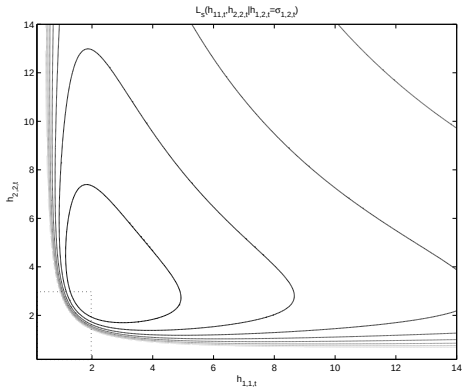
(a) $L_S(h_{1,1,t}|h_{2,2,t}, h_{1,2,t})$



(b) $L_S(h_{2,2,t}|h_{1,1,t}, h_{1,2,t})$



(c) $L_s(h_{1,1,t}, h_{2,2,t}|h_{1,2,t} = \sigma_{1,2,t})$ (loss)



(d) $L_s(h_{1,1,t}, h_{2,2,t}|h_{1,2,t} = \sigma_{1,2,t})$ (contours)

Figure 2.8: Stein loss function

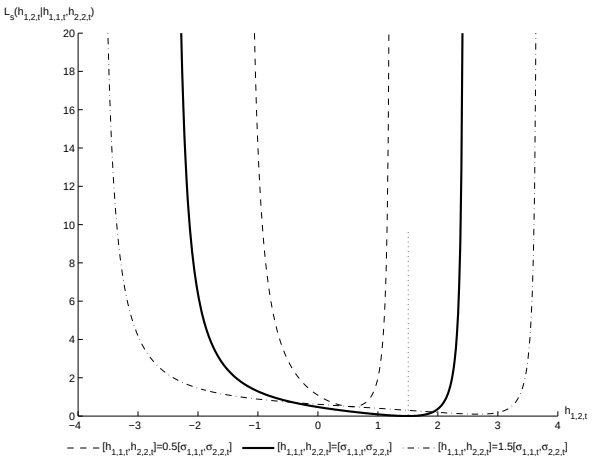


Figure 2.9: $L_s(h_{1,2,t}|h_{1,1,t}, h_{2,2,t})$

Chapter 3

On the Forecasting Accuracy of Multivariate GARCH Models[1]

3.1 Introduction

Most financial applications are multivariate problems with volatility forecasts as one of the inputs. Forecasting sequences of variance matrices is relatively easily done using a multivariate GARCH model, i.e. the conditional variance matrix is modelled as a function of past returns. A large number of multivariate GARCH models have been proposed in the literature, see Bauwens, Laurent, and Rombouts (2006) and Silvennoinen and Terasvirta (2009b) for extensive surveys. The first generation of models, for example the VEC model of Bollerslev, Engle, and Wooldridge (1988a) and the BEKK model of Engle and Kroner (1995), are direct extensions of the univariate GARCH model of Bollerslev (1986). These models are very general and allow for rich and flexible dynamics for the conditional variance matrix. They have been extensively used to model volatility spillovers and in applications such as conditional CAPM and futures hedging. Examples are respectively Karolyi (1995) and Bali (2008). However, being heavily parameterized, they are tractable only for a small number of series, typically lower than four.

More recently, the focus has turned to larger scale problems such as dynamics of correlations between equity and bond returns, portfolio selection and Value at Risk, see Engle (2009) for examples. In these applications, the numerical evaluation of first generation models becomes unfeasible. Both, the number of parameters and the number of operations required to evaluate the likelihood function tend to explode rapidly with the number of

[1] *This chapter has been adapted from Laurent S., Rombouts J.V.K. and Violante F. (2009), On the Forecasting Accuracy of Multivariate GARCH Models. CORE discussion paper 2010-25*

series. Alternative approaches for achieving more manageable and parsimonious specifications have been proposed. Feasible specifications can be obtained by imposing strong parameter restrictions on the BEKK model, which includes the scalar BEKK model and the exponentially weighted moving average model proposed by J.P.Morgan (1996). On the one hand, factor structures like in Engle, Ng, and Rothschild (1990), the orthogonal models of Alexander and Chibumba (1997), Alexander (2000), van der Weide (2002), Lanne and Saikkonen (2007), and Fan, Wang, and Yao (2008) have been proposed. On the other hand, increasing attention has been devoted to conditional correlation models because they can be estimated using a multi-step procedure. The first models have been introduced by Engle (2002a) and Tse and Tsui (2002). Extensions of Engle (2002a) are the asymmetric conditional correlation model of Cappiello, Engle, and Sheppard (2006), the consistent DCC of Aielli (2006) and the sequential DCC model of Palandri (2009).

A priori it is difficult, if not impossible, to identify which model has the best out-of-sample forecasting performance. The evaluation of univariate volatility forecasts is well understood, see Hansen and Lunde (2005), Hansen, Lunde, and Nason (2003), Becker and Clements (2008) among others. In the multivariate setting, although many models are available, from an applied viewpoint, there are no clear guidelines available on model evaluation and selection.

This chapter addresses the selection of multivariate GARCH models in terms of conditional variance matrix out-of-sample forecasting accuracy by providing a large scale analysis. We consider 10 assets from the NYSE, 125 multivariate GARCH specifications, 3 forecast horizons (1, 5 and 20-day ahead), 6 ex-post estimators (proxies) of the conditional covariance matrix, 4 statistical loss functions to measure model performances and 2 statistical tests for identification of the models with superior predictive performances. We also condition the analysis to the forecast sample period. We consider 3 different periods homogeneous in their volatility dynamics (calm, volatile and extremely volatile markets). We proceed as follows. First, we estimate a large variety of models and produce a set of out-of-sample model based forecasts. This can be easily done using standard econometric software packages which are today readily available to the forecaster. Second, we identify the set of models that show superior forecasting performance. These models can then be used either to produce combined forecasts or to select a particular preferred model. Recent somewhat related studies include Clements, Doolan, Hurn, and Becker (2009), Caporin and McAleer (2010) and Chiriac and Voev (2010), though their analysis usually involves a small number of alternative parameterizations and/or small cross sectional dimensions.

Several approaches have been proposed with respect to the inference on the set of superior models. Testing procedure of equal predictive ability (EPA) based on pairwise com-

parison of forecast performances have been introduced by Diebold and Mariano (1995) and then generalized by West (1996), Clark and McCracken (2001), Clark and West (2006) and Clark and West (2007). See West (2006) for a survey. Giacomini and White (2006) develop a general framework that allows to construct tests based on conditional expectations of forecasts and forecasts errors and therefore for a unified treatment of nested and non-nested models. Focussing on the forecast method rather than the model their approach allows to take into account the estimation technique, parameter uncertainty, the choice of the sample size, model misspecification and data heterogeneity. Since our aim is to compare a large number of model based forecasts in order to obtain a joint confidence interval for all possible pairwise comparisons, other alternatives based on multiple comparisons seem to be better suited to our analysis. The reality check test for data snooping of White (2000) and the improved version proposed by Hansen (2005) are based on superior predictive ability (SPA) and allow for multiple comparison against a prespecified benchmark model. Apart from the SPA test, we mainly follow the model confidence set (MCS) approach proposed by Hansen, Lunde, and Nason (2010b). The MCS allows to identify, from a universe of model based forecasts, a subset of models, equivalent in terms of superior ability, which outperform all the other competing models.

To measure out-of-sample forecasting performance, model based forecasts are usually compared to ex-post realizations as they become available. To do this, the forecaster needs to select a loss function and a proxy for the true conditional variance matrix which is unobservable even ex-post. The question arises on which proxy to use and to what extent this substitution affects the forecast evaluation. Building on Hansen and Lunde (2006a) and Patton (2009), in Chapter 2 we have addressed these questions in the case of the comparison of multivariate volatility models using statistical loss functions. They show that the substitution of the underlying volatility by a proxy may induce a distortion in the ranking i.e., the evaluation based on the proxy differs from the ranking that would be obtained if the true target was observable. However, such distortion can be avoided if the loss function has a particular functional form. In this chapter, we use four robust loss functions which allow for various types of asymmetry in the way variances and variance matrix predictions are evaluated. With respect to the choice of the loss function, and within the MCS framework, we find that the Euclidean and Frobenius loss functions (both symmetric) appear to deliver a relatively large MCS, while the asymmetric loss functions, and in particular the Stein loss function, allow to identify sets of superior models which are systematically smaller. These results are consistent with the findings of Clements, Doolan, Hurn, and Becker (2009) in the multivariate setting and Hansen, Lunde, and Nason (2003) in the univariate settings.

Model performances are evaluated using the realized covariance estimator based on intraday returns sampled at the 5 minute frequency which serves as a proxy for the latent covariance matrix. Apart from the popular 5-minute frequency, which, given the characteristics of the assets selected should strike a good compromise between accuracy and microstructure bias (Andersen, Bollerslev, Diebold, and Labys, 1999 and Russell and Bandi, 2004), a robustness check with respect to the choice and the accuracy of the proxy is performed using the realized covariance estimator based on intraday returns sampled at 1 and 30 minutes and a realized kernel estimator based on intraday returns sampled at 1, 5 and 30 minutes, see de Pooter, Martens, and van Dijk (2008). Our results are robust to the choice and the accuracy of the volatility proxy.

As pointed out by Hansen, Lunde, and Nason (2003), the MCS is specific to the set of candidate models and the sample period. By considering not only the full sample (from April 1, 1999 to December 27, 2008, totalling 2486 trading days) but also three sub-samples which are homogenous in their volatility dynamics, we illustrate how highly sensitive the MCS is with respect to the forecast evaluation sample. We find that over the dot-com bubble, the set of superior models is composed of more sophisticated specifications such as orthogonal and dynamic conditional correlations, both with leverage effect in the conditional variances. Over calm periods, a simple assumption like constant conditional correlation and symmetry in the conditional variances cannot be rejected. Over the 2007-2008 financial crisis, accounting for non-stationarity in the conditional variance process generates superior forecasts.

With respect to the longer forecast horizons (5 and 20 day ahead), we find that while the composition of the MCS is in line with the one-step ahead case, the MCS reduces in size. The performances of models with similar properties and structure tend to cluster but differences between clusters increase. This, together with a substantial reduction of the variability of sample performances, due to the smoothness of longer horizon forecasts, makes it easier to separate between superior and inferior models.

In the last part of our study, we assess, using SPA tests, the predictive ability of six popular and parsimonious specifications selected with respect to two dimensions, the multivariate structure and symmetry in the dynamics of the variance processes. We find that the most valid alternative is represented by the Dynamic Conditional Correlation model of Engle (2002a) when coupled with leverage effect in the conditional variances of the marginal processes. This model seems to capture well the dynamics of the conditional variance matrix consistently across the different sample periods. However, in line with the MCS results, simple hypotheses like constant correlation and/or symmetric variance process cannot be rejected over periods of calm markets.

An alternative approach to evaluate variance matrix forecasts is to use an economic loss function such as asset allocation in Engle and Colacito (2006). Other examples are Value-at-Risk forecasting and derivative pricing. See also Voev (2009) for a related setting. However, as pointed out by Patton and Sheppard (2009) the main drawback of an evaluation of volatility forecasts based on economic criteria is that it generally relies on additional and application-specific assumptions, the ordering may not depend exclusively on the accuracy of the conditional variance matrix forecast and the criteria are generally non-robust in the sense that imperfect forecasts can outperform the true conditional variance matrix.

The rest of the chapter is organized as follows. Section 3.2 discusses the multivariate GARCH specifications, the proxies for the conditional variance, the loss functions and the MCS approach. Section 3.3 provides a description of the data and outlines some stylized facts. Section 3.4 presents the results for the multiple comparison based on the MCS and Section 3.5 for the comparison based on the SPA test. Section 3.6 concludes.

3.2 Methodology

In this section, we first introduce the multivariate GARCH models used for the forecasting exercise. Second, we define estimators of the underlying variance matrix used to compare the volatility forecasts. We conclude with a discussion on the properties of the loss functions used to evaluate the forecast errors and with a brief summary of the MCS approach.

3.2.1 Forecasting models set

Consider a N-dimensional vector stochastic process $r_t = \mu_t + \varepsilon_t$ and denote $\Im_{t-1}$ as the information set available at $t-1$. We are interested in modeling its conditional variance matrix $H_t = E(\varepsilon_t \varepsilon_t' | \Im_{t-1})$. Since the conditional mean μ_t is typically of minor importance for GARCH-type models, following Hansen and Lunde (2005) and Becker and Clements (2008), we assume a constant conditional mean for all assets.

We consider parametric specifications for the conditional variance of the multivariate GARCH (MGARCH) type, i.e., H_t is a parametric function of past returns. To control for the number of parameters, we impose covariance or correlation targeting when possible, see Engle and Mezrich (1995). This means that H_t can be expressed in terms of the unconditional variance/correlation and other parameters, provided that the process is covariance stationary. Hence, it is possible to reparameterize the model and replace the unconditional covariance and/or correlation by a consistent estimator before maximizing the likelihood. The targeting ensures a reasonable value of the model-implied unconditional variance and, although it is not a maximum likelihood estimator (therefore asymptotically inefficient),

the long run variance will be consistent even if the MGARCH model is misspecified. This solution also facilitates the numerical optimization of the remaining parameters by reducing the dimensionality of the parameter space. For the properties of the variance targeting estimator and a comparison with the standard quasi-maximum likelihood estimator in the univariate case, see Francq, Horvath, and Zakoian (2009).

We consider several families of MGARCH models which are revealed to be feasible in terms of numerical evaluation when the dimension of r_t is relatively large. According to the classification in Bauwens, Laurent, and Rombouts (2006), among the generalizations of the univariate standard GARCH model, we consider three specifications, namely the diagonal and scalar BEKK of Engle and Kroner (1995) and the multivariate RiskMetrics (RM) model of J.P.Morgan (1996). In the fully parameterized BEKK model with all orders set to 1, the conditional variance is given by

$$H_t = C + A\epsilon_{t-1}\epsilon'_{t-1}A' + BH_{t-1}B', \tag{3.1}$$

where C is a positive definite matrix and A and B are square parameter matrices. The full BEKK specification is not considered as it is not feasible for large cross-sectional dimensions. In the diagonal BEKK (DBEKK), the matrices of parameters A and B are diagonal, while in the scalar BEKK (SBEKK), $A = aI_N$, $B = bI_N$, where a and b are scalars and I_N is the identity matrix. In these models, variance targeting is imposed by setting $H = E(\epsilon_t\epsilon'_t)$ and $C = H - AHA' - BHB'$ which implies $E(H_t) = H$. Note that the scalar BEKK model imposes the same dynamics to all the elements of H_t (and thus is equivalent to the scalar VEC model of Bollerslev, Engle, and Wooldridge, 1988). The RM model has the same parametric form as defined in (3.1) but assumes that the conditional variance matrix is an integrated process, i.e., $a + b = 1$ and $C = 0$, governed by a fixed smoothing parameter, b, equal to 0.96. This model, widely used by practitioners, does not require parameter estimation.

Among the MGARCH models that can be represented as linear combinations of univariate GARCH models, we consider the orthogonal GARCH model of Kariya (1988) and Alexander and Chibumba (1997). In this model, the data are generated by an orthogonal transformation of N (or a smaller number of) uncorrelated factors, f_t, which can be separately defined as any stationary univariate GARCH process. The model can be expressed

as

$$H_t = V^{1/2} P L^{1/2} S_t P L^{1/2} V^{1/2}, \tag{3.2}$$

$$S_t = E_{t-1}(f_t f_t') = diag(\sigma^2_{f_{1,t}}, \ldots, \sigma^2_{f_{N,t}}) \tag{3.3}$$

$$f_t = L^{-1/2} P V^{-1/2} \epsilon_t \tag{3.4}$$

where $V = diag(v_1, ..., v_N)$, with $v_i = E(\epsilon^2_{i,t})$, $i = 1, ..., N$, L and P are $m \times m$ and $N \times m$ matrices of the $m \leq N$ largest eigenvalues of the unconditional correlation matrix and associated orthogonal eigenvectors, respectively. Other specifications belonging to this group are the generalized orthogonal GARCH model by van der Weide (2002) and Lanne and Saikkonen (2007), the full factor GARCH model by Vrontos, Dellaportas, and Politis (2003) and the conditionally uncorrelated components GARCH by Fan, Wang, and Yao (2008). However, these models are computationally challenging when the dimension is large and thus are not considered here.

The last category of models can be viewed as nonlinear combinations of univariate GARCH models. They allow to specify separately N possibly different univariate models for the conditional variances, $\sigma^2_{i,t}$, $i = 1, ..., N$, and a model for the conditional correlation matrix, R_t. The dynamic conditional correlation (DCC) model, in the formulation of Engle (2002a) (DCCE), is defined as

$$H_t = D_t^{1/2} \, R_t \, D_t^{1/2} \tag{3.5}$$

$$R_t = (Q_t \odot I_N)^{-1/2} \, Q_t \, (Q_t \odot I_N)^{-1/2} \tag{3.6}$$

$$Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha u_{t-1} u_{t-1}' + \beta Q_{t-1}, \tag{3.7}$$

where $D_t = diag(\sigma^2_{1,t}, \ldots, \sigma^2_{N,t})$ and $u_{i,t} = \epsilon_{i,t}/\sigma_{i,t}$, $i = 1, ..., N$ defines the devolatilized innovations. The constant conditional correlation (CCC) model of Bollerslev (1990), the asymmetric DCC (ADCC) model of Cappiello, Engle, and Sheppard (2006), the Dynamic Conditional Equi-Correlation (DECO) model of Engle and Kelly (2008) also belong to this family. To ensure positive definiteness, the correlation matrix is modeled as a transformation of a latent matrix Q_t which is a function of past devolatilized innovations.

While the CCC model of Bollerslev (1990) assumes time invariant, but pairwise specific, correlations, which can be estimated by a consistent estimator for the unconditional correlation, the DECO model of Engle and Kelly (2008) assumes that correlations are time varying but equal across the N assets ($R_{ij,t} = \rho \; \forall i \neq j$). Interestingly, under some suitable conditions, the DECO model gives consistent estimators of the correlation dynamics (α, β) in (3.7) even when the equicorrelation assumption is not supported by the data. Since

the hypothesis of equicorrelation is likely to be rejected, in this chapter we use the DECO approach as a technique to estimate the correlation parameters α and β. We then use the DECO estimates to predict and forecast time varying and pairwise specific correlations. The ADCC extends the DCCE by accounting for asymmetries in the correlation dynamics through the additional term $\gamma(u_{t-1}u'_{t-1} \odot 1_{u_{t-1}<0}1'_{u_{t-1}<0})$ in (3.7) where $1_{u_{t-1}<0}$ is a vector of dimension N such that $[1_{u_{t-1}<0}]_i = 1$ if $u_{i,t-1} < 0$ and 0 otherwise. The main drawback of the DCCE, the DECO and the ADCC, is that, under variance/correlation targeting, the choice of the estimator for the long run target $\bar{Q}$ is not obvious as Q_t is neither a conditional variance nor a correlation. Although inconsistent for the target, since the recursion in Q_t does not have a martingale difference representation, Engle and Sheppard (2001) suggest the use of the unconditional expectation of the outer product of devolatilized innovations, arguing that the impact of this choice is very small in practice.

An alternative formulation of the DCC model has been suggested by Tse and Tsui (2002) (DCCT). The conditional correlation R_t is defined as:

$$R_t = (1 - \theta_1 - \theta_2)\bar{R} + \theta_1\Psi_{t-1} + \theta_2 R_{t-1}, \tag{3.8}$$

with Ψ_{t-1} the $N \times N$ correlation matrix of ϵ_τ for $\tau = t-K, t-K+1, \ldots, t-1$ and $K \geq N$. Its i, j-th element is given by

$$\psi_{ij,t-1} = \frac{\sum_{m=1}^{K} u_{i,t-m}u_{j,t-m}}{\sqrt{(\sum_{m=1}^{K} u_{i,t-m}^2)(\sum_{m=1}^{K} u_{j,t-m}^2)}}, \tag{3.9}$$

where u_{it} is defined as above. In the DCCT the correlation matrix is modeled directly and depends on past local correlations of devolatilized innovations. Also in this case, under variance/correlation targeting, the choice of $\bar{R}$ is not obvious. We set $\bar{R}$ equal to the unconditional correlation of the devolatilized innovations.

One of the advantages of the conditional correlation models relies on the fact that the estimation problem can be carried out sequentially. This requires first the estimation of the N conditional variances of the assets, second the estimation of the correlation target and the conditional correlation process. Although inefficient, this procedure is consistent and it dramatically reduces the computational burden of the likelihood. The univariate specification for the conditional variance that we include in the conditional correlation models are ARCH (Engle, 1982), GARCH (Bollerslev, 1986), GJR (Glosten, Jagannathan, and Runkle, 1992), Exponential GARCH (Nelson, 1991), Asymmetric Power ARCH (Ding, Granger, and Engle, 1993), Integrated GARCH (Engle and Bollerslev, 1986), RiskMetrics (J.P.Morgan, 1996), Hyperbolic GARCH (Davidson, 2004) and fractionally integrated GARCH (Baillie,

Bollerslev, and Mikkelsen, 1996). With respect to the number of lags in the models, we fix both the ARCH (p) and the GARCH (q) orders to 1 for the scalar BEKK, multivariate RiskMetrics and the correlation specification in the DCC models. The univariate GARCH models for the conditional variances in the Orthogonal GARCH and DCC specifications include various combinations of the orders p, q. Table 3.1 summarizes the 125 multivariate GARCH configurations we consider in the forecasting exercise.

Table 3.1: Forecasting models set

Conditional correlation type				Orthogonal GARCH				BEKK type			
Corr.	Variance	p	q		Variance	p	q			p	q
	Arch	1,2	-		Arch	1,2	-	BEKK	scalar	1	1
	Aparch	1	1		Aparch	1	1		diagonal	1	1
CCC,	Egarch	0,1,2	1,2	Orth.	Egarch	0,1,2	1,2	RM	-	1	1
DCCA,	Garch	1,2	1,2		Garch	1,2	1,2				
DCCE,	Gjr	1,2	1,2		Gjr	1,2	1,2				
DCCT,	Hgarch	1	1								
DECO	Igarch	1	1								
	Figarch	1	1								
	Rm	1	1								

3.2.2 Proxies for the conditional variance matrix

In our application, the daily realized covariance serves as a proxy for the true conditional variance matrix, Σ_t, when evaluating the forecasting performance of the different MGARCH models. Recent literature suggests several estimators. Examples are the well known realized variance, and its jump robust version bi-power covariation, see Barndorff-Nielsen and Shephard (2004b) and Barndorff-Nielsen and Shephard (2004d), the realized kernel estimators proposed by Zhou (1996), Hansen and Lunde (2006b), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b) which account for serial correlation in the high frequency returns. Parametric models, like vector moving average realized variance can be found in Hansen and Lunde (2008). Intraday returns are defined as $r_i = p_i - p_{i-\Delta}$ for $i = \Delta, 2\Delta, ..., T$, with $1/\Delta$ intervals per day. The daily realized variance ($\hat{\Sigma}^{(\Delta)}$) matrix (Andersen, Bollerslev, Diebold, and Labys, 2003 and Barndorff-Nielsen and Shephard, 2004b) is defined as

$$\hat{\Sigma}^{(\Delta)} = \sum_{i=1}^{\lfloor 1/\Delta \rfloor} r_{i\Delta} r'_{i\Delta}. \tag{3.10}$$

where $\lfloor 1/\Delta \rfloor$ represents the integer part of $1/\Delta$. As the sampling frequency of the intraday returns increases ($\Delta \to 0$), $\hat{\Sigma}^{(\Delta)}$ converges almost surely to Σ_t. See Barndorff-Nielsen and Shephard (2004d), Mykland and Zhang (2006), Andersen, Bollerslev, and Diebold (2002) and related references for details.

The definition of $\hat{\Sigma}^{(\Delta)}$ requires the assumption that intraday returns are uncorrelated. However, failing this assumption, $\hat{\Sigma}^{(\Delta)}$ would result in a biased estimator of Σ_t. Hence, we also consider a simple kernel estimator, defined as

$$\hat{\Sigma}^{(\Delta)}_{AC,q} = \lambda_0 + \sum_{i=1}^{q} (\lambda_{-i} + \lambda_i) + \sum_{i=q+1}^{2q} \left(1 - \frac{i-q}{q+1}\right)(\lambda_{-i} + \lambda_i) \tag{3.11}$$

$$\lambda_q = \begin{cases} \frac{1}{(1-q\Delta)} \sum_{i=q+1}^{\lfloor 1/\Delta \rfloor} r_{i\Delta} r'_{(i-q)\Delta} & q \geq 0 \\ \frac{1}{(1-|q|\Delta)} \sum_{i=|q|+1}^{\lfloor 1/\Delta \rfloor} r_{(i-|q|)\Delta} r'_{i\Delta} & q < 0 \end{cases}. \tag{3.12}$$

This estimator (see Zhou, 1996, Zhang, Mykland, and Ait-Sahalia, 2005, Hansen and Lunde, 2006b and Hansen and Lunde, 2008), based on the Newey and West (1987) variance estimator, is equal to the $\hat{\Sigma}^{(\Delta)}$ plus a term that is a Bartlett-type weighted sum of higher-order autocovariances (lags and leads). More refined realized kernel estimators have been recently proposed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b). Throughout the chapter, unless explicitly mentioned, we will use the $\hat{\Sigma}^{(5min)}$ estimator. $\hat{\Sigma}^{(1min)}$, $\hat{\Sigma}^{(30min)}$, $\hat{\Sigma}^{(1min)}_{AC,q}$, $\hat{\Sigma}^{(5min)}_{AC,q}$ and $\hat{\Sigma}^{(30min)}_{AC,q}$ will serve to check the robustness of the results to different proxies.

3.2.3 Loss functions

At the core of the forecasting comparison is the choice of the loss function. In this chapter, we use the following loss functions

$$L_E = (\sigma_t - h_t)'(\sigma_t - h_t) \tag{3.13}$$

$$L_F = Tr[(\Sigma_t - H_t)'(\Sigma_t - H_t)] \tag{3.14}$$

$$L_S = Tr[H_t^{-1}\Sigma_t] - \log\left|H_t^{-1}\Sigma_t\right| - N \tag{3.15}$$

$$L_d = \frac{1}{d(d-1)} Tr(\Sigma_t^d - H_t^d) - \frac{1}{(d-1)} Tr(H_t^{d-1}(\Sigma_t - H_t)) \quad d \geq 3. \tag{3.16}$$

The first two loss functions belong to a family of quadratic loss functions based on the forecast error. L_E is the Euclidean distance in the vector space of $\sigma_t - h_t = vech(\Sigma_t - H_t)$, where $vech()$ is the operator that stacks the lower triangular portion of a matrix into a vector. Hence, L_E only considers the unique elements of the variance matrix and these

elements are equally weighted. The Frobenius distance, L_F, is defined as the sum of the element-wise square differences of $\Sigma_t - H_t$ and is the natural extension to matrix spaces of the mean squared error. The relevant variable in the comparison is in this case the variance matrix itself and it corresponds to the loss function implied by the matrix Normal likelihood. Although closely related, it differs from L_E for double counting the loss associated to the conditional covariances. The Stein loss function L_S of (James and Stein, 1961) is a scale invariant loss function based on the standardized (in matrix sense) forecast error. It is the loss function implied by the Wishart density.

Note that since L_E only considers the unique elements of the forecast error matrix, it is symmetric in the sense that variances and covariances over/under predictions are equally penalized. On the other hand, L_F equally weights all elements of the forecast error matrix, thus over/under predictions for a given element of the variance matrix are equally penalized, i.e. symmetric with respect to the sign of the forecast error. Though covariances forecast errors are more penalized then variances ones, i.e. asymmetric with respect to the nature of the forecast error. The loss function L_S also considers the whole variance matrix as the variable of interest. This loss function is homogeneous of degree 0 (errors are measured in relative terms) and asymmetric with respect to over/under predictions (in matrix sense) and, in particular, under predictions are heavily penalized. Finally, in the same spirit, L_d also accounts for asymmetry with respect to over/under predictions, but in the opposite direction, i.e. over predictions are penalized instead. L_d also allows to tune the degree of asymmetry, i.e. the weights given to over/under prediction, through the choice of the parameter d, which also represents its degree of homogeneity. In this chapter we set $d = 3$ which implies a mild degree of asymmetry comparable to L_S. See Laurent, Rombouts, and Violante (2009) for further details and examples.

3.2.4 The model confidence set

A review of the MCS procedure has been provided in Chapter 1. In this section we recall the basic ideas and provide some more details on the implementation and the properties of the test.

Let us denote the initial set of h-step ahead conditional variance forecasts $\mathcal{M}^0 : \{H_{i,t+h} \in \mathcal{M}^0 \ \forall i = 1, \ldots, M\}$, where $t = 0, 1, ..., T-1$ and T is the forecast sample size. The MCS procedure is based on a sequence of equivalence tests. The starting hypothesis is that all forecasts in $\mathcal{M}^0$ have equal forecasting performance as measured by a loss function $L_{i,t} = L(\Sigma_t, H_{i,t})$. Let $d_{ij,t} = L_{i,t} - L_{j,t} \ \forall i, j = 1, ..., M$ define the relative performance of forecast i and j. The null hypothesis takes the form $H_{0,\mathcal{M}^0} : E(d_{ij,t}) = 0, \ \forall \ i, j = 1, ..., M$. We use the 'deviation' statistic defined as $T_D = M^{-1} \sum_{i \in \mathcal{M}^0} t_i^2$, where $t_i = \sqrt{T}\bar{d}_i / \sqrt{Var(\sqrt{T}\bar{d}_i)}$

represents the standardized relative performance of forecast i with respect to the average across forecasts, $\bar{d}_i = M^{-1}\Sigma_{j\in\mathcal{M}^0}\bar{d_{ij}}$ and $\bar{d_{ij}} = T^{-1}\Sigma_{t=1}^{T}d_{ij,t}$ is the sample loss difference between forecast i and j. A block bootstrap scheme is used to obtain the distribution under the null. If the null of equal predictive ability is rejected, an elimination rule removes the forecast with the largest t_i. This process is repeated until non-rejection of the null occurs, thus allowing to construct a $(1-\alpha)$-confidence set for the best forecast in $\mathcal{M}^0$.

Being the statistic based on studentized quantities, the analysis of the variance of $\bar{d}_i$ plays a central role for evaluating the informativeness of the MCS. Hansen, Lunde, and Nason (2010b) point out that an inferior forecast (i.e., with sample loss larger the average, $\bar{d}_i > 0$) may be included in the MCS if the variance of $\bar{d}_i$ is large enough, i.e. t_i, is sufficiently small to avoid being discarded. Consider the following decomposition of $Var(\bar{d}_i)$

$$\begin{aligned} Var(\bar{d}_i) &= Var(\bar{L}) + Var(\bar{L}_i) - 2Cov(\bar{L}_i, \bar{L}) \\ &= Var(\bar{L}) + \left(1 + \frac{Var(\bar{L}_i)}{Var(\bar{L})} - 2\sqrt{\frac{Var(\bar{L}_i)}{Var(\bar{L})}}Corr(\bar{L}_i, \bar{L})\right). \end{aligned} \tag{3.17}$$

where $\bar{L}_i = T^{-1}\Sigma_{t=1}^{T}L_{i,t}$ and $\bar{L} = M^{-1}\Sigma_{i\in\mathcal{M}}\bar{L}_i$. Thus, an inferior forecast may enter the MCS if $Var(\bar{L}_i)$ is large enough and/or $Corr(\bar{L}_i, \bar{L})$ is small. However, in some cases the risk of inclusion of poor models does not arise or it only marginally affects the elimination process. This is the case when the set under the null contains only two forecasts. Since $|\bar{d}_1| = |\bar{d}_2|$, then $Var(\bar{d}_1) = Var(\bar{d}_2)$ and thus the variance plays no role in the elimination: if the null is rejected, given the elimination rule defined above the forecast with the best sample performance is always preferred. In the case when the set contains more than two forecasts, an inferior forecast can only be preferred to another inferior forecast with better sample performance but never to a forecast for which $\bar{d}_i < 0$. By the same reasoning, if there is only one forecast in the set with $\bar{d}_i > 0$, it will always be excluded no matter how large its variance is.

3.3 Data and forecasting scheme

We consider stock returns from 10 assets traded in the NYSE as detailed in Table 3.2. The sample period spans March 02, 1988 to December 27, 2008, which amounts to 5226 trading days. The dataset has been cleaned from weekends, holidays and early closing days. Days with missing values and/or constant prices have also been removed. The assets have been selected among the most liquid over the period analyzed in order to minimize microstructure noise components such as non-trading and non-synchronous trading which may induce bias

in the volatility proxy. In particular, the phenomenon of asynchronous trading may lead to the so called Epps effect, i.e., the empirical correlation between stock returns converges to zero as the sampling frequency of the data increases (Epps, 1979). Preliminary analysis show some evidence of this phenomenon when the covariance proxies are computed using 1-minute returns, while it is negligible when the sampling frequency used to compute the covariance proxy is 5-minute frequency or lower.

Following the approach of Andersen, Bollerslev, Frederiksen, and Nielsen (2010), the MGARCH models are estimated using daily open-to-close returns. As explained above, to reduce the computational burden, unconditional means are subtracted from each series of returns before proceeding to the estimation of the 125 multivariate GARCH models by quasi maximum likelihood. The initial estimation sample consists of the first 2740 daily observations, i.e. March 02, 1988 to March 31, 1999. The last 2486 trading days constitute the sample for which we compute 1, 5 and 20-day ahead forecasts. For computational convenience, we re-estimate the model parameters every month (22 days) using a rolling window of the last 2740 observations, while within the 22-day window, the parameters are kept fixed and only the data is updated. This mix of fixed and rolling forecast scheme satisfies the assumptions required by the MCS test (Hansen, Lunde, and Nason, 2010b), allows the comparison of nested models and to account for data heterogeneity (Giacomini and White, 2006, West, 2006), as well as to compare results over sub-samples (since forecasts over different period are conditioned on the most recent information). The proxies for the conditional variance are based on intraday returns computed from 1-minute intervals last trade prices. Since the daily trading period of the NYSE is 6.5 hours, this amounts to 390 1-minute intraday observations per day. The source of the data is the One-Minute Equity Data (OMED) database provided by Tick Data. All programs have been written by the authors using OxMetrics 6 (Doornik, 2009) and are available on request.

Table 3.2: Stock names and descriptive statistics

Name	Sector	Mean	Std. Dev.	Max	Min	Skewness	Kurtosis
Abbott Labs	Health Care	0.085	1.53	10.26	-9.47	-0.05	2.43
BP plc	Energy	0.013	1.17	10.27	-13.96	-0.22	11.83
Colgate-Palmolive	Consumer Stap.	0.073	1.40	16.51	-8.59	0.35	6.48
Eastman Kodak	Consumer Disc.	-0.043	1.74	12.76	-14.13	-0.14	6.42
FedEx Corp.	Industrials	0.068	1.79	12.58	-9.67	0.39	2.93
Coca Cola Co.	Consumer Stap.	0.067	1.38	8.92	-11.08	0.06	3.79
PepsiCo Inc.	Consumer Stap.	0.127	1.44	12.14	-13.78	-0.11	5.97
Procter & Gamble	Consumer Stap.	0.100	1.33	10.50	-9.05	0.00	5.01
Wal-Mart	Consumer Stap.	0.008	1.64	14.75	-8.71	0.27	4.35
Wyeth	Health Care	0.027	1.65	12.32	-15.42	-0.31	6.67

Note. Statistics based on the full sample (estimation plus forecast) of 5229 observations

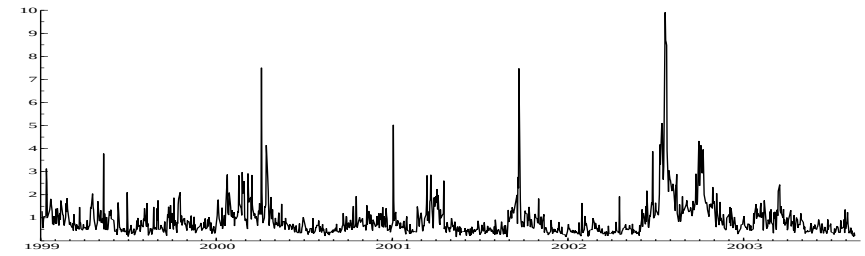
The sample period we consider is characterized by dramatic changes in the volatility dynamics. To investigate the impact of this on the MCS results, the forecasting sample has been divided into three sub-samples. The first sub-sample (1050 obs.) identifies a period of widespread turbulence on the markets. Starting in April, 1999, and ending in March 2003, it includes the peak of the Dot-com boom (until March 2000), the burst and the aftermath of the bubble burst. Peaks in the volatility over this period correspond to the burst of the speculative bubble (March, 2000) and the attack to the twin towers (September, 2001). Towards the end of the period, the turmoil started with the bankruptcy of WorldCom (July, 2002) and ended in October, 2002, with a record low of the Dow Jones Industrial and Nasdaq (5- and 6-years low respectively). The second sub-sample (1080 obs.), from April 2003 to July, 2007, corresponds to a period of stable and upward trending markets. The third sub-sample (356 obs.) corresponds to the recent financial crisis. The beginning of the sample, August, 2007, coincides with the fall of Northern Rock when it became apparent that the financial turmoil, started with the subprime crisis in the US, had spread beyond US's borders. It is also the period when the crisis hits its peak in September and October 2008. To visualize the difference among the three sub periods, Figure 3.1 shows the daily realized variance computed using 5-minute returns of an equally weighted portfolio made of the 10 assets used in the application. It is clear from this figure that the volatility dynamics as well as its scale varies widely between periods.

3.4 Model confidence set results

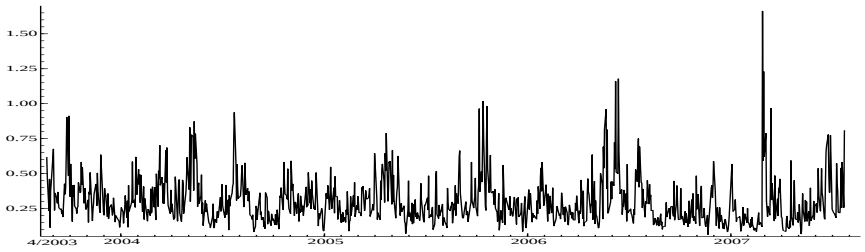
In this section, we describe the MCS results based on one day ahead variance forecasts for four different forecasting samples described in the previous section, i.e. the full sample, the dot-com speculative bubble burst and aftermath, calm markets and the 2007-2008 financial crisis. The section concludes with a discussion on the multi-step ahead (5 and 20 days) forecast evaluation.

3.4.1 Full sample

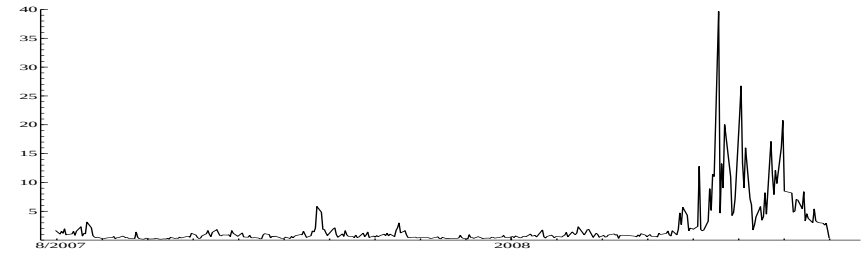
The MCS results for the full forecast sample (2486 obs.) are reported in Table 3.3. To save space, results for the Frobenius loss function (L_F) are not reported. Because if its similarity with the Euclidean loss function (L_E), results based on L_F are very similar in terms of ordering and, in general, we remark that the more conservative L_E MCS always includes the MCS obtained under the L_F loss function. Following Hansen, Lunde, and Nason (2003), we set the confidence level for the MCS to $\alpha = 0.25$. The number of bootstrap samples used to obtain the distribution under the null is set to $10,000$. The values reported for L_E



(a) Dot-com bubble (1/04/99 - 31/03/03) - 1050 Obs.



(b) Calm period (1/04/03 - 31/07/07) - 1080 Obs.



(c) 2007-2008 financial crisis (1/08/07 - 27/12/08) - 356 Obs.

Figure 3.1: Daily realized volatility (computed from 5-min returns) of the 10 asset equally weighted portfolio

and L_3 are the average loss per element of $vech(\Sigma_t - H_t)$, i.e. the total loss is divided by $N(N+1)/2$ and N^2 respectively. For L_S, where the distance is measured in relative terms, the total loss is reported.

The MCS includes 39 models for L_E and is largely dominated by orthogonal and DECO models. With respect to the composition of the MCS, we remark, first, that the family of orthogonal models shows the best sample performances. The flexibility of the orthogonal GARCH model seems therefore able to adapt to a sample that alternates periods of calm with periods of extremely high instability. The results also suggest the rejection of the hypothesis of constant conditional correlation. Second, although the difference is not statistically significant, models allowing for asymmetry/leverage in the conditional variance systematically perform better than symmetric models with Gjr specifications showing the

best sample performances. The same consideration holds for longer versus shorter lags, with higher order models showing in general better sample performances. Third, the MCS includes some specifications that allow for long memory and integrated conditional variances. This is the case for the DECO, DCCA and DCCE with hyperbolic GARCH conditional variances, DECO, DCCA and DCCT with fractionally integrated GARCH conditional variances, DECO with Rm conditional variances and the RM model. Furthermore, if we focus on the sample performances, the specifications allowing for fractional integration or hyperbolic decay of shocks in the conditional variances exhibit the best performances within each family of models.

Table 3.3: MCS on full sample (1/04/99 - 27/12/08)

Euclidean distance (39 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DCCA	Egarch (1,2)	48	3.880	1.165	0.27	1.302	0.999
	Figarch (1,1)	20	3.673	0.521	0.67	1.076	0.996
	Hgarch (1,1)	25	3.720	0.803	0.45	1.052	0.996
DCCT	Figarch (1,1)	38	3.823	1.089	0.30	1.159	0.994
DCCE	Egarch (1,2)	53	3.901	1.207	0.25	1.325	0.998
	Figarch (1,1)	18	3.661	0.406	0.71	1.075	0.996
	Hgarch (1,1)	24	3.719	0.766	0.47	1.057	0.996
DECO	Aparch (1,1)	27	3.735	0.848	0.42	1.111	0.998
	Egarch (0,1)	29	3.742	0.825	0.43	1.172	0.999
	(0,2)	30	3.747	0.877	0.40	1.163	0.999
	(1,2)	33	3.762	0.936	0.37	1.176	0.999
	Figarch (1,1)	2	3.478	0.004	0.94	0.906	0.997
	Garch (1,1)	34	3.768	0.906	0.38	1.171	0.998
	(1,2)	31	3.750	0.965	0.35	1.137	0.999
	(2,1)	28	3.737	0.993	0.34	1.125	0.999
	(2,2)	32	3.759	1.061	0.31	1.159	0.999
	Gjr (1,1)	22	3.692	0.603	0.60	1.090	0.998
	(1,2)	21	3.676	0.706	0.50	1.046	0.999
	(2,1)	14	3.635	0.521	0.67	0.991	0.999
	(2,2)	19	3.667	0.668	0.54	1.036	0.999
	Hgarch (1,1)	5	3.535	0.103	0.89	0.886	0.997
	Igarch (1,1)	35	3.783	1.018	0.33	1.061	0.993
	Rm (1,1)	23	3.699	0.545	0.64	1.117	0.998
Orth.	Aparch (1,1)	7	3.575	0.197	0.89	0.921	0.996
	Egarch (0,1)	17	3.660	0.628	0.58	1.019	0.998
	(0,2)	13	3.623	0.567	0.64	0.945	0.999
	(1,1)	15	3.647	0.735	0.50	0.933	0.998
	(1,2)	12	3.593	0.517	0.67	0.872	0.997
	(2,1)	26	3.726	1.037	0.32	1.066	0.999
	(2,2)	6	3.539	0.175	0.89	0.793	0.996
	Garch (1,1)	16	3.656	0.724	0.50	0.964	0.998
	(1,2)	11	3.589	0.594	0.67	0.870	0.998
	(2,1)	9	3.586	0.549	0.67	0.885	0.999
	(2,2)	8	3.580	0.466	0.69	0.865	0.998
	Gjr (1,1)	10	3.587	0.412	0.73	0.817	0.997
	(1,2)	3	3.507	0.169	0.89	0.713	0.996
	(2,1)	1	3.468	-	1.00	0.672	0.995
	(2,2)	4	3.509	0.116	0.89	0.730	0.996
RM	(1,1)	36	3.810	1.127	0.28	0.967	0.993

Stein distance (10 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Figarch (1,1)	7	3.528	0.346	0.57	0.730	0.932
	Garch (2,1)	10	3.548	1.302	0.25	1.211	0.988
	Igarch (1,1)	3	3.501	0.546	0.69	1.119	0.985
DCCA	Igarch (1,1)	4	3.516	0.680	0.57	1.254	0.986
DCCT	Figarch (1,1)	5	3.518	0.232	0.69	0.743	0.931
	Garch (2,1)	9	3.541	0.880	0.36	1.223	0.989
	Igarch (1,1)	1	3.496	-	1.00	1.130	0.987
DCCE	Figarch (1,1)	6	3.525	0.381	0.57	0.789	0.929
	Garch (2,1)	8	3.535	0.561	0.49	1.255	0.989
	Igarch (1,1)	2	3.500	0.235	0.69	1.228	0.986

L_3 loss function (20 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DECO	Figarch (1,1)	15	102.3	0.844	0.41	1.092	0.999
	Gjr (1,1)	23	105.0	1.063	0.29	1.156	1.000
	Hgarch (1,1)	17	102.5	0.883	0.39	1.082	0.999
	Igarch (1,1)	24	105.1	0.969	0.33	1.142	0.999
Orth.	Aparch (1,1)	10	98.83	0.689	0.50	1.023	0.999
	Egarch (0,1)	14	101.3	0.864	0.43	1.083	1.000
	(0,2)	11	99.98	1.001	0.43	1.033	1.000
	(1,1)	12	100.4	0.848	0.43	1.017	1.000
	(1,2)	6	98.45	0.988	0.43	0.977	0.999
	(2,1)	16	102.4	0.834	0.43	1.104	1.000
	(2,2)	4	95.68	0.609	0.53	0.898	0.999
	Arch (1)	18	103.8	0.923	0.43	0.914	0.990
	Garch (1,1)	13	101.1	0.851	0.43	1.053	1.000
	(1,2)	9	98.60	1.055	0.43	0.984	1.000
	(2,1)	7	98.59	0.928	0.43	0.997	1.000
	(2,2)	8	98.60	1.099	0.43	0.980	1.000
	Gjr (1,1)	5	97.69	0.887	0.43	0.954	0.999
	(1,2)	2	94.06	1.032	0.53	0.852	1.000
	(2,1)	1	91.98	-	1.00	0.801	1.000
	(2,2)	3	94.54	0.782	0.53	0.872	0.999

Note. Rnk: model i's ranking position based on average sample performances (out of 125 models); $\bar{L}_i$: model i's average sample performance; T_D: deviation statistic; p-val: MCS p-value; $VR : V(\bar{L}_i)/V(\bar{L})$ ratio between the variance of model i's loss and the average loss (across models); Corr: $Corr(\bar{L}_i, \bar{L})$ correlation between model i's loss and the average loss (across models).

We next turn to the MCS under the two asymmetric loss functions. Under L_S, the MCS includes 10 models all belonging to the DCC family. Interestingly, the selected models focus on the long memory properties of the conditional variances rather than leverage, asymmetry or even time varying correlation. In fact, the MCS includes models from the CCC, DCCE, DCCA and DCCT families all with fractionally integrated and integrated GARCH or high order GARCH models for the conditional variances, with integrated models showing the best sample performances. When the evaluation is based on the L_3 loss function, the MCS contains 20 models. The MCS is in fact dominated by the orthogonal family of MGARCH, which scores the best sample performances. In line with the previous results, it includes also other specifications, all of which in the DECO family, which allow for long memory and integrated conditional variances.

It is worth noting that the results in terms of MCS are specific to the sample period (and the set of candidate models). As described in Section 3.3, the sample considered is characterized by dramatic changes in volatility dynamics, favoring long memory type models. Furthermore, relatively large average sample performances though close across models indicate that either all models under comparison fail in predicting accurately the conditional variance or, most likely, that this failure refers only to particular periods of time where the inadequacy of the the GARCH estimator is so striking to drive the result even when very long evaluation samples are considered. In the next sections, MCS results are presented for three sub-samples. The aim is to verify to what extent different levels of market instability affect the forecasting performance of the models and the ability of the MCS procedure to separate between superior and inferior models.

3.4.2 Dot-com speculative bubble burst and aftermath

The MCS results are reported in Table 3.4 for the Euclidean (L_E), Stein (L_S) and L_3 loss functions. The MCS under L_E contains 38 models. As expected, there are differences with the MCS obtained for the full sample. First, modelling directly the conditional correlation and accounting for the leverage effect in the conditional variances becomes more important. To be precise, DCC type models with Egarch conditional variances dominate the MCS and show the smallest losses. This result is also confirmed by the fact that the MCS also contains two CCC specifications, both with Egarch dynamics for the conditional variances, which suggests that adequately modelling asymmetry in the conditional variances can in some cases compensate the restrictive assumption of no dynamics in the conditional correlation. Furthermore, the exclusion of other specifications that also specifically account for asymmetry/leverage in the variance, i.e. DCC type models with Aparch and Gjr dynamics for the conditional variances, suggests that the choice of the specific parametrization is

important. Finally, as expected the relative importance of accounting for a (fractionally) integrated variance process, although still present, becomes less noticeable. In this case, we find only 4 specifications (out of the 38 models in the MCS) allowing for long memory and integrated conditional variances (against 10 out of 39 for the full sample).

Table 3.4: MCS on first sub-sample. Dot-com bubble burst (1/04/99 - 31/03/03)

Euclidean distance (38 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Egarch (0,1)	27	2.821	0.985	0.37	1.031	0.999
	(1,1)	41	2.844	1.170	0.29	1.150	0.996
DCCA	Egarch (0,1)	6	2.776	0.335	0.83	0.988	0.999
	(0,2)	18	2.801	0.588	0.65	1.030	0.999
	(1,1)	20	2.806	0.510	0.68	1.117	0.997
	(1,2)	17	2.799	0.545	0.66	1.012	0.999
	Figarch (1,1)	22	2.810	0.372	0.79	0.820	0.989
DCCT	Egarch (0,1)	23	2.811	0.658	0.57	1.026	0.999
	(1,1)	31	2.834	0.779	0.49	1.146	0.996
	Figarch (1,1)	44	2.849	0.839	0.45	0.855	0.989
DCCE	Egarch (0,1)	4	2.769	0.226	0.84	1.011	0.999
	(0,2)	13	2.794	0.404	0.77	1.052	0.999
	(1,1)	19	2.804	0.430	0.75	1.127	0.997
	(1,2)	10	2.783	0.331	0.83	1.019	0.999
	(2,2)	33	2.837	1.028	0.35	1.118	0.997
	Figarch (1,1)	14	2.796	0.343	0.83	0.832	0.990
	Gjr (2,1)	39	2.841	1.242	0.26	0.967	0.994
DECO	Egarch (0,1)	1	2.751	-	1.00	0.948	0.999
	(0,2)	7	2.776	0.290	0.83	0.991	0.999
	(1,1)	5	2.775	0.281	0.84	1.066	0.998
	(1,2)	2	2.760	0.322	0.88	0.961	0.999
	(2,1)	30	2.832	0.721	0.53	1.136	0.996
	(2,2)	21	2.807	0.605	0.62	1.055	0.998
	Figarch (1,1)	26	2.818	0.470	0.71	0.779	0.985
	Gjr (1,1)	43	2.848	1.125	0.30	0.875	0.993
	(2,1)	37	2.838	0.934	0.40	0.900	0.994
Orth.	Aparch (1,1)	3	2.764	0.089	0.88	0.976	0.992
	Egarch (0,1)	12	2.789	0.303	0.83	1.047	0.994
	(0,2)	16	2.797	0.364	0.79	1.083	0.996
	(1,1)	29	2.831	0.847	0.45	1.133	0.997
	(1,2)	25	2.817	0.604	0.62	1.099	0.996
	(2,2)	34	2.837	0.983	0.37	1.135	0.996
	Garch (2,1)	35	2.837	0.723	0.53	1.052	0.991
	(2,2)	24	2.815	0.567	0.65	1.044	0.993
	Gjr (1,1)	8	2.779	0.242	0.83	0.926	0.991
	(1,2)	9	2.780	0.256	0.83	0.933	0.992
	(2,1)	15	2.797	0.392	0.77	0.995	0.995
	(2,2)	11	2.785	0.280	0.83	0.991	0.994

Stein distance (2 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DCCE	Igarch (1,1)	1	3.268	-	1.00	0.999	0.999
DCCT	Igarch (1,1)	2	3.274	1.212	0.27	1.003	1.000
CCC	*Igarch (1,1)*	*3*	*3.283*	-	-	-	
DCCA	*Igarch (1,1)*	*4*	*3.293*	-	-	-	
DCCE	*Figarch (1,1)*	*5*	*3.439*	-	-	-	
DCCT	*Figarch (1,1)*	*6*	*3.444*	-	-	-	
DCCE	*Hgarch (1,1)*	*7*	*3.446*	-	-	-	
DCCT	*Hgarch (1,1)*	*8*	*3.454*	-	-	-	
DCCE	*Rm (1,1)*	*9*	*3.455*	-	-	-	
DCCE	*Egarch (1,2)*	*10*	*3.456*	-	-	-	

L_3 loss function (11 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
Orth.	Aparch (1,1)	1	16.394	-	1.00	0.918	0.999
	Egarch (0,1)	2	16.568	0.887	0.47	0.983	0.999
	(0,2)	3	16.664	0.688	0.47	1.031	1.000
	(1,1)	9	17.035	1.192	0.27	1.117	0.999
	(1,2)	7	16.918	0.996	0.33	1.082	0.999
	(2,2)	11	17.086	1.353	0.27	1.121	0.998
	Garch (2,2)	13	17.235	1.235	0.27	1.007	0.991
	Gjr (1,1)	4	16.733	1.255	0.33	0.876	0.998
	(1,2)	5	16.737	2.285	0.33	0.891	0.999
	(2,1)	8	17.012	1.394	0.27	0.998	0.998
	(2,2)	6	16.797	1.288	0.33	0.985	0.998

Note. See Table 3.3.

The Stein loss function delivers a small MCS. The MCS consists of 2 models, namely the DCCE and the DCCT with integrated GARCH conditional variances. Although the MCS does not overlap with the one found under the symmetric loss function it is clear that when overweighting underpredictions the focus centers on the long memory properties of the conditional variance process. Table 3.4 also reports the best 10 models ordered in terms of sample performances. Although statistically inferior, it is worth noting that the top of the classification is dominated by models accounting for this feature. On the other hand, the MCS under the L_3 loss function includes 8 models, all from orthogonal GARCH family.

Most models account for asymmetry in the variance processes of the components.

3.4.3 Calm markets

Results for the MCS for the second sub-sample are reported in Tables 3.5 and 3.6. With the exception of the Stein loss function, the MCS obtained for this sample is the largest. This is not surprising because this period is characterized by relatively small and slow moving volatility. It is therefore reasonable to expect most of the MGARCH model based forecasts under comparison to show an adequate fit. The average sample performances over this period are close to zero showing a dramatic improvement over the full sample evaluation.

The MCS under L_E contains 74 models, about 60% of the models considered, and includes specifications from all the families of MGARCH models. As a general result, the data does not show evidence of dynamics in the correlation process or asymmetry/leverage or long memory in the conditional variance. Looking at the composition of the MCS, we can draw the following three conclusions. First, DECO type models are excluded from the set of superior models with the exception of DECO-Aparch and DECO-Rm. The decomposition of the variance (columns 7 and 8) together with the ranking position, suggests in both cases that the information content of these models is doubtful. Both models show a relatively small correlation with the average across models, $\bar{L}$. The same remark holds for the DCC type specifications with Rm conditional variances. Second, although only Orth.-Gjr(p,q) models are statistically inferior, the remaining orthogonal specifications show the highest relative variance and smallest correlation with the average loss. Hence, it is likely that the orthogonal models end up in the MCS because the data does not contain sufficient information to infer that these models are inferior within the MCS. Third, similar conclusions hold for CCC/DCC type models with Rm and Gjr(p,q) ($p = 1$ and $q = 1, 2$) conditional variances. In particular, CCC/DCC-Gjr models show, together with by far the poorest sample performances within the MCS, the largest relative variance (in average 25% larger than $Var(\bar{L})$) and the smallest correlation with $\bar{L}$.

We consider now the two asymmetric loss functions. Under L_S, the MCS contains 12 models. In line with previous results, the MCS shows no evidence of particular features in the variance process as dynamics in the correlation process or asymmetry/leverage or long memory in the conditional variance. The set of superior models is dominated by CCC, DCCT and DCCE specifications, with Garch conditional variances, confirming that the hypothesis of constant conditional correlation is difficult to reject. The MCS also includes two asymmetric specifications, i.e. DCCE-Gjr(1,1) and DCCT-Gjr(1,1), although both characterized by weaker sample performances within the MCS. Finally, under L_3, we obtain results similar to L_E both for size and composition of the MCS. However, although

Table 3.5: MCS on second sub-sample. Calm period (1/04/03 - 31/07/07)

Euclidean distance (74 models)							
MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Aparch (1,1)	2	0.328	6.224	0.73	0.884	0.969
	Egarch (0,1)	9	0.345	0.695	0.73	0.975	0.997
	(0,2)	33	0.348	0.456	0.73	1.042	0.992
	(1,1)	18	0.346	0.895	0.73	1.098	0.982
	(1,2)	28	0.347	0.771	0.73	1.061	0.980
	(2,1)	7	0.344	0.967	0.73	1.082	0.987
	Figarch (1,1)	25	0.347	0.447	0.73	0.992	0.996
	Garch (1,1)	50	0.350	0.467	0.67	1.009	0.997
	(1,2)	46	0.350	0.446	0.68	1.021	0.997
	(2,1)	26	0.347	0.439	0.73	1.014	0.997
	(2,2)	11	0.345	0.612	0.73	0.982	0.998
	Gjr (1,1)	91	0.374	0.923	0.37	1.237	0.957
	(1,2)	85	0.372	0.619	0.54	1.260	0.961
	Hgarch (1,1)	55	0.351	0.454	0.68	0.940	0.995
	Rm (1,1)	65	0.356	0.507	0.63	0.990	0.967
DCCA	Aparch (1,1)	4	0.329	3.590	0.73	0.884	0.970
	Egarch (0,1)	20	0.346	0.497	0.73	0.977	0.997
	(0,2)	40	0.349	0.422	0.71	1.044	0.991
	(1,1)	32	0.348	0.626	0.73	1.101	0.981
	(1,2)	38	0.349	0.517	0.73	1.064	0.980
	(2,1)	16	0.346	0.869	0.73	1.084	0.986
	Figarch (1,1)	30	0.347	0.445	0.73	0.992	0.996
	Garch (1,1)	56	0.351	0.497	0.64	1.010	0.997
	(1,2)	53	0.351	0.488	0.65	1.021	0.997
	(2,1)	35	0.348	0.437	0.68	1.015	0.997
	(2,2)	19	0.346	0.504	0.73	0.983	0.998
	Gjr (1,1)	93	0.374	1.134	0.27	1.238	0.957
	(1,2)	89	0.373	0.830	0.41	1.262	0.960
	Hgarch (1,1)	49	0.350	0.439	0.69	0.942	0.995
	Rm (1,1)	64	0.356	0.482	0.65	0.989	0.967
DCCT	Aparch (1,1)	1	0.328	-	1.00	0.884	0.970
	Egarch (0,1)	8	0.345	0.710	0.73	0.975	0.997
	(0,2)	31	0.348	0.471	0.73	1.042	0.991
	(1,1)	17	0.346	1.031	0.73	1.098	0.982
	(1,2)	29	0.347	0.723	0.73	1.061	0.980
	(2,1)	6	0.344	0.959	0.73	1.082	0.987
	Figarch (1,1)	22	0.347	0.490	0.73	0.991	0.997
	Garch (1,1)	48	0.350	0.442	0.67	1.009	0.997
	(1,2)	39	0.349	0.439	0.67	1.021	0.997
	(2,1)	23	0.347	0.452	0.73	1.013	0.997
	(2,2)	10	0.345	0.678	0.73	0.982	0.998
	Gjr (1,1)	88	0.373	0.749	0.46	1.237	0.957
	(1,2)	82	0.372	0.570	0.57	1.261	0.960
	Hgarch (1,1)	43	0.350	0.440	0.67	0.940	0.995
	Rm (1,1)	5	0.340	1.288	0.73	0.957	0.971
DCCE	Aparch (1,1)	3	0.329	3.631	0.73	0.884	0.970
	Egarch (0,1)	15	0.346	0.598	0.73	0.977	0.997
	(0,2)	36	0.349	0.427	0.73	1.045	0.991
	(1,1)	24	0.347	0.813	0.73	1.101	0.981
	(1,2)	34	0.348	0.546	0.73	1.064	0.980
	(2,1)	12	0.345	0.943	0.73	1.084	0.986
	Figarch (1,1)	21	0.347	0.509	0.73	0.992	0.996
	Garch (1,1)	51	0.350	0.472	0.67	1.010	0.997
	(1,2)	47	0.350	0.450	0.68	1.022	0.997
	(2,1)	27	0.347	0.423	0.73	1.013	0.997
	(2,2)	13	0.345	0.587	0.73	0.984	0.998
	Gjr (1,1)	92	0.374	1.023	0.32	1.239	0.957
	(1,2)	86	0.373	0.678	0.50	1.262	0.960
	Hgarch (1,1)	42	0.349	0.434	0.67	0.942	0.995
	Rm (1,1)	63	0.355	0.461	0.67	0.989	0.967
DECO	Aparch (1,1)	14	0.346	0.956	0.73	0.902	0.970
	Rm (1,1)	45	0.350	0.459	0.73	0.974	0.973
Orth.	Aparch (1,1)	37	0.349	0.839	0.73	1.088	0.960
	Egarch (0,1)	44	0.350	0.611	0.73	1.095	0.960
	(0,2)	54	0.351	0.499	0.73	1.091	0.960
	(1,1)	57	0.351	0.450	0.73	1.097	0.960
	(1,2)	41	0.349	0.712	0.73	1.096	0.960
	(2,1)	60	0.352	0.432	0.71	1.087	0.961
	(2,2)	59	0.352	0.430	0.69	1.092	0.963
	Garch (1,1)	58	0.352	0.425	0.72	1.087	0.961
	(1,2)	61	0.352	0.441	0.67	1.090	0.960
	(2,1)	52	0.351	0.550	0.73	1.088	0.961
	(2,2)	62	0.353	0.443	0.67	1.086	0.962
SBEKK	(1,1)	67	0.363	0.534	0.60	0.955	0.952

L_3 loss function (74 models)							
MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Aparch (1,1)	2	0.631	1.090	0.49	0.792	0.910
	Egarch (0,1)	22	0.718	3.320	0.49	1.005	0.997
	(0,2)	52	0.756	1.586	0.49	1.391	0.961
	(1,1)	63	0.784	0.882	0.49	1.777	0.918
	(1,2)	61	0.783	1.015	0.49	1.745	0.917
	(2,1)	57	0.772	1.327	0.49	1.615	0.936
	Figarch (1,1)	31	0.731	0.731	0.49	1.025	0.998
	Garch (1,1)	34	0.732	0.725	0.48	0.998	0.997
	(1,2)	40	0.739	0.752	0.44	1.050	0.998
	(2,1)	42	0.740	0.731	0.46	1.081	0.996
	(2,2)	23	0.718	0.873	0.49	0.980	0.999
	Gjr (1,2)	95	0.876	0.947	0.34	2.098	0.897
	Hgarch (1,1)	49	0.747	0.825	0.39	1.043	0.996
	Igarch (1,1)	93	0.874	0.852	0.38	1.352	0.791
	Rm (1,1)	15	0.674	4.335	0.49	0.823	0.912
DCCA	Aparch (1,1)	4	0.638	6.954	0.49	0.790	0.910
	Egarch (0,1)	29	0.727	0.840	0.49	1.009	0.997
	(0,2)	55	0.767	0.723	0.49	1.407	0.958
	(1,1)	71	0.794	0.748	0.49	1.801	0.915
	(1,2)	70	0.794	0.730	0.49	1.770	0.913
	(2,1)	60	0.781	0.787	0.49	1.635	0.934
	Figarch (1,1)	37	0.737	0.741	0.45	1.031	0.998
	Garch (1,1)	38	0.738	0.767	0.43	1.000	0.997
	(1,2)	47	0.745	0.859	0.38	1.054	0.998
	(2,1)	50	0.748	0.811	0.40	1.086	0.996
	(2,2)	28	0.727	0.736	0.49	0.983	0.999
	Hgarch (1,1)	51	0.749	0.840	0.38	1.051	0.995
	Igarch (1,2)	97	0.882	1.097	0.26	1.345	0.792
	Rm (1,1)	18	0.677	3.386	0.49	0.818	0.914
DCCT	Aparch (1,1)	1	0.631	-	1.00	0.791	0.910
	Egarch (0,1)	24	0.718	4.498	0.49	1.007	0.997
	(0,2)	53	0.757	1.150	0.49	1.398	0.960
	(1,1)	64	0.784	0.934	0.49	1.785	0.916
	(1,2)	62	0.784	0.880	0.49	1.754	0.915
	(2,1)	58	0.773	2.041	0.49	1.622	0.935
	Figarch (1,1)	30	0.731	0.749	0.49	1.027	0.998
	Garch (1,1)	32	0.732	0.733	0.49	0.999	0.997
	(1,2)	39	0.738	0.745	0.44	1.052	0.998
	(2,1)	41	0.740	0.732	0.47	1.083	0.996
	(2,2)	25	0.719	0.872	0.49	0.981	0.999
	Gjr (1,2)	94	0.876	0.894	0.36	2.107	0.897
	Hgarch (1,1)	45	0.744	0.775	0.42	1.047	0.996
	Igarch (1,1)	92	0.873	0.844	0.38	1.349	0.791
	Rm (1,1)	5	0.653	7.324	0.49	0.814	0.914
DCCE	Aparch (1,1)	3	0.636	6.797	0.49	0.790	0.910
	Egarch (0,1)	27	0.724	0.877	0.49	1.010	0.997
	(0,2)	54	0.765	0.769	0.49	1.408	0.958
	(1,1)	68	0.792	0.814	0.49	1.801	0.915
	(1,2)	67	0.791	0.810	0.49	1.771	0.913
	(2,1)	59	0.779	0.843	0.49	1.635	0.934
	Figarch (1,1)	35	0.733	0.728	0.47	1.029	0.998
	Garch (1,1)	36	0.736	0.736	0.46	1.000	0.997
	(1,2)	44	0.743	0.799	0.40	1.054	0.998
	(2,1)	43	0.743	0.759	0.43	1.074	0.996
	(2,2)	26	0.724	0.786	0.49	0.983	0.999
	Gjr (1,2)	96	0.881	1.003	0.31	2.119	0.895
	Hgarch (1,1)	48	0.746	0.786	0.42	1.051	0.996
	Rm (1,1)	12	0.671	4.628	0.49	0.819	0.913
DECO	Aparch (1,1)	21	0.694	1.725	0.49	0.815	0.910
	Rm (1,1)	20	0.686	2.173	0.49	0.827	0.918
Orth.	Aparch (1,1)	6	0.661	6.560	0.49	0.874	0.900
	Egarch (0,1)	8	0.666	5.965	0.49	0.880	0.899
	(0,2)	9	0.666	5.685	0.49	0.867	0.900
	(1,1)	16	0.676	4.113	0.49	0.887	0.899
	(1,2)	14	0.673	4.443	0.49	0.885	0.899
	(2,1)	17	0.676	3.965	0.49	0.879	0.899
	(2,2)	19	0.679	3.378	0.49	0.883	0.903
	Garch (1,1)	7	0.665	5.720	0.49	0.877	0.900
	(1,2)	11	0.670	4.466	0.49	0.880	0.898
	(2,1)	10	0.668	5.554	0.49	0.881	0.899
	(2,2)	13	0.672	4.258	0.49	0.882	0.900
DBEKK	(1,1)	46	0.745	0.841	0.49	0.820	0.898
SBEKK	(1,1)	33	0.732	0.93	0.486	0.837	0.891
RM	(1,1)	56	0.772	0.73	0.466	0.879	0.913

Note. See Table 3.3.

over this sample the type of asymmetry accounted for by L_3 is not statistically relevant, i.e., does not impact on the composition of the MCS, we observe changes in the ordering of the models. For example, the Orthogonal type models included in both the MCSs, ranking between 37th and 62nd under L_E, figure between the 6th and the 19th position under L_3. Given the asymmetry of L_3, we can conclude that Orthogonal models tend to underestimate the conditional variance. The differences in terms of MCS with the outcome obtained under L_E are: i) the inclusion of DCC type specifications with integrated conditional variances, which however show very poor sample performances within the MCS together with the largest relative variances and the smallest correlations with $\bar{L}$; ii) the inclusion of all BEKK type models.

Table 3.6: MCS-second sub-sample. Calm period (1/04/03 - 31/07/07) (Cont.)

Stein distance (12 models)							
MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Garch (1,1)	5	3.180	0.285	0.72	0.948	0.999
	(1,2)	10	3.193	1.253	0.26	1.168	0.996
	(2,1)	3	3.175	0.476	0.74	1.033	0.998
DCCT	Garch (1,1)	6	3.183	0.413	0.61	0.935	0.999
	(1,2)	8	3.191	0.683	0.47	1.154	0.996
	(2,1)	2	3.174	0.265	0.74	1.022	0.998
	(2,2)	7	3.189	1.265	0.29	1.027	0.998
	Gjr (1,1)	16	3.203	1.171	0.26	0.806	0.982
DCCE	Garch (1,1)	4	3.179	0.307	0.74	0.967	0.998
	(1,2)	12	3.194	1.101	0.30	1.198	0.996
	(2,1)	1	3.171	-	1.00	1.065	0.998
	Gjr (1,1)	15	3.201	1.084	0.29	0.834	0.982

Note. See Table 3.3.

3.4.4 2007-08 financial crisis

Results for the MCS for the last sub-sample are reported in Table 3.7. The MCS under L_E contains 39 models. In line with the results obtained for full sample, the MCS is dominated by specifications in the DECO and the Orthogonal families. Other DCC type specifications are included only when they account for long memory and integrated conditional variances. Indeed, in line with the results for the full sample (and in sharp contrast with the Dot-com speculative bubble burst period) modeling long memory and integrated conditional variances becomes more important. Furthermore, although we find models that account for asymmetry/leverage, contrary to the Dot-com bubble burst period, models with Egarch dynamics for the conditional variances are systematically rejected.

Under L_S the results are also consistent with the ones obtained for the full sample, though the MCS is larger (26 models). The models in the MCS belong to the DCC family and account for long-memory in volatility or leverage effect. The non-rejection of some

Table 3.7: MCS-third sub-sample: 2007-2008 financial crisis (1/08/07 - 27/12/08)

Euclidean distance (39 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Hgarch (1,1)	40	17.172	1.034	0.32	1.171	0.995
DCCA	Figarch (1,1)	28	16.345	0.880	0.39	1.099	0.997
	Hgarch (1,1)	21	16.162	0.678	0.50	1.072	0.997
	Rm (1,1)	35	16.954	0.892	0.38	1.264	0.998
DCCT	Figarch (1,1)	43	17.283	1.207	0.25	1.184	0.995
	Hgarch (1,1)	38	17.086	0.992	0.33	1.154	0.995
DCCE	Figarch (1,1)	25	16.305	0.826	0.42	1.097	0.997
	Hgarch (1,1)	22	16.208	0.797	0.44	1.076	0.997
	Rm (1,1)	44	17.376	1.157	0.27	1.307	0.999
DECO	Aparch (1,1)	27	16.317	0.886	0.39	1.122	0.997
	Figarch (1,1)	5	14.919	0.063	0.90	0.922	0.998
	Garch (1,1)	32	16.661	0.884	0.39	1.187	0.997
	(1,2)	29	16.492	0.887	0.39	1.153	0.998
	(2,1)	31	16.583	0.938	0.36	1.141	0.999
	(2,2)	33	16.713	0.962	0.34	1.175	0.999
	Gjr (1,1)	23	16.237	0.828	0.42	1.104	0.998
	(1,2)	16	16.043	0.787	0.44	1.058	0.999
	(2,1)	14	15.879	0.780	0.44	1.001	0.999
	(2,2)	17	16.048	0.892	0.39	1.048	0.999
	Hgarch (1,1)	2	14.816	0.061	0.90	0.899	0.997
	Igarch (1,1)	24	16.275	0.808	0.44	1.071	0.992
	Rm (1,1)	19	16.076	0.444	0.68	1.132	0.998
Orth.	Aparch (1,1)	13	15.791	0.596	0.58	0.918	0.996
	Egarch (0,1)	26	16.308	0.914	0.38	1.020	0.998
	(0,2)	15	16.026	0.890	0.39	0.942	0.999
	(1,1)	20	16.088	0.881	0.39	0.928	0.998
	(1,2)	12	15.757	0.891	0.44	0.868	0.998
	(2,1)	30	16.562	1.068	0.30	1.067	0.999
	(2,2)	6	15.316	0.282	0.79	0.784	0.996
	Arch (2)	71	18.210	1.128	0.28	1.101	0.971
	Garch (1,1)	18	16.052	0.869	0.39	0.963	0.997
	(1,2)	9	15.618	0.827	0.44	0.867	0.998
	(2,1)	10	15.644	0.814	0.44	0.884	0.999
	(2,2)	11	15.666	0.874	0.44	0.861	0.998
	Gjr (1,1)	7	15.391	0.405	0.71	0.812	0.996
	(1,2)	3	14.853	0.120	0.90	0.705	0.997
	(2,1)	1	14.577	-	1.00	0.660	0.996
	(2,2)	4	14.895	0.070	0.90	0.720	0.997
RM	(1,1)	8	15.464	0.153	0.86	0.973	0.992

Stein distance (26 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Aparch (1,1)	21	4.773	0.992	0.32	1.098	0.990
	Egarch (0,1)	14	4.712	0.579	0.46	0.991	0.986
	(0,2)	16	4.716	0.569	0.46	1.006	0.985
	(1,2)	10	4.665	0.587	0.48	0.954	0.990
	Figarch (1,1)	2	4.531	3.442	0.48	0.781	0.942
	Hgarch (1,1)	9	4.663	0.623	0.47	0.784	0.931
DCCA	Aparch (1,1)	30	4.843	1.099	0.29	1.417	0.991
	Egarch (0,1)	20	4.766	0.626	0.44	1.286	0.987
	(0,2)	23	4.787	0.678	0.42	1.313	0.984
	(1,2)	17	4.722	0.586	0.46	1.229	0.991
	Figarch (1,1)	6	4.585	1.143	0.48	0.959	0.939
	Hgarch (1,1)	8	4.631	0.684	0.48	0.861	0.930
DCCT	Aparch (1,1)	19	4.758	0.814	0.37	1.145	0.992
	Egarch (0,1)	11	4.669	0.550	0.48	1.031	0.989
	(0,2)	13	4.678	0.534	0.48	1.048	0.987
	(1,2)	7	4.623	0.636	0.48	0.995	0.993
	Figarch (1,1)	1	4.511	-	1.00	0.816	0.940
	Gjr (1,2)	24	4.802	1.192	0.26	1.214	0.990
	Hgarch (1,1)	4	4.566	4.693	0.48	0.737	0.931
DCCE	Aparch (1,1)	22	4.787	0.743	0.40	1.337	0.991
	Egarch (0,1)	15	4.714	0.578	0.47	1.203	0.987
	(0,2)	18	4.727	0.562	0.47	1.228	0.984
	(1,2)	12	4.671	0.635	0.48	1.151	0.991
	Figarch (1,1)	3	4.543	2.103	0.48	0.927	0.939
	Garch (2,1)	28	4.834	0.903	0.34	1.306	0.981
	Hgarch (1,1)	5	4.578	0.897	0.48	0.824	0.934

L_3 loss function (26 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DECO	Aparch (1,1)	26	682.5	1.116	0.28	1.128	1.000
	Figarch (1,1)	17	660.0	0.769	0.48	1.062	1.000
	Garch (1,1)	29	687.4	1.189	0.26	1.147	1.000
	Gjr (1,1)	24	680.2	1.076	0.30	1.121	1.000
	(1,2)	22	677.6	1.152	0.27	1.105	1.000
	Hgarch (1,1)	15	656.9	0.820	0.48	1.052	1.000
	Igarch (1,1)	21	675.9	0.896	0.38	1.108	0.999
	Rm (1,1)	25	681.9	1.025	0.32	1.136	1.000
Orth.	Aparch (1,1)	10	641.6	0.795	0.48	0.991	0.999
	Egarch (0,1)	16	658.6	0.782	0.47	1.050	1.000
	(0,2)	11	648.8	0.787	0.48	1.001	1.000
	(1,1)	12	650.5	0.808	0.44	0.986	1.000
	(1,2)	8	637.2	1.057	0.48	0.949	0.999
	(2,1)	18	664.1	0.841	0.42	1.072	1.000
	(2,2)	4	617.4	0.667	0.49	0.871	0.999
	Arch (1)	19	665.5	0.962	0.48	0.890	0.989
	(2)	27	684.5	0.952	0.35	1.013	0.994
	Garch (1,1)	13	652.6	0.797	0.48	1.021	0.999
	(1,2)	6	635.9	1.087	0.48	0.955	1.000
	(2,1)	7	637.1	0.993	0.48	0.968	1.000
	(2,2)	9	637.4	1.146	0.48	0.952	1.000
	Gjr (1,1)	5	631.2	0.980	0.48	0.924	0.999
	(1,2)	2	605.8	1.224	0.49	0.825	0.999
	(2,1)	1	590.6	-	1.00	0.776	0.999
	(2,2)	3	609.2	0.858	0.49	0.846	0.999
RM	(1,1)	14	654.4	0.886	0.48	1.044	0.998

Note. See Table 3.3

CCC specifications, which is surprising in this case, shows that adequately modeling the conditional variances of the returns can compensate the loss in forecasting accuracy induced by the restrictive assumption of constant conditional correlation.

For L_3 the results are also in line with the full sample. The MCS contains 26 models and is dominated by orthogonal and DECO specifications with the former showing the best sample performances. Among the DECO specifications included in the MCS we find both evidence of long memory/integrated conditional variances and leverage effect (Aparch and Gjr).

Finally, the average loss over the last sub-sample is much larger than in the first two periods (irrespectively of the choice of the loss function). We conclude first that in turbulent periods GARCH models do not seem to be well suited to adequately estimate the conditional variance. Second, the large losses accumulated over short periods of high instability tend to drive the MCS results even when long forecasting periods are considered. Hence, a careful evaluation of the trade off between forecast sample length (to reduce sampling variability) and the informativeness and accuracy of the selection appears to be crucial in this setting.

3.4.5 Robustness check to the use of alternative proxies

To verify the robustness of our results to the choice of the volatility proxy, we repeat the analysis using $\hat{\Sigma}^{(\Delta)}$, see (3.10), computed using 1 and 30 minute returns and $\hat{\Sigma}^{(\Delta)}_{AC,q=1}$, see (3.12), computed using 1, 5 and 30 minutes returns. The MCS is robust in terms of size and composition to the alternative volatility proxies. In particular, when the proxy is based on higher frequency returns we generally find smaller MCS.

As an example (complete results are available upon request), if we consider the Euclidean distance (L_E), under $\hat{\Sigma}^{(1min)}$ ($\hat{\Sigma}^{(1min)}_{AC,q=1}$) we find 25 (35) models for the full sample, 26 (33) for the dot-com bubble burst period, 60 (71) for the calm period and 47 (38) for the 2007-2008 financial crisis sub-sample. In accordance with the literature, the robustness of these results is implied by the consistency of the loss function. The higher accuracy of the proxy only translates into a lower variability of the sample evaluation of the models which makes easier to effectively discriminate between models. Along the same line, and consistently with the results obtained under $\hat{\Sigma}^{(5min)}$, when the evaluation is based on $\hat{\Sigma}^{(5min)}_{AC,q=1}$ and L_E we find 40 models for the full sample, and 30, 71 and 38 for the three sub-samples respectively. Finally, when we use proxies based on 30 minutes returns we find 41 (40) models for the full sample and 41 (59), 73 (66) and 37 (35) for the three sub-samples respectively.

Our results show that the use of a higher frequency proxy ensures the elimination of uninformative models while the consistency with the results obtained using relatively low frequency proxies shows that the potential microstructure bias is negligible. This result

underlines the value of high precision proxies, in particular when the set of competing models is characterized by a high degree of similarity, see Laurent, Rombouts, and Violante (2009) and Patton and Sheppard (2009).

3.4.6 Multiple comparison based on longer forecast horizons

Results for the multi-step (5 and 20 days) forecast evaluation are reported in Table 3.8 for the L_E loss function. The composition of the MCS is in line with the one-step ahead case. As expected, when the forecast horizon increases the average loss increases, and this irrespectively of the evaluation period or the choice of the loss function. Furthermore, the MCS reduces in size, which seems to be a specific feature here. This result is due to two reasons. First, the performances of models with similar properties and structure tend to cluster (convergence to the same long run variance matrix) but differences between clusters increase (different specifications can imply different levels for the long run variance). Second, longer horizon forecasts are generally smoother, which substantially reduces the variability of the relative performances, $\bar{d}_i$, making it easier to separate between models. The interaction between the two effects is particularly strong for the calm period. The results are in line with the conclusion drawn for the one-step ahead forecast comparison, i.e., the constant correlation hypothesis cannot be rejected, but the size of the MCS reduces to only the CCC-Egarch(2,1) model for 5-day ahead horizon and the CCC-Egarch(2,1), the CCC-Garch(1,2) and the DCCT-Garch(1,2) models for the 20-day ahead horizon. This is because over this period the variability of $\bar{d}_i$ reduces so much that even small differences in performances become highly significant.

In sharp contrast with the one-step ahead case, we find that non stationary models are rejected most of the time for longer forecast horizons. In fact, longer horizon forecasts for these types of models typically exhibit an explosive pattern. An exception is the RiskMetrics type model and the conditional correlation models with Igarch conditional variances, when the evaluation is based on the 2007-2008 financial crisis period. The non exclusion of these specifications indicates the inadequacy of GARCH-type models in periods of extreme market instability. In fact, the k-step ahead forecast for the RiskMetrics type models (except for the correlation component in the DCC-type) is uninformative because it coincides with the 1-step ahead forecast independently from the forecast horizon. This also holds for the models allowing for integrated conditional variances, whose intercept over this period is insignificant in most cases and numerically close to zero (0.003 on average).

Similar considerations and qualitatively the same results, not reported but available upon request, are also found for the L_S and L_3 loss functions.

Table 3.8: MCS on multistep ahead covariance forecasts - Euclidean distance

5-day ahead forecast horizon
Full sample (16 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DCCA	Rm (1,1)	15	4.508	0.907	0.40	1.039	0.998
DECO	Gjr (1,1)	11	4.468	1.082	0.31	1.043	0.999
	(2,1)	14	4.495	1.211	0.25	1.047	0.999
	Rm (1,1)	4	4.366	0.224	0.85	0.962	0.998
Orth.	Aparch (1,1)	2	4.312	0.073	0.94	0.963	0.998
	Egarch (0,1)	6	4.375	0.446	0.69	1.007	0.999
	(0,2)	23	4.569	1.160	0.27	1.141	0.996
	(1,1)	9	4.410	0.608	0.60	1.026	0.999
	(2,1)	10	4.465	0.662	0.57	1.061	0.998
	(2,2)	12	4.470	0.819	0.45	0.985	0.999
	Garch (1,1)	5	4.369	0.677	0.60	0.967	0.998
	(1,2)	13	4.472	1.001	0.35	1.035	0.999
	(2,1)	8	4.399	0.737	0.57	0.974	0.999
	Gjr (1,1)	1	4.300	-	1.00	0.912	0.996
	(1,2)	7	4.390	0.616	0.60	0.990	0.999
	(2,1)	3	4.316	0.052	0.94	0.915	0.997

20-day ahead forecast horizon
Full sample (12 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DECO	Rm (1,1)	9	5.117	1.821	0.25	0.960	0.999
Orth.	Aparch (1,1)	1	5.020	-	1.00	1.006	1.000
	Egarch (0,1)	7	5.102	1.131	0.29	1.026	0.999
	(0,2)	13	5.172	1.209	0.25	1.044	0.999
	(1,1)	10	5.121	1.482	0.25	1.031	0.999
	(2,1)	11	5.144	1.308	0.25	1.034	0.999
	Garch (1,1)	3	5.039	0.351	0.69	0.989	1.000
	(1,2)	5	5.092	1.495	0.25	1.002	1.000
	(2,1)	8	5.105	1.630	0.25	0.990	1.000
	Gjr (1,1)	2	5.024	0.019	0.89	0.967	0.999
	(1,2)	4	5.068	0.963	0.38	0.991	1.000
	(2,1)	6	5.096	1.432	0.25	0.978	0.999

Dot-Com bubble burst (5 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
Orth.	Aparch (1,1)	1	2.959	-	1.00	0.982	1.000
	Egarch (0,1)	3	2.969	0.402	0.65	1.038	1.000
	(0,2)	5	2.996	0.991	0.32	1.095	0.999
	Gjr (1,1)	2	2.967	0.152	0.69	0.909	0.999
	(1,2)	4	2.984	1.126	0.30	0.986	1.000

Dot-Com bubble burst (8 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
Orth.	Aparch (1,1)	8	3.274	0.618	0.552	1.028	1.000
	Egarch (0,2)	5	3.262	0.293	0.818	1.021	1.000
	(1,1)	2	3.251	0.021	0.978	1.020	1.000
	(2,1)	4	3.257	0.072	0.953	1.028	1.000
	(0,1)	1	3.250	-	1.000	1.011	1.000
	Gjr (1,1)	3	3.254	0.015	0.978	0.965	0.999
	(1,2)	6	3.265	0.256	0.808	0.971	1.000
	(2,1)	7	3.273	0.318	0.751	0.963	0.999

Calm period (1 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Egarch (2,1)	1	0.610	-	1.00	-	-

Calm period (3 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Egarch (2,1)	3	0.635	0.598	0.440	1.010	1.000
	Garch (1,2)	1	0.633	-	1.000	0.995	1.000
DCCT	Garch (1,2)	2	0.633	1.614	0.440	0.995	1.000

2007-2008 financial crisis (30 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
DCCA	Garch (2,1)	39	21.09	1.216	0.25	1.138	0.997
	Gjr (1,2)	36	20.98	1.181	0.26	1.133	0.997
	(2,1)	29	20.72	0.959	0.35	1.109	0.998
	Igarch (1,1)	27	20.53	1.156	0.27	1.026	0.998
	Rm (1,1)	7	19.56	0.303	0.76	1.020	0.999
DCCT	Rm (1,1)	26	20.52	0.937	0.37	1.067	0.998
DCCE	Gjr (2,1)	32	20.82	0.985	0.34	1.115	0.997
	Igarch (1,1)	25	20.51	1.074	0.30	1.025	0.998
	Rm (1,1)	15	19.92	0.951	0.40	1.045	0.999
DECO	Garch (1,2)	24	20.51	1.029	0.32	1.073	0.999
	(2,1)	20	20.16	0.918	0.38	1.053	0.999
	Gjr (1,1)	13	19.88	0.915	0.39	1.022	1.000
	(1,2)	19	20.13	0.923	0.38	1.052	0.999
	(2,1)	12	19.83	0.875	0.40	1.030	0.999
	(2,2)	18	20.04	0.918	0.38	0.981	0.998
	Igarch (1,1)	10	19.67	0.814	0.44	0.934	0.996
	Rm (1,1)	1	18.81	-	1.00	0.942	0.999
Orth.	Aparch (1,1)	5	19.48	0.725	0.48	0.938	0.997
	Egarch (0,1)	14	19.89	0.918	0.38	0.980	0.999
	(1,1)	17	20.02	0.901	0.39	1.002	0.999
	(2,1)	22	20.34	0.947	0.36	1.037	0.999
	(2,2)	16	19.94	0.873	0.40	0.961	0.999
	Garch (1,1)	6	19.54	0.873	0.40	0.938	0.997
	(1,2)	21	20.27	1.117	0.28	1.011	0.999
	(2,1)	8	19.58	0.908	0.40	0.949	0.998
	Gjr (1,1)	3	18.93	0.022	0.98	0.889	0.994
	(1,2)	11	19.71	0.906	0.39	0.969	0.999
	(2,1)	4	18.97	0.032	0.98	0.896	0.996
	(2,2)	9	19.64	0.897	0.40	0.938	0.998
RM	(1,1)	2	18.83	0.001	0.98	0.887	0.994

2007-2008 financial crisis (18 models)

MCS		Rnk	$\bar{L}_i$	T_D	p-val	VR	Corr
CCC	Rm (1,1)	18	23.86	1.105	0.32	1.046	0.999
DCCA	Garch (2,1)	20	23.99	1.208	0.27	1.049	0.998
	Igarch (1,1)	15	23.56	1.009	0.37	0.941	0.999
	Rm (1,1)	3	22.87	0.394	0.73	1.009	1.000
DCCT	Rm (1,1)	12	23.48	1.117	0.37	1.031	0.998
DCCE	Igarch (1,1)	13	23.54	1.098	0.37	0.939	0.999
	Rm (1,1)	5	23.13	0.579	0.58	1.023	1.000
DECO	Garch (2,1)	9	23.43	1.149	0.37	1.020	0.999
	Igarch (1,1)	6	23.17	0.390	0.73	0.906	0.997
	Rm (1,1)	1	22.64	-	1.00	0.993	0.999
Orth.	Aparch (1,1)	8	23.32	1.194	0.37	1.041	1.000
	Garch (1,1)	7	23.29	1.177	0.37	1.023	1.000
	(2,1)	14	23.56	1.250	0.26	1.022	1.000
	Gjr (1,1)	4	22.97	0.465	0.73	1.003	0.999
	(1,2)	10	23.45	0.991	0.37	1.026	0.999
	(2,1)	11	23.45	1.057	0.37	1.011	0.999
	(2,2)	16	23.67	0.989	0.37	0.995	0.999
RM	(1,1)	2	22.73	0.048	0.82	0.964	0.998

Note. See Table 3.3.

3.5 Setting a benchmark: the predictive ability of the DCCE

In this section, we focus on the predictive ability of a predefined benchmark model with respect to all other alternatives. As benchmarks we choose simple and parsimonious specifications and take into account two dimensions: the assumption on the multivariate structure (CCC, DCCE and Orthogonal) and on the dynamics of variance of the marginal processes/principal components (Garch(1,1) and Egarch(0,1)). The CCC-Garch(1,1) model represents the simplest alternative and allows to test simple hypotheses such as constant correlation and symmetric variances for the marginal processes. The choice of the DCCE among the DCC specification introduced in Section 3.2.1 is not coincidental: this model has been increasingly popular because of its flexibility and straightforward interpretation. The DCCE-Garch(1,1) therefore serves as a benchmark to assess whether relaxing the assumption of constant correlation is sufficient to improve predictive ability. Finally, the Orthogonal-Garch(1,1) model represents a simple and parsimonious alternative to direct modeling of the dynamics of the conditional covariance and correlation. In a univariate setting, Hansen and Lunde (2005) suggest that the absence of leverage effect is likely to be rejected on stock market returns. To validate this result in the multivariate framework, we also couple the three multivariate models with the Egarch(0,1) specifications for the conditional variance processes.

The predictive ability of our benchmarks is evaluated using the test for superior predictive ability (SPA) proposed by Hansen (2005). This test generates the probability distribution of the model which performs best relative to the benchmark. Using the notation introduced in Section 3.2.4, let us define $d_{0j,t} = L_{0,t} - L_{j,t}$, $j = 1, ..., M$, the relative performance of model j with respect to the benchmark model (indexed by 0). Under reasonable assumptions $\lambda_j = E[d_{0j,t}]$ is well defined. The null hypothesis is expressed with respect to the best alternative model, i.e. $H_{0,M} : \max_{j \in M} \frac{\lambda_j}{\omega_j} \leq 0$, where ω_j^2 denotes the asymptotic variance of λ_j. The test statistic is $\sqrt{T}\left[\max_{j \in M} \frac{\bar{d}_{0j}}{\hat{\omega}_j}\right]$ where $\bar{d}_{0j} = T^{-1} \sum_{t=1}^{T} d_{0j,t}$ is the sample loss differential between the benchmark and model j. P-values for the test are obtained by bootstrap.

The results for the six different benchmarks are reported in Tables 3.9 and 3.10. Consistently with the MCS results in Section 3.4, the hypothesis of constant correlation (Benchmark 1 and 4), as well as of symmetric dynamics for the variance matrix (Benchmark 2 and 5) is always rejected except when forecasts are compared over calm periods. However, the hypothesis of symmetric dynamics for the variances of the assets returns considered is rather weak. Evidence of the leverage effect is much stronger (e.g., Benchmark 5) when the com-

parison is taken over periods of market instability. Also, allowing for dynamic correlation significantly improves models' forecasting ability.

Table 3.9: SPA test (symmetric dynamics for the marginal variances)

Benchmark 1: CCC-Garch(1,1)									
	L_E			L_S			L_3		
	p_L	p_C	p_U	p_L	p_C	p_U	p_L	p_C	p_U
Full sample	0.003	0.003	0.003	0.000	0.000	0.000	0.026	0.027	0.027
Dot-com bubble	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Calm period	0.018	0.020	0.023	0.434	**0.817**	0.963	0.170	**0.211**	0.259
07-08 financial crisis	0.015	0.016	0.016	0.000	0.000	0.000	0.019	0.019	0.019
Benchmark 2: DCCE-Garch(1,1)									
	L_E			L_S			L_3		
	p_L	p_C	p_U	p_L	p_C	p_U	p_L	p_C	p_U
Full sample	0.061	0.064	0.067	0.000	0.000	0.000	0.095	0.098	0.101
Dot-com bubble	0.001	0.002	0.002	0.000	0.000	0.000	0.003	0.003	0.003
Calm period	0.108	**0.115**	0.170	0.384	**0.825**	0.982	0.092	**0.102**	0.141
07-08 financial crisis	0.023	0.024	0.024	0.008	0.009	0.009	0.037	0.038	0.038
Benchmark 3: Orth.-Garch(1,1)									
	L_E			L_S			L_3		
	p_L	p_C	p_U	p_L	p_C	p_U	p_L	p_C	p_U
Full sample	0.087	**0.118**	0.120	0.000	0.000	0.000	0.191	**0.276**	0.280
Dot-com bubble	0.070	0.081	0.090	0.000	0.000	0.000	0.031	0.034	0.037
Calm period	0.001	0.002	0.003	0.000	0.000	0.000	0.010	0.013	0.021
07-08 financial crisis	0.257	**0.321**	0.332	0.003	0.003	0.003	0.357	**0.488**	0.494

Note. p_C consistent p-value, p_L and p_U lower and upper bound for the consistent p-value respectively. See Hansen (2005) for further details. Consistent p-values in bold indicate the non-rejection of the null at confidence level $\alpha = 0.10$.

With respect to the type of multivariate model, the Orthogonal approach (in particular with leverage) exhibits superior performance exclusively over turbulent periods while it is systematically outperformed over calm periods. As underlined in Section 3.4 the fact that this model is preferred under the L_3 criterion suggests that it is likely to underestimate the covariance matrix (Benchmark 3 and 6).

In this application, the most valid specification is the DCCE-Egarch(0,1). It captures well the dynamics of the covariance matrix across the different samples. Its performances are not statistically worse than any of the 124 competing models, both when considering the full sample or any of the sub-samples. For the 2007-08 financial crisis period the null is rejected under L_E but not under L_S, i.e. the DCCE-Egarch(0,1) possibly tends overestimate the variance matrix during periods of extreme market instability.

Table 3.10: SPA test (asymmetric dynamics for the marginal variances)

	L_E			L_S			L_3		
	p_L	p_C	p_U	p_L	p_C	p_U	p_L	p_C	p_U
Benchmark 4: CCC-Egarch(0,1)									
Full sample	0.014	0.016	0.016	0.000	0.000	0.000	0.043	0.043	0.044
Dot-com bubble	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Calm period	0.100	**0.164**	0.237	0.000	0.000	0.000	0.242	**0.423**	0.547
07-08 financial crisis	0.016	0.016	0.016	0.046	0.056	0.085	0.018	0.019	0.019
Benchmark 5: DCCE-Egarch(0,1)									
Full sample	0.100	**0.115**	0.136	0.001	0.002	0.002	0.082	0.084	0.086
Dot-com bubble	0.403	**0.746**	0.909	0.000	0.000	0.000	0.080	0.091	0.131
Calm period	0.154	**0.227**	0.386	0.000	0.000	0.000	0.023	0.029	0.035
07-08 financial crisis	0.035	0.037	0.037	0.165	**0.235**	0.459	0.035	0.036	0.036
Benchmark 6: Orth.-Egarch(0,1)									
Full sample	0.243	**0.372**	0.400	0.000	0.000	0.000	0.297	**0.524**	0.546
Dot-com bubble	0.341	**0.522**	0.597	0.000	0.000	0.000	0.217	**0.723**	0.838
Calm period	0.004	0.006	0.009	0.000	0.000	0.000	0.000	0.000	0.000
07-08 financial crisis	0.189	**0.220**	0.229	0.003	0.003	0.004	0.336	**0.489**	0.503

Note. See Table 3.9.

3.6 Conclusion

Several multivariate GARCH models exist in the literature. However, from an applied viewpoint no guidelines are available on forecasting performances evaluation and model selection. We apply the model confidence set approach (MCS), which allows to isolate superior models in terms of predictice ability, to 125 multivariate GARCH model based forecasts. We consider 10 assets from NYSE for which we forecast 1, 5 and 20-day ahead conditional variance matrices from April 1, 1999 to December 27, 2008. The evaluation is based on two symmetric and two asymmetric loss functions and the ex-post underlying volatility is approximated by the realized covariance estimator based on intraday returns sampled at 5 minute frequency.

In line with recent literature, we find the Euclidean and Frobenius loss functions (both symmetric) to deliver relatively large MCS, from about one half to one fourth of the total number of models, while the two asymmetric loss functions identify sets of superior models systematically smaller. The MCS is composed of sophisticated specifications such as orthogonal and dynamic conditional correlation (DCC), both with long memory in the conditional variances. With respect to the properties of the loss function, we conclude that Orthogonal and DECO models tend to underestimate the conditional covariance, the DCC

of Engle (2002a) (as well as its asymmetric version) and the DCC of Tse and Tsui (2002) tend to overestimate.

We illustrate how sensitive the MCS is with respect to the forecast sample under investigation by considering not only the full forecast sample but also by investigating sub-samples which are homogenous in their volatility dynamics. We find that over the dot-com bubble burst and aftermath period, the set of superior models is composed by rather sophisticated models such as DCC and Orthogonal, both with leverage effect in the conditional variances of returns and principal components, respectively. Over calm periods, a simple assumption like constant conditional correlation and symmetry in the conditional variances cannot be rejected. Finally, over the 2007-2008 financial crisis, accounting for non-stationarity in the conditional variance process significantly improves models' forecasting performances.

With respect to the longer forecast horizons (5 and 20 day ahead), we find that while the composition of the MCS is in line with the one-step ahead case, the MCS reduces in size. The performances of models with similar properties and structure tend to cluster but differences between clusters increase. This, together with a substantial reduction of the variability of sample performances, due to the smoothness of longer horizon forecasts, makes it easier to separate between superior and inferior models.

Focussing on the DCC class of models we can draw the following conclusions. First, the DECO model, which is estimated under the assumption of cross sectional equicorrelation, delivers superior forecasts over periods of market instability, but performs rather poorly during calm periods. Second, modeling the asymmetric response of shocks in the conditional correlation with a single parameter does not seem to significantly improve models' forecasting performances with respect to the standard DCC of Engle (2002a). Third, when comparing the DCC of Engle (2002a) with the DCC of Tse and Tsui (2002), we can conclude that, although statistically equivalent in terms of forecasting ability, while the first shows better sample performances over turbulent periods, the second performs better over calm periods. Fourth, we find that the most valid specification is represented by the DCC model of Engle (2002a) when coupled with leverage effect in the conditional variances of the marginal processes. This model captures well the dynamics of the variance matrix consistently across the different sample periods. The latter result is confirmed by the Superior Predictive Ability (SPA) test. The null hypothesis that the DCC of Engle (2002a) with exponential GARCH dynamics is not outperformed by the other 124 specifications cannot be rejected at standard levels irrespectively of the evaluation period.

This chapter considers only forecasts based on multivariate GARCH models. It would be interesting to compare the performances of this class of volatility models with other approaches such as heterogeneous autoregression based on historical values of ex-post measures

of the conditional variance as in the model proposed by Corsi (2009), models that combine ARMA structures for both the conditional variance and realized measures of volatility as in Hansen, Huang, and Shek (2010) or yet multivariate stochastic volatility (Gourieroux, Jasiak, and Sufana, 2009) and regime switching models as in Silvennoinen and Terasvirta (2009a). Other problems like the evaluation of forecast performances of correlation matrices and high dimensional applications (hundreds of series) also merit more attention.

Chapter 4

The diffusion Limit of Dynamic Conditional Correlation Models[1]

4.1 Introduction

Continuous and discrete time volatility models are often considered as two competitive views to modeling financial time series. Thanks to the analytical tractability ensured by Ito calculus, continuous time models have played a central role in theoretical finance. The continuous time setting allows to have a deeper understanding of the properties of the corresponding discrete time model and to assess probabilistic and statistic properties of discrete time sequences such as stationary, moment finiteness or distributional results which are otherwise intractable in discrete time, see Nelson (1990), Nelson and Foster (1994) and Nelson (1994) for some examples.

From an applied viewpoint, inference on continuous time parameters of stochastic volatility models represents an important issue. The intractable likelihood functions and the unobservable volatility process require sophisticated estimation procedures. Several estimation methods have been proposed, such as the simulation based method of moments, Duffie and Singleton (1993), the quasi-indirect inference of Broze, Scaillet, and Zakoian (1998) or Bayesian Markov chain Monte Carlo methods, Jones (2003). Bollerslev, Engle, and Nelson (1994) and Ghysels, Harvey, and Renault (1996) provide exhaustive surveys on stochastic volatility models. For this reason, discrete time volatility models have been most often preferred by the applied econometrician. Rather than estimating and forecasting with a diffusion model observed at discrete points in time, it is in fact often easier to use a discrete

[1] *This chapter has been adapted from Hafner, C.M., Laurent S. and Violante F. (2009), The diffusion Limit of Dynamic Conditional Correlation Models.*

model directly.

The theory of convergence of discrete time Markov sequences towards continuous time diffusion process, see Stroock and Varadhan (1979), Kushner (1984) and Ethier and Kurtz (1986b), provides the theoretical foundations to establish mutual complementarities, possible inter-changeability and connections between the two approaches. Nelson (1990) provides conditions ensuring the weak convergence of a discrete time Markov chain, defined by a system of stochastic difference equations, towards a diffusion, defined by a system of stochastic differential equations. The proposed approach, dubbed moment matching procedure, requires the convergence, as the interval between observations shrinks to zero, of a number of conditional moments of the increments of the system of interest, i.e., log-price and variance, to well defined limits at an appropriate rate. In the context of GARCH-type models, Nelson (1990), illustrates the convergence through various GARCH specifications. This approach has been used by Duan (1997) to derive the diffusion limit of the Augmented GARCH model, by Fornari and Mele (1997) to study the continuous time behavior of the class of non linear ARCH models proposed by Ding, Granger, and Engle (1993), by Alexander and Lazar (2005) to derive the diffusion limit of a weak GARCH process and in a related setting by Trifi (2006) to illustrate the convergence results for the CEV-ARCH model of Fornari and Mele (2006) and the CMSV model of Jeantheau (2004) and Hobson and Rogers (1998).

The advantage of this type of approximation lies essentially in estimating and forecasting. Considering the discrete time model as a diffusion approximation allows to infer the parameters of the diffusion model by the parameter estimates of a discrete time GARCH-type model. In this chapter we propose a natural alternative to the direct estimation of the diffusion parameters, consisting in inferring the diffusion parameters by means of a tractable likelihood function of a discrete time multivariate GARCH process which can be considered as an approximation of the diffusion process. Since the resulting likelihood function refers to a process converging in distribution to the solution of a system of stochastic differential equations that is not an Euler approximation of it, following Fornari and Mele (2006), we call the resulting criterion quasi-approximated likelihood function (QAML). Requiring a minimal computational effort, this approach has been advocated by many authors to avoid more sophisticated estimation procedures, see Engle and Lee (1996), Lewis (2000), Barone-Adesi, Rasmussen, and Ravanelli (2005) and Stentoft (2008) among others. This advantage becomes even more striking in the multivariate case. Multivariate volatility models within the conditional correlation class can be estimated easily and inference results to be accurate even for large dimensions. Clearly this approach also has some drawbacks. First, the QAML estimator is not necessarily consistent because the discrete time approximation is typically not closed under temporal aggregation, see Drost and Nijman (1993) and Drost

and Werker (1996). Second, except for some specific cases, a one-to-one correspondence between weak convergence of the discrete time model and disaggregation from its diffusion is not guaranteed, see Corradi (2000) and Wang (2002).

In this chapter we focus on conditional correlation models with GARCH dynamics for the variances of the marginal processes. Apart from Nelson (1994) in the context of asymptotic filtering theory, to our knowledge a comprehensive investigation of the the relationship between multivariate discrete and continuous time models, and in particular to conditional correlation models, has not been addressed yet.

We recover the diffusion limit of a modified version of the well known Dynamic Conditional Correlation (DCC) model of Engle (2002b), dubbed consistent DCC (cDCC), proposed by Aielli (2006). The cDCC is based on a more natural representation of correlation driving process which, unlike the standard DCC model, preserves the martingale difference property. For this specification, we point out the existence of a degenerate diffusion limit. The degeneracy of the cDCC-GARCH diffusion limit is due to the particular structure of the discrete time model in which the noise propagation system of the variances and the one of the correlation driving process are perfectly correlated. This structure is preserved in the diffusion limit which is characterized by a diffusion matrix singular by construction. More precisely, the diffusion of the variances and of the diagonal elements of the correlation driving process are pairwise governed by the same Brownian motion.

As a particular case, we also consider the Constant Conditional Correlation model (CCC) of Bollerslev (1990), which can be obtained from the cDCC under suitable parameter restrictions. The CCC-GARCH model is particularly interesting because, unlike the cDCC-GARCH process, it admits a non-degenerate diffusion and, in the bivariate specification, a closed form solution for the diffusion limit.

Finally, we propose and discuss different sets of conditions regarding the speed of convergence of parameters of the cDCC-GARCH model. In this way, we are able to recover other types of degenerated diffusions which are characterized by a stochastic price process while variances and/or correlations remain time varying but deterministic. In the same spirit of Corradi (2000), we then discuss what type of process can be obtained as Euler approximation of the different diffusions recovered.

The chapter is completed by a comprehensive simulation study evaluating the accuracy of our convergence results.

The chapter is organized as follows. In Section 4.2 we present the theorem of weak convergence of discrete time Markov chains. In Section 4.3 we study the continuous time behavior of the cDCC and CCC models. We also present the degenerate diffusions induced by a reparameterization of the convergence conditions. In Section 4.4, we illustrate through

a Monte Carlo simulation our convergence results. In Section 4.5 we conclude and discuss directions for further research. All proofs are provided in the Appendix.

4.2 Weak convergence of stochastic systems

In this section we introduce a set of conditions for the convergence of a system of discrete time stochastic difference equations towards system of stochastic differential equations based on the work of Stroock and Varadhan (1979), Kushner (1984), Ethier and Kurtz (1986b) and Nelson (1990).

Let us define $D\left([0,\infty),\mathbb{R}^N\right)$ the space of cadlag mappings from $[0,\infty)$ into $\mathbb{R}^N$ and $B\left(\mathbb{R}^N\right)$ the Borel sets on $\mathbb{R}^N$. P_h is the probability measure on $D\left([0,\infty),\mathbb{R}^N\right)$ for each $h>0$. Let M_{kh} be the σ-field generated by $(kh,\ {}_hX_0,\ {}_hX_h,\ {}_hX_{2h},...,\ {}_hX_{kh})$, where ${}_hX_{kh}$ is a N-dimensional discrete time Markov chain indexed by $h>0,\ k\in\mathbb{N}$, with ν_h a probability measure on $\left(\mathbb{R}^N,B(\mathbb{R}^N)\right)$, such that $P_h[{}_hX_0\in\Gamma]=\nu_h(\Gamma)$ for any $\Gamma\in B\left(\mathbb{R}^N\right)$ defines the distribution of the starting point ${}_hX_0$, and with transition probabilities $P_h[{}_hX_{(k+1)h}\in\Gamma|M_{kh}]=\Pi_{h,kh}({}_hX_{kh},\Gamma)$, $\forall k\in\mathbb{N}$, $\Gamma\in B\left(\mathbb{R}^N\right)$. Let us now define ${}_hX_t$ a continuous time process, formed from the discrete time process ${}_hX_{kh}$ as a cadlag step function with jumps at $h,2h,3h,...$, such that $P_h[{}_hX_t={}_hX_{kh},kh<t<(k+1)h]=1$. Finally, let X_t be a continuous time process obtained from ${}_hX_t$ by shrinking the frequency h towards zero. X_t represents the limiting diffusion process to which, under Assumption 4.1 to 4.4 given below, the discrete time process ${}_hX_t$ weakly converges as $h\to 0$.

For the convergence results we need the following assumptions

Assumption 4.1. *There exist a continuous mapping $a(x,t)$ from $\mathbb{R}^N\times[0,\infty)\to N\times N$ space of the symmetric positive semi-definite matrices and a continuous measurable mapping $b(x,t)$ from $\mathbb{R}^N\times[0,\infty)\to\mathbb{R}^N$ such that for all $r>0$ and $(k-1)h<t<kh$*

$$a)\ \lim_{h\to 0}\sup_{\|x\|\leqslant r}\left\|h^{-1}E\left[{}_hX_{(k+1)h}-{}_hX_{kh}\middle|{}_hX_{kh}=x\right]-b(x,t)\right\|=0, \tag{4.1}$$

$$b)\ \lim_{h\to 0}\sup_{\|x\|\leqslant r}\left\|h^{-1}E\left[({}_hX_{(k+1)h}-{}_hX_{kh})({}_hX_{(k+1)h}-{}_hX_{kh})'\middle|{}_hX_{kh}=x\right]-a(x,t)\right\|=0, \tag{4.2}$$

$$c)\ \exists\delta>0:\lim_{h\to 0}\sup_{\|x\|\leqslant r}\left\|h^{-1}E\left[\left|({}_hX_{(k+1)h}-\ {}_hX_{kh})_i\right|^{2+\delta}\middle|{}_hX_{kh}=x\right]\right\|=0,$$

where $(.)_i$ is the i^{th} element of the vector $({}_hX_{(k+1)h}-{}_h\,X_{kh})$.

Assumption 4.2. *There exists a continuous mapping $\sigma(x,t)$ from $\mathbb{R}^N\times[0,\infty)\to N\times N$ space of matrices such that for all $x\in\mathbb{R}^N$ and $t\geq 0$, $a(x,t)=\sigma(x,t)\sigma(x,t)'$.*

Assumption 4.3. *${}_hX_0$ converges in distribution, as $h \to 0$, to a random variable X_0 with probability measure ν_0 on $(\mathbb{R}^N, B(\mathbb{R}^N))$.*

Assumption 4.4. *$\nu_0, b(x,t), a(x,t)$ uniquely specify the distribution of a diffusion process X_t with initial distribution ν_0, drift vector $b(x,t)$ and diffusion matrix $a(x,t)$.*

We can now state the following theorem for the weak convergence of discrete time stochastic sequences.

Theorem of weak convergence (Nelson, 1990). *Under Assumptions 1 to 4, the sequence of discrete time process ${}_hX_{kh}$ indexed by $h > 0$ $k \in \mathbb{N}$, converges in distribution, as $h \to 0$, to the diffusion process X_t solution of the system of stochastic differential equations*

$$dX_t = b(X_t, t)\,dt + \sigma(X_t, t)\,dW_t, \tag{4.3}$$

where dW_t is a N-dimensional vector of mutually independent standard Brownian motion, independent from X_0 and with initial distribution ν_0. The process X_t exists, it is finite in finite time intervals almost surely, it is distributionally unique and its distribution does not depend on the choice of $\sigma(x,t)$.

For the proof we refer to Nelson (1990). Conditions under which ν_0, $b(x,t)$ and $a(x,t)$ ensure finiteness of the process in finite time intervals and uniqueness of the limiting diffusion are extensively discussed in Stroock and Varadhan (1979), Ethier and Kurtz (1986b), and Nelson (1990). To ensure weak existence, uniqueness and non-explosion of the diffusion process X_t on compact sets we rely on 'Condition A' of Nelson (1990), i.e.,

Condition 1 (weak existence and uniqueness)**.** *Let $a(x,t)$ and $b(x,t)$ be continuous in both x and t with two partial derivatives with respect to x.*

Following Theorem 10.2.2 of Stroock and Varadhan (1979), we impose the following conditions of non-explosiveness of X_t.

Condition 2 (non-explosiveness)**.** *For each $T > 0$, there is a $C_T < \infty$ such that*

$$\sup_{0 \le t \le T} \|a(x,t)\| \le C_T(1 + |x|^2), \quad x \in \mathbb{R}^N$$

and

$$\sup_{0 \le t \le T} \langle x, b(x,t) \rangle \le C_T(1 + |x|^2), \quad x \in \mathbb{R}^N.$$

4.3 Main theoretical results

Let ${}_hY_{kh}$ be an N-dimensional vector of logarithmic prices indexed by kh, $k \in \mathbb{N}$, $h > 0$. The pre-subscript h represents the sampling frequency or said differently, the time interval between two consecutive observations, i.e. for given h, prices are observed at times $h, 2h, 3h, ...$ We let the parameters depend on the sampling frequency. Furthermore, the variance of the innovations is made proportional to h. In this chapter we focus on the covariance stationary case, hence usual suitable positivity and stationarity constraints on the parameters of the variances and correlation driving process apply, see Bollerslev (1986), Engle and Sheppard (2001), Engle (2002b), Aielli (2006) and Aielli (2009).

In the remainder of the chapter we use the following operators: vec() stacks the columns of a matrix into a vector, vech() stacks the lower triangular portion of a square matrix into a vector, vechl() stacks the strictly lower triangular portion of a square matrix into a vector (i.e., excluding the diagonal elements), diag() stacks the diagonal of a square matrix into a vector. Furthermore, 1_N is a $(N \times 1)$ vector of ones and I_N is the $(N \times N)$ identity matrix. We also make use of the following special matrices: D_N denotes the $(N^2 \times N(N+1)/2)$ duplication matrix, which for any symmetric matrix A transforms vech(A) into vec(A), D_N^+ its generalized inverse, see e.g. Lütkepohl (1996) for details, I^* is defined such that diag$(A) = I^*$vech(A) with $I^* = I^{+\prime} D_N$ and $I^+ = (1_N \otimes I_N) \odot [1'_N \otimes \text{vec}(I_N)]$ is the $(k^2 \times k)$ matrix I^+ transforms vec(A) into diag(A). Finally, I^- is defined such that vechl(A)vechl$(A)' = I^-(\text{vech}(A)\text{vech}(A)')I^{-\prime} = I^- D_N^+ (A \otimes A) D_N^{+\prime} I^{-\prime}$.[2]

4.3.1 The cDCC-GARCH process

The discrete time cDCC-GARCH process of Aielli (2006) for the log return of a N-dimensional portfolio of assets Y_t is specified as follows:

$$Y_t = Y_{t-1} + S_t \, \eta_t, \qquad \eta_t \sim \text{N}(0, R_t) \tag{4.4}$$

$$V_{t+1} = c + A \, S_t^2 (\eta_t \odot \eta_t) + B \, V_t, \tag{4.5}$$

$$Q^*_{t+1} = \bar{Q} + \vartheta (P_t \, \eta_t \, \eta_t{}' \, P_t) + \gamma \, Q^*_t, \tag{4.6}$$

$$P_t = (Q^*_t \odot I_N)^{1/2}, \tag{4.7}$$

$$R_t = P_t^{-1} \, Q^*_t \, P_t^{-1}, \tag{4.8}$$

[2]Examples: $I^+_{N=3} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}'$ and $I^-_{N=3} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$.

where $\{R_t^{-1/2}\eta_t, t = 1, 2, ...\}$ is a sequence of N-dimensional vectors of i.i.d. gaussian innovations. Let us now consider the properties of the system of stochastic difference equations (4.4)-(4.8) as the time is partitioned more and more finely. We let the parameters of the system to depend on sampling frequency h as well as the covariance matrix of the vector of idiosyncratic innovations. We start partitioning time in (4.4)-(4.8), for $h > 0$ and $k \in \mathbb{N}$, according to the following scheme

$$
\begin{aligned}
{}_hY_{kh} &= {}_hY_{(k-1)h} + {}_hS_{kh}\,{}_h\eta_{kh}, && (4.9)\\
{}_hV_{(k+1)h} &= c_h + A_h h^{-1}\,{}_hS_{kh}^2({}_h\eta_{kh} \odot {}_h\eta_{kh}) + B_h\,{}_hV_{kh}, && (4.10)\\
{}_hQ_{(k+1)h} &= \bar{Q}_h + \vartheta_h h^{-1}\mathrm{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta_{kh}'\,{}_hP_{kh}) + \gamma_h\,{}_hQ_{kh}, && (4.11)\\
{}_hP_{kh} &= ({}_hQ_{kh}^* \odot I_N)^{1/2}, && (4.12)\\
{}_hR_{kh} &= {}_hP_{kh}^{-1}\,{}_hQ_{kh}^*\,{}_hP_{kh}^{-1}, && (4.13)
\end{aligned}
$$

and

$$
P_h[({}_hY_0, {}_hV_0, {}_hQ_0) \in \Gamma] = \nu_h(\Gamma) \text{ for any } \Gamma \in B\left(\mathbb{R}^{N(N+5)/2}\right) \tag{4.14}
$$

where c_h is an $(N \times 1)$ parameter vector, A_h, B_h are $(N \times N)$ diagonal parameter matrices, ϑ_h and γ_h are scalars and ${}_h\eta_{kh}$ is an $(N \times 1)$ vector of devolatilized but correlated innovations, such that ${}_hR_{kh}^{-1/2}\,{}_h\eta_{kh} \sim$ i.i.d. N$(0, h\ I_N)$. Further, ${}_hS_{kh}$ is an $(N \times N)$ diagonal matrix of rescaled conditional standard deviations with $\mathrm{diag}({}_hS_{kh}^2) = {}_hV_{kh}$ an $(N \times 1)$ vector of rescaled conditional variances. The correlation driving process ${}_hQ_{kh}^*$ represents the conditional expectation of the outer product of the corrected devolatilized innovations, see Aielli (2009) for details, and it is therefore symmetric and positive semi-definite. For analytical tractability we express the model in terms of the vector of the unique elements of ${}_hQ_{kh}^*$, i.e. ${}_hQ_{(k+1)h} = \mathrm{vech}({}_hQ_{kh}^*)$. Finally, (4.14) defines the distribution of the starting point. The system (4.9)-(4.14) describe a discrete time Markov process.

The advantage of the cDCC over the standard DCC model of Engle (2002b) stands in the fact that the recursion in ${}_hQ_{kh}$ preserves the martingale difference property, i.e., $h^{-1}\mathrm{E}[({}_hP_{kh}\ {}_h\eta_{kh}\ {}_h\eta_{kh}'\ {}_hP_{kh})|M_{kh}] = {}_hP_{kh}\ {}_hR_{kh}\ {}_hP_{kh} = {}_hQ_{kh}^*$. Hence, for a given h, the process $(h^{-1}\ {}_hP_{kh}\ {}_h\eta_{kh},\ {}_hQ_{kh}^*)$ is a multivariate semi-strong GARCH process in the sense of Hafner (2008).

Without loss of generality, we reparameterize the drift in the recursion ${}_hQ_{kh}$ as a combination of a frequency invariant component and frequency dependent parameters. The drift $\bar{Q}_h$ can be expressed as $\bar{Q}_h = (1 - \vartheta_h - \gamma_h)\bar{Q}$.[3] This transformation will turn par-

[3]The same transformation can be carried out also for the intercept of the ${}_hV_{(k+1)h}$ process, i.e., $c_h =$

ticularly useful when deriving the diffusion limit of the Constant Conditional Correlation (CCC) model of Bollerslev (1990). In fact, under the parameter restriction $\vartheta_h = \gamma_h = 0$, ${}_hQ_{kh} = \bar{Q}_h = \bar{Q}$ and therefore ${}_hR_{kh} = \bar{Q}\ \forall h$.[4]

Before deriving the diffusion limit of cDCC-GARCH process we determine the convergence rates of the discrete time parameters for the moment conditions to converge as the sampling frequency increases, as required by Assumption 4.1.

Proposition 4.1. *Under the following convergence rates for the parameters of the discrete time cDCC-GARCH process (4.9)-(4.11),*

$$c_h = h\,c + o(h) \tag{4.15}$$

$$(A_h + B_h - I_N) = -h\,\Lambda + o(h) \tag{4.16}$$

$$A_h = \sqrt{h}\,A + o(\sqrt{h}) \tag{4.17}$$

$$(1 - \vartheta_h - \gamma_h) = h\,\phi + o(h) \tag{4.18}$$

$$\vartheta_h = \sqrt{h}\,\vartheta + o(\sqrt{h}), \tag{4.19}$$

for some $(N \times 1)$ vector c, $(N \times N)$ diagonal matrices A and Λ and scalars ϕ and ϑ with positive and finite elements, Assumption 4.1 holds.

The convergence rates in Proposition 4.1 ensure that the first and the second conditional moments per unit of time converge, as $h \to 0$, to well-behaved limits and that the first difference of the process $[{}_hY_{kh},\ {}_hV_{kh},\ {}_hQ_{kh}]$ has an absolute moment higher than two converging to zero at an appropriate rate as required by Assumption 4.1.

Note that $c > 0$ (elementwise) ensures positivity of the variance process, $A > 0$ and $\vartheta > 0$ ensure that the rescaled second conditional moment does not vanish as $h \to 0$,[5] while $\Lambda > 0$ and $\phi > 0$ ensure covariance stationarity of the return process.[6]

Proposition 4.2. *Under the convergence rates given in Proposition 4.1, the first and second*

$(I_N - A_h - B_h)\bar{c}$. The vector $\bar{c}$ is frequency invariant and holds the (rescaled) unconditional variances of the return process $({}_hY_{(k+1)h} - {}_hY_{kh})$, i.e., $\bar{c} = \mathrm{E}[({}_hY_{(k+1)h} - {}_hY_{kh}) \odot ({}_hY_{(k+1)h} - {}_hY_{kh})]/h = \mathrm{E}[{}_hV_{(k+1)h}]$, $\forall h$.

[4]Note that even though in general $\bar{Q}$ does not need to be a correlation matrix, i.e. $\mathrm{diag}(\bar{Q}) = 1_N$, under the CCC parameter restrictions the diagonal elements of $\bar{Q}$ are not identifiable together with the intercept of the ${}_hV_{(k+1)h}$ process. Fixing $\mathrm{diag}(\bar{Q}) = 1_N$ ensures that: (i) $\mathrm{E}[{}_hV_{(k+1)h}] = (I_N - A_h - B_h)^{-1}c_h$ is the rescaled unconditional variance of the return process $(\mathrm{E}[({}_hY_{(k+1)h} - {}_hY_{kh}) \odot ({}_hY_{(k+1)h} - {}_hY_{kh})]/h)$, (ii) $\bar{Q}$ can be directly interpreted as the (un)conditional correlation of $({}_hY_{(k+1)h} - {}_hY_{kh})$.

[5]These conditions imply that the diffusion limit of the cDCC-GARCH process converges to a continuous time stochastic volatility process.

[6]In the univariate setting, two special cases, $\Lambda = 0$ (integrated variance) and $\Lambda < 0$ (strictly stationary but not covariance stationary GARCH process) are also discussed in Nelson (1990). In this chapter we restrict the analysis to the covariance stationary case.

moment conditions of the cDCC-GARCH process (4.9)-(4.11) converge, as $h \to 0$*, to*

$$b(X_t, t) = \begin{bmatrix} 0 \\ c - \Lambda V_t \\ \phi(\bar{Q} - Q_t) \end{bmatrix}, \tag{4.20}$$

and

$$vech(a(X_t, t)) = \begin{bmatrix} S_t R_t S_t \\ 0 \\ 0 \\ 0 \\ 2AS_t^2(R_t \odot R_t)S_t^2 A \\ \vartheta \left[I^*(D_N^+ K_t D_N^{+\prime}) \left(D_N^+ (P_t \otimes P_t) D_N^{+\prime} \right) - 1_N Q_t' \right]' S_t^2 A \\ \vartheta^2 [(D_N^+ (P_t \otimes P_t) D_N^{+\prime})(D_N^+ K_t D_N^{+\prime})(D_N^+ (P_t \otimes P_t) D_N^{+\prime}) - Q_t Q_t'] \end{bmatrix}, \tag{4.21}$$

where $X_t = [Y_t, V_t, Q_t]'$*. The matrix* $a(X_t, t)$ *is singular and its rank is equal to* $N(N+3)/2 < \dim(a(X_t, t)) = N(N+5)/2$*. The conditional correlation,* R_t*, is computed at each point in time as in (4.13).*

In the cDCC-GARCH model the matrix $a(X_t, t)$ is singular by construction. The singularity is due to the particular structure of the model in which the noise propagation system of the variance processes and the one of the diagonal elements of the correlation driving processes are pairwise perfectly correlated. This is because, although (possibly) different in terms of level and dynamics, (4.10) and (4.11) are driven by the same source of noise. In this sense the discrete time cDCC-GARCH model is somewhat redundant, as will be discussed at the end of this section[7].

It is also worth noting that, similarly to the univariate case, $a(X_t, t)$ is block diagonal given the gaussianity assumption for ${}_h\eta_{kh}$.

Endowed with the rates of convergence of the parameters and the limits of the moments conditions we can now state in Theorem 4.1 the diffusion limit of the cDCC-GARCH process.

Theorem 4.1 (**Diffusion limit of the cDCC-GARCH model**)**.** *Under (4.15) to (4.19), the discrete time cDCC-GARCH process (4.9)-(4.11) weakly converges to the diffusion process* $X_t = [Y_t,\ V_t,\ Q_t]'$ *which is the solution to the system of stochastic differential equations*

$$dX_t = b(X_t, t)\,dt + \sigma(X_t, t)\,dW_t. \tag{4.22}$$

[7]For an example, consider the simple case where $(I_N - A_h - B_h)^{-1} c_h = \text{diag}(\bar{Q})$, $A_h = \vartheta_h I_N$, $B_h = \gamma_h I_N$, then the model reduces a standard scalar VEC with N redundant equations.

The drift, $b(X_t, t)$, is given in (4.20) while the scale, $\sigma(X_t, t)$, is a continuous mapping such that, for all $X_t \in \mathbb{R}^{N(N+5)/2}$ and $t \geq 0$, $a(X_t, t) = \sigma(X_t, t)\sigma(X_t, t)'$ where $a(X_t, t)$ is given in (4.21). The rank of $\sigma(X_t, t)$ is $N(N+3)/2$ thus leading to a degenerate diffusion limit driven by a vector, dW_t, of $N(N+3)/2$ mutually independent standard Brownian motions, independent from the initial value $X_0 = [Y_0,\ V_0,\ Q_0]'$.

Thus, the cDCC-GARCH process converges to a degenerate diffusion, in the sense that it is driven by as many Brownian motions as the number of stochastic differential equations in the system but whose covariance matrix is singular.

It is interesting to elaborate on the implications of singularity of the diffusion matrix $a(X_t, t)$. To this end, let us rearrange the order of the elements of the diffusion process X_t as $[Y_t, V_t, Q_t^{(d)}, Q_t^{(l)}]'$, where $Q_t^{(d)} = [Q_{ii,t}]_{i=1,\dots,N} = \text{diag}(Q_t^*)$ and $Q_t^{(l)} = [Q_{ij,t}]_{i<j=2,\dots,N} = \text{vechl}(Q_t^*)$. The two partial diffusion processes $[Y_t, V_t, Q_t^{(l)}]'$ and $[Y_t, Q_t^{(d)}, Q_t^{(l)}]'$ share the same correlation structure, while $\text{Corr}(\text{d}V_{t,i}, \text{d}Q_{t,ii}) = 1\ \forall i$ implies that the two partial diffusions are driven by the same vector of Brownian innovations. Thus, the relevant part in terms of noise propagation system of the diffusion limit of the cDCC-GARCH process consists of a system of $N(N+3)/2$ stochastic differential equations, either $[Y_t, V_t, Q_t^{(l)}]$ or $[Y_t, Q_t^{(d)}, Q_t^{(l)}]$, while the remaining N diffusion processes, $Q_t^{(d)}$ or V_t respectively, are characterized by a specific deterministic part (drift) but a common, though appropriately rescaled, stochastic component. To illustrate this point, let us consider the following partition of the diffusion matrix in (4.21), whose elements have been opportunely reordered (the time index has been dropped to simplify the notation)

$$a(X_t, t) = \begin{bmatrix} \Sigma_{YY} & 0 & 0 & 0 \\ 0 & \Sigma_{VV} & \Sigma_{VQ^{(d)}} & \Sigma_{VQ^{(l)}} \\ 0 & \Sigma'_{VQ^{(d)}} & \Sigma_{Q^{(d)}Q^{(d)}} & \Sigma_{Q^{(d)}Q^{(l)}} \\ 0 & \Sigma'_{VQ^{(l)}} & \Sigma'_{Q^{(d)}Q^{(l)}} & \Sigma_{Q^{(l)}Q^{(l)}} \end{bmatrix} \tag{4.23}$$

where

$$\begin{aligned} \Sigma_{YY} &= S_t R_t S_t \\ \Sigma_{VV} &= 2AS_t^2(R_t \odot R_t)S_t^2 A \\ \Sigma_{VQ^{(d)}} &= 2\vartheta AS_t^2(R_t \odot R_r)(P_t P_t) \\ &= \vartheta \Sigma_{VV}(S_t^2 A)^{-1}(P_t P_t) \\ \Sigma_{Q^{(d)}Q^{(d)}} &= 2\vartheta^2(P_t P_t)(R_t \odot R_r)(P_t P_t) \\ &= \vartheta^2(P_t P_t)(AS_t^2)^{-1}\Sigma_{VV}(S_t^2 A)^{-1}(P_t P_t) \end{aligned}$$

$$
\begin{aligned}
\Sigma_{VQ^{(l)}} &= \vartheta A S^2 \left[I^* (D_N^+ K_t D_N^{+\prime}) \left(D_N^+ (P_t \otimes P_t) D_N^{+\prime} \right) I^{-\prime} - 1_N Q_t' I^{-\prime} \right] \\
\Sigma_{Q^{(d)}Q^{(l)}} &= \vartheta^2 (P_t P_t) [I^* (D_N^+ K_t D_N^{+\prime}) (D_N^+ (P_t \otimes P_t) D_N^{+\prime}) I^{-\prime} - 1_N Q_t' I^{-\prime}] \\
&= \vartheta (P_t P_t) (A S_t^2)^{-1} \Sigma_{VQ^{(l)}} \\
\Sigma_{Q^{(l)}Q^{(l)}} &= \vartheta^2 I^- [(D_N^+ (P_t \otimes P_t) D_N^{+\prime}) (D_N^+ K_t D_N^{+\prime}) (D_N^+ (P_t \otimes P_t) D_N^{+\prime}) - Q_t Q_t'] I^{-\prime}.
\end{aligned}
$$

Let us also define $C_t = \vartheta (P_t P_t)(A S_t^2)^{-1}$. We can rewrite (4.23) as

$$
\begin{bmatrix}
\Sigma_{YY} & 0 & 0 & 0 \\
0 & \Sigma_{VV} & \Sigma_{VV} C_t' & \Sigma_{VQ^{(l)}} \\
0 & C_t \Sigma_{VV} & C_t \Sigma_{VV} C_t' & C_t \Sigma_{VQ^{(l)}} \\
0 & \Sigma_{VQ^{(l)}}' & \Sigma_{VQ^{(l)}}' C_t' & \Sigma_{Q^{(l)}Q^{(l)}}
\end{bmatrix}.
$$

Therefore,

$$
a([Y_t, V_t, Q_t^{(l)}], t) = \begin{bmatrix}
\Sigma_{YY} & 0 & 0 \\
0 & \Sigma_{VV} & \Sigma_{VQ^{(l)}} \\
0 & \Sigma_{VQ^{(l)}}' & \Sigma_{Q^{(l)}Q^{(l)}}
\end{bmatrix} \tag{4.24}
$$

$$
a([Y_t, Q_t^d, Q_t^{(l)}], t) = \begin{bmatrix}
\Sigma_{YY} & 0 & 0 \\
0 & C_t \Sigma_{VV} C_t' & C_t \Sigma_{VQ^{(l)}} \\
0 & \Sigma_{VQ^{(l)}}' C_t' & \Sigma_{Q^{(l)}Q^{(l)}}
\end{bmatrix}. \tag{4.25}
$$

The decomposition in (4.24) and (4.25) shows that the two partial processes $[V_t, Q_t^{(l)}]'$ and $[Q_t^{(d)}, Q_t^{(l)}]'$, both uncorrelated with Y_t, share the same correlation structure. Furthermore, from Proposition 4.2, it immediately follows that $[V_t, Q_t^{(l)}]'$ and $[Q_t^{(d)}, Q_t^{(l)}]'$ are elementwise perfectly correlated[8], which implies that the two partial diffusion are driven by the same vector of Brownian motions. However, although either partial diffusion process $[Y_t, V_t, Q_t^{(l)}]$ or $[Y_t, Q_t^{(d)}, Q_t^{(l)}]'$ is sufficient alone to fully characterize the noise propagation system of the cDCC diffusion limit, they are both necessary to characterize the distributions of Y_t and V_t which depend on both V_t and $Q_t = [Q_t^{(d)}, Q_t^{(l)}]'$ through the correlation process R_t.[9] A practical implementation of the diffusion (4.20)-(4.21) and (4.22) is discussed in Section 2.4.

[8] More generally, $\mathrm{Corr}(\mathrm{d}V_{t,i}, \mathrm{d}V_{t,j}) = \mathrm{Corr}(\mathrm{d}Q_{t,i}^{(d)}, \mathrm{d}Q_{t,j}^{(d)}) = \mathrm{Corr}(\mathrm{d}V_{t,i}, \mathrm{d}Q_{t,j}^{(d)}) = (R_t \odot R_t)_{ij} \ \forall i, j = 1, ..., N$.

[9] Note that the partial system $[Y_t, Q_t^{(d)}, Q_t^{(l)}]'$ is however sufficient to characterize the distribution of the correlation driving process Q_t and hence of the correlation R_t.

4.3.2 A special case: the CCC-GARCH process

As a particular case, we now consider the Constant Conditional Correlation (CCC) model of Bollerslev (1990). The cDCC process nests the CCC process under the following parameter restrictions

$$\vartheta_h = \gamma_h = 0 \quad \forall h.$$

Thus, the CCC-GARCH process can be written as

$$ {}_hY_{kh} = {}_hY_{(k-1)h} + {}_hS_{kh}\,{}_h\eta_{kh}, \tag{4.26}$$

$$ {}_hV_{(k+1)h} = c_h + A_h h^{-1}\,{}_hS^2_{kh}({}_h\eta_{kh} \odot {}_h\eta_{kh}) + B_h\,{}_hV_{kh}, \tag{4.27}$$

where c_h, A_h, B_h are defined as before and ${}_h\eta_{kh}$ is a $(N \times 1)$ vector of devolatilized but correlated innovations, such that ${}_h\eta_{kh} \sim \mathrm{N}(0, h\ R)$, where R represents the (frequency invariant) constant conditional correlation matrix, i.e., $\mathrm{Var}({}_h\eta_{kh}) = h\,R \Rightarrow \mathrm{Corr}({}_h\eta_{kh}) = R\ \forall h$. This model, though rather restrictive in practice, is particularly interesting because, unlike the cDCC-GARCH process, it allows for a non-degenerate diffusion and, in the bivariate specification, a closed form solution for the diffusion limit. The rates of convergence for the parameters and the CCC-GARCH process are stated in Proposition 4.3 and Theorem 4.2.

Proposition 4.3. *Under the following convergence rates for the parameters of the discrete time CCC-GARCH process (4.26)-(4.27)*

$$c_h = h\,c + o(h) \tag{4.28}$$

$$(A_h + B_h - I_N) = -h\,\Lambda + o(h) \tag{4.29}$$

$$A_h = \sqrt{h}\,A + o(\sqrt{h}), \tag{4.30}$$

for some $(N \times 1)$ vector c, $(N \times N)$ diagonal matrices A and Λ with positive and finite elements, Assumption 4.1 holds.

The same considerations on the parameters as in Proposition 4.1 hold by symmetry with the cDCC-GARCH process.

Theorem 4.2 (**Diffusion limit of the CCC-GARCH model**). *Under the convergence conditions in Proposition 4.3, the CCC-GARCH process (4.26)-(4.27) weakly converges to the non-degenerate diffusion process $X_t = [Y_t\ V_t]'$ solution to a system of stochastic differential equations of the form (4.22), with drift*

$$b(X_t, t) = \begin{bmatrix} 0 \\ c - \Lambda V_t \end{bmatrix} \tag{4.31}$$

and diffusion matrix

$$a(X_t, t) = \begin{bmatrix} S_t\, R\, S_t & 0 \\ 0 & 2AS_t^2(R \odot R)S_t^2 A \end{bmatrix} \tag{4.32}$$

and driven by a vector, dW_t, of $2N$ mutually independent Brownian motions, independent of the initial value $X_0 = [Y_0\ V_0]'$.

The diffusion limit of the CCC model is clearly non-degenerate because it is driven by as many Brownian motions as the number of variables in the system and whose covariance matrix is non-singular.

It is clear that the diffusion limit of the cDCC-GARCH process (as well as of the CCC-GARCH process) is a continuous time stochastic volatility model (i.e., stochastic variances and correlations). We discuss next the case when rates of convergence other than the ones introduced in Proposition 4.1, but still satisfying Assumption 4.1, are used.

4.3.3 cDCC-GARCH diffusion with deterministic variance/correlation

In this section we reconsider the continuous time approximation of the cDCC-GARCH process (4.9)-(4.11). The convergence rate $h^{1/2}$, suggested in Proposition 4.1, represents the slowest rate of convergence for the parameters A_h and ϑ_h satisfying Assumption 4.1. More generally, the rate $h^{1/2}$ represents the only rate ensuring that the second conditional moments $\text{Var}({}_hV_{(k+1)h} - {}_hV_{kh}|M_{kh})$, $\text{Var}({}_hQ_{(k+1)h} - {}_hQ_{kh}|M_{kh})$ and $\text{Cov}[({}_hV_{(k+1)h} - {}_hV_{kh}), ({}_hQ_{(k+1)h} - {}_hQ_{kh})|M_{kh}]$ scaled by h^{-1}, do not vanish as $h \to 0$. As shown in Theorem 4.1, the resulting diffusion limit is characterized by stochastic variances of the marginal processes and stochastic correlation driving process.

However, there are other admissible convergence rates for A_h and ϑ_h which also satisfy Assumption 4.1. Thus, depending on the continuous time approximation we consider, we can recover different types of diffusion for the process (4.9)-(4.11).[10] This alternative set of results is shown in Proposition 4.4 and Theorem 4.3.

Proposition 4.4. *Under the following convergence rates for the parameters A_h and ϑ_h*

$$\lim_{h\to 0} h^{-(\frac{1}{2}+\delta)} A_h = \tilde{A} < \infty \tag{4.33}$$

and/or

$$\lim_{h\to 0} h^{-(\frac{1}{2}+\delta)} \vartheta_h = \tilde{\vartheta} < \infty, \tag{4.34}$$

[10]The following arguments can be easily extended to the CCC-GARCH process, although this case is not explicitly treated here.

for some $(N \times N)$ *diagonal matrix* $\tilde{A} > 0$ *(elementwise),* $\tilde{\vartheta} > 0$ *and some* $\delta > 0$*, Assumption 4.1 holds.*

Note that under (4.33) and (4.34), A_h and ϑ_h are of order $h^{1/2+\delta}$, $\delta > 0$.

Proposition 4.4 suggests alternative sets of conditions regarding the speed of convergence of the discrete time parameters under which Assumption 4.1 holds. The implications of Proposition 4.4 are straightforward. In fact, it is immediate to see that if either A_h or ϑ_h or both are of order $h^{1/2+\delta}$, $\delta > 0$, then the terms depending on ${}_h\eta_{kh}$ on the right hand side of (4.10) and/or (4.11) are also of order $h^{1/2+\delta}$. Consequently, the conditional second moments rescaled by h^{-1} are of order $h^{2\delta}$ and therefore converge to zero as $h \to 0$. The resulting diffusion limits are clearly degenerate and are characterized by time varying but deterministic variances of the marginal processes and/or a deterministic correlation driving process. The conditions in (4.33) and (4.34) can also be seen as a special case of (4.17) and (4.19), obtained by setting $A = 0I_N$ and $\vartheta = 0$, respectively, when the rate of convergence of A_h and ϑ_h is $\sqrt{h}$, i.e., $A_h = o(\sqrt{h})$ and $\vartheta_h = o(\sqrt{h})$.

Theorem 4.3 (**cDCC-GARCH diffusion with deterministic variance/correlation**)**.** *Under (4.15), (4.16), (4.18) and (4.33)-(4.34), the discrete time cDCC-GARCH process (4.9)-(4.11) admits a degenerate diffusion limit with time varying but deterministic variances* (V_t) *and stochastic correlations, stochastic variances and time varying but deterministic correlation driving process* (Q_t)*, or both deterministic variances and correlations. The diffusion process* $X_t = [Y_t\ V_t\ Q_t]'$ *is solution to a system of stochastic differential equations of the form (4.22), with drift given by (4.20) and diffusion matrix given respectively by*

i) (deterministic variances but stochastic correlation) under (4.15), (4.16), (4.18), (4.19) and (4.33)

$$a(X_t,t) = \begin{bmatrix} S_tR_tS_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \vartheta^2[(D_N^+(P_t \otimes P_t)D_N^{+\prime})(D_N^+K_tD_N^{+\prime}) \\ & & (D_N^+(P_t \otimes P_t)D_N^{+\prime}) - Q_tQ_t'] \end{bmatrix}. \tag{4.35}$$

The diffusion process defined by (4.22), (4.20) and (4.35) is driven by $N(N+3)/2$ *independent standard Brownian motions;*

ii) (stochastic variance but deterministic correlation) under (4.15), (4.16), (4.17), (4.18) and (4.34)

$$a(X_t,t) = \begin{bmatrix} S_tR_tS_t & 0 & 0 \\ 0 & 2AS_t^2(R_t \odot R_r)S_t^2A & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.36}$$

The diffusion process defined by (4.22), (4.20) and (4.36) is driven by $2N$ *independent standard Brownian motions;*

iii) (deterministic variances and correlation) under (4.15), (4.16), (4.18) and both (4.33) and (4.34)

$$a(X_t, t) = \begin{bmatrix} S_t R_t S_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.37}$$

The diffusion process defined by (4.22), (4.20) and (4.37) is driven by N *independent standard Brownian motions.*

We now discuss what type of process can be obtained as Euler approximation of the different diffusion processes recovered in Theorem 4.3. Following the arguments of Corradi (2000), it is easy to show, using conventional algebra of stochastic calculus, see, e.g., Steele (2001) p.123, that Euler approximations of the three diffusions defined in Theorem 4.3 are: *i*) a process with stochastic correlation and GARCH variances, *ii*) a process with stochastic variance and cDCC correlation and *iii*) a cDCC-GARCH process as in (4.9)-(4.11), respectively.

Furthermore, when either the variances, the correlation driving process or both are M_0 measurable, i.e., deterministic given V_0, Q_0 and (V_0,Q_0) respectively, it is possible to recover some special cases.

In the first case, under Theorem 4.3(*i*) and constant variance (since $\Lambda > 0$, ${}_hV_{kh} \to V = \Lambda^{-1} c \ \forall h$ as $kh \to \infty$), the process ${}_hP_{kh}\, S^{-1}\, ({}_hY_{(k+1)h} - {}_hY_{kh})$ is a multivariate stochastic volatility process with stochastic covariance driving process ${}_hQ^*_{kh}$.

In the second case, under Theorem 4.3(*ii*) and constant correlation driving process, following the same argument, the process $({}_hY_{(k+1)h} - {}_hY_{kh})$ is a constant correlation type process with stochastic volatility for the marginal processes.

The third case is perhaps the more interesting one. In fact, under Theorem 4.3(*iii*), constant variances and cDCC correlation driving process, the process ${}_hP_{kh}\, S^{-1}\, ({}_hY_{(k+1)h} - {}_hY_{kh})$ is a scalar VEC process (Bollerslev, Engle, and Wooldridge, 1988b), under Theorem 4.3(*iii*), GARCH variances and constant correlation driving process, the process $({}_hY_{(k+1)h} - {}_hY_{kh})$ is a CCC-GARCH process as in (4.26)-(4.27), while under Theorem 4.3(*iii*) and both constant variances and correlation driving process, $({}_hY_{(k+1)h} - {}_hY_{kh}) \sim \mathrm{N}(0, h\, S\, R\, S)$.

4.4 Simulation

An interesting aspect of the convergence of the system of stochastic difference equations (4.9)-(4.11) to the system of stochastic differential equations of the type defined in Theorem 4.1 is that it allows to exploit the relationships established in Proposition 4.1 to infer the parameters of the underlying continuous time process through the parameter estimates of the corresponding discrete time model. In this section we validate by Monte Carlo simulation the convergence theory detailed above and investigate the accuracy of the continuous time parameter estimates inferred by the discrete parameter estimates. We estimate the parameters of a sequence, indexed by h (level of aggregation of the data), of discrete time GARCH models with i.i.d. innovations. Then, for each h, we use the relationships given in Proposition 4.1 to obtain the diffusion parameters and we investigate the behavior of the latter as $h \to 0$. Since these relationships are not exact but depend on the sampling frequency, we expect that the better the approximation (higher sampling frequency) the more accurate the inference. We focus on the bivariate cDCC-GARCH process defined by (4.9)-(4.11), that we have shown to weakly converge, when $h \to 0$, to the continuous time process (4.22), with drift defined in (4.20) and diffusion matrix defined in (4.21). For $N = 2$ and using the results of Section 4.3.1, the cDCC-GARCH diffusion can be written as

$$\begin{bmatrix} \mathrm{d}Y_{1t} \\ \mathrm{d}Y_{2t} \end{bmatrix} = \begin{bmatrix} \sqrt{V_{1t}} & 0 \\ 0 & \sqrt{V_{2t}} \end{bmatrix} \Upsilon^{(1)}(\rho_t)^{\frac{1}{2}} \mathrm{d}W_t^{(1)} \tag{4.38}$$

$$\begin{bmatrix} \mathrm{d}V_{1t} \\ \mathrm{d}V_{2t} \\ \mathrm{d}Q_{12t} \end{bmatrix} = \begin{bmatrix} \Lambda_{11}(c_1 - V_{1t}) \\ \Lambda_{22}(c_2 - V_{2t}) \\ \phi(\bar{Q}_{22} - Q_{22t}) \end{bmatrix} \mathrm{d}t + \sqrt{2} \begin{bmatrix} \alpha_1 V_{1t} & 0 & 0 \\ 0 & \alpha_2 V_{2t} & 0 \\ 0 & 0 & \vartheta Q_{12t} \sqrt{\frac{1+\rho_t^2}{2\rho_t^2}} \end{bmatrix} \Upsilon^{(2)}(\rho_t)^{\frac{1}{2}} \mathrm{d}W_t^{(2)} \tag{4.39}$$

$$\begin{bmatrix} \mathrm{d}Q_{11t} \\ \mathrm{d}Q_{22t} \\ \mathrm{d}Q_{12t} \end{bmatrix} = \begin{bmatrix} \phi(\bar{Q}_{11} - Q_{11t}) \\ \phi(\bar{Q}_{22} - Q_{22t}) \\ \phi(\bar{Q}_{12} - Q_{12t}) \end{bmatrix} \mathrm{d}t + \sqrt{2}\vartheta \begin{bmatrix} Q_{11t} & 0 & 0 \\ 0 & Q_{22t} & 0 \\ 0 & 0 & Q_{12t} \sqrt{\frac{1+\rho_t^2}{2\rho_t^2}} \end{bmatrix} \Upsilon^{(2)}(\rho_t)^{\frac{1}{2}} \mathrm{d}W_t^{(2)}, \tag{4.40}$$

where

$$\Upsilon^{(1)}(\rho_t) = \begin{bmatrix} 1 & \rho_t \\ \rho_t & 1 \end{bmatrix}, \qquad \Upsilon^{(2)}(\rho_t) = \begin{bmatrix} 1 & \rho_t^2 & \sqrt{\frac{2\rho_t^2}{1+\rho_t^2}} \\ \rho_t^2 & 1 & \sqrt{\frac{2\rho_t^2}{1+\rho_t^2}} \\ \sqrt{\frac{2\rho_t^2}{1+\rho_t^2}} & \sqrt{\frac{2\rho_t^2}{1+\rho_t^2}} & 1 \end{bmatrix}$$

and

$$\rho_t = \frac{Q_{12t}}{\sqrt{Q_{11t}Q_{22t}}}.$$

Note that the drift in $[\mathrm{d}V_{1t}, \mathrm{d}V_{2t}]$ has been reparameterized such that c represents the rescaled unconditional variance of the return process. The two partial systems $[\mathrm{d}V_{1t}, \mathrm{d}V_{2t}, \mathrm{d}Q_{12t}]$ and $[\mathrm{d}Q_{11t}, \mathrm{d}Q_{22t}, \mathrm{d}Q_{12t}]$ share the same correlation structure, $\Upsilon(\rho_t)$, and stochastic component, $\mathrm{d}W_t^{(2)}$. Note also that Q_{12t} is distributionally (and also path by path) equivalent in both systems, see Section 4.3.1 for further details. We use an Euler discretization scheme of (4.38)-(4.40), and we simulate 500 sample paths using a discretization interval $\Delta t = 1/640$ which corresponds to one observation every 2 minutes 15 seconds (640 obs/day) and of length (k) equal to 2000 days. All programs have been written by the authors and are available upon request.

The data is generated using the following parameterization: $c = [1, 1.5]'$, $A_{11} = 0.07$, $A_{22} = 0.10$, $\Lambda_{11} = 0.13$, $\Lambda_{22} = 0.10$, $\vartheta = 0.08$ and $\phi = 0.04$. The unconditional expectation of the correlation driving process, $\bar{Q}$, has been parameterized such that it represent the unconditional correlation, i.e., $\mathrm{diag}(\bar{Q}) = 1_N$. Under this parameterization, the target for the unconditional correlation, $\bar{Q}_{12}$, is set to 0.5, while to reduce the number of parameters to be estimated, $\bar{Q}_{ii}$, $i = 1, 2$, are fixed to their true value. The square root of the correlation matrices of the diffusion, $\Upsilon^{(1)}(\rho_t)$ and $\Upsilon^{(2)}(\rho_t)$, are computed by spectral decomposition.

For each sample path we estimate (4.9)-(4.11) using returns sampled at daily ($h = 1$), 12-hour ($h = 1/2$), 6-hour ($h = 1/4$), 3-hour ($h = 1/8$), 90-minutes ($h = 1/16$), 45-minutes ($h = 1/32$) and 22.5-minutes ($h = 1/64$) frequency. The sequences of discrete time models are estimated by QAML. Although the QAML estimator is known to be biased in this setting, in similar though unrelated frameworks Fornari and Mele (2006), Broze, Scaillet, and Zakoian (1998), Hafner and Rombouts (2007) and Barone-Adesi, Rasmussen, and Ravanelli (2005) among others, have shown that the bias tend to disappear as the sample size and the precision of the discretization become sufficiently large.

Finally, our choice of k and h should reveal to be adequate to overcome finite sample problems and regard the estimates as asymptotic at all frequencies.

Inference on the continuous time parameters based on the discrete time estimates is obtained using the relationships provided in Proposition 4.1, i.e. up to $o(1)$ (except for (4.41) which is frequency invariant)

$$c = (I_N - A_h - B_h)^{-1} c_h, \tag{4.41}$$

$$A = A_h/\sqrt{h}, \tag{4.42}$$

$$\Lambda = (I_N - A_h - B_h)/h, \tag{4.43}$$

$$\vartheta = \vartheta_h/\sqrt{h}, \tag{4.44}$$

$$\phi = (1 - \vartheta_h - \gamma_h)/h. \tag{4.45}$$

Table 4.1 shows that the relationships between discrete and continuous time parameters given in (4.41)-(4.45) hold.

Table 4.1: Inference on the diffusion parameters of the cDCC process

Diffusion parameter estimates

Sampling freq.		c_1	A_{11}	Λ_{11}	c_2	A_{22}	Λ_{22}	Q_{12}	ϑ	ϕ
1-day	(h=1)	1.0005	0.0217	0.1763	1.5022	0.0395	0.1281	0.5020	0.0445	0.0488
12-hour	(h=1/2)	0.9991	0.0268	0.1527	1.5013	0.0489	0.1190	0.5008	0.0525	0.0459
6-hour	(h=1/4)	1.0007	0.0330	0.1478	1.4985	0.0586	0.1103	0.5019	0.0588	0.0442
3-hour	(h=1/8)	1.0009	0.0386	0.1417	1.4986	0.0681	0.1081	0.5024	0.0645	0.0439
90-min	(h=1/16)	1.0015	0.0462	0.1384	1.4960	0.0754	0.1068	0.5029	0.0686	0.0436
45-min	(h=1/32)	1.0014	0.0530	0.1356	1.4940	0.0814	0.1060	0.5029	0.0719	0.0434
22.5-min	(h=1/64)	1.0008	0.0575	0.1349	1.4973	0.0870	0.1058	0.5023	0.0742	0.0433
true	($h \to 0$)	1.0000	0.0700	0.1300	1.5000	0.1000	0.1000	0.5000	0.0800	0.0400

Notes: Inference on the continuous time parameters based on (4.42), (4.43), (4.44), (4.45). The true values of the parameters are reported in the last line (denoted $h \to 0$) for reference

Bias

Sampling freq.		c_1	A_{11}	Λ_{11}	c_2	A_{22}	Λ_{22}	Q_{12}	ϑ	ϕ
1-day	(h=1)	0.0005	-0.0483	0.0463	0.0022	-0.0605	0.0281	0.0020	-0.0354	0.0088
12-hour	(h=1/2)	-0.0009	-0.0432	0.0227	0.0013	-0.0511	0.0190	0.0008	-0.0275	0.0059
6-hour	(h=1/4)	0.0007	-0.0370	0.0178	-0.0015	-0.0414	0.0103	0.0019	-0.0212	0.0042
3-hour	(h=1/8)	0.0009	-0.0314	0.0117	-0.0014	-0.0319	0.0081	0.0024	-0.0155	0.0039
90-min	(h=1/16)	0.0015	-0.0238	0.0084	-0.0040	-0.0246	0.0068	0.0029	-0.0114	0.0036
45-min	(h=1/32)	0.0014	-0.0170	0.0056	-0.0060	-0.0186	0.0060	0.0029	-0.0081	0.0034
22.5-min	(h=1/64)	0.0008	-0.0125	0.0049	-0.0027	-0.0130	0.0058	0.0023	-0.0058	0.0033

RMSE

Sampling freq.		c_1	A_{11}	Λ_{11}	c_2	A_{22}	Λ_{22}	Q_{12}	ϑ	ϕ
1-day	(h=1)	0.0389	0.0502	0.0878	0.0733	0.0623	0.0889	0.0344	0.0373	0.0277
12-hour	(h=1/2)	0.0282	0.0446	0.0539	0.0611	0.0525	0.0589	0.0313	0.0288	0.0168
6-hour	(h=1/4)	0.0233	0.0381	0.0271	0.0556	0.0423	0.0315	0.0292	0.0222	0.0129
3-hour	(h=1/8)	0.0201	0.0321	0.0212	0.0514	0.0327	0.0247	0.0274	0.0164	0.0107
90-min	(h=1/16)	0.0187	0.0243	0.0187	0.0491	0.0252	0.0182	0.0269	0.0122	0.0091
45-min	(h=1/32)	0.0174	0.0173	0.0234	0.0486	0.0190	0.0156	0.0263	0.0089	0.0084
22.5-min	(h=1/64)	0.0170	0.0129	0.0220	0.0482	0.0136	0.0148	0.0245	0.0064	0.0082

As the sampling frequency increases, the inference on all parameters of the diffusion appear to be more and more accurate and the bias tends to disappear at an appropriate rate. In conclusion, our simulation provides strong evidence that the convergence conditions implemented here prove adequate and the simple and computationally feasible QAML estimator implemented here provides a valid inference on the diffusion parameters.

4.5 Conclusions

In this chapter we investigate the convergence of discrete time conditional correlation GARCH models towards continuous time diffusion processes and therefore possible substitutability and complementarity between the two processes. This approach allows us on one hand to have a deeper understanding of the properties of the discrete time model and on the other hand to simplify the inference on the continuous time model by using the discrete time one as a diffusion approximation.

We consider two conditional correlation GARCH specifications, namely the cDCC of Aielli (2006) and the CCC of Bollerslev (1990). For the cDCC-GARCH model, as the time step shrinks to zero and under the conditions advocated by the theory of weak convergence of Markov chains, the diffusion limit recovered is degenerate in the sense that the diffusion of the variances and that of the diagonal elements of the correlation driving process are pairwise governed by the same Brownian motion. We show that this result is due to the particular structure of the noise propagation system of the variances and of the correlation driving process which are perfectly correlated.

The CCC model, which can be obtained from the cDCC under suitable parameter restrictions, admits a non-degenerate diffusion.

Under a different set of conditions regarding the speed of convergence of the parameters, we identify other degenerate diffusion limits characterized by stochastic price process but where either the variances, the correlations, or both, are time varying but deterministic.

In a Monte Carlo simulation we study how and to what extent considering the discrete time model as a diffusion approximation allows to infer the parameters of the diffusion model by the parameter estimates of a discrete time GARCH-type model. Our results show that the convergence theory presented in this chapter proves adequate and that quasi approximated maximum likelihood estimator provides valid inference on the diffusion parameters.

There are several possible extension of this work. Apart from the diffusion limit of several univariate GARCH specifications, Nelson (1990) also derives a closed form for the stationary distribution of the continuous time limit of variances and innovations. One direction for future research is to extend these results to the conditional correlation specifications and to elaborate on the distribution of variances and correlations. A second direction for research is to generalize the results of Drost and Nijman (1993) and Drost and Werker (1996) which allow to establish exact, rather than approximated, relationships between discrete and continuous time parameters. Our results show that using approximated relationships allows to accurately infer the diffusion parameters provided the data used to estimate the discrete time model are sampled at a sufficiently high frequency. This may be problematic

in empirical applications where parameter estimates may be biased in presence of market microstructure noise, intra-period periodicity or simply when sufficiently high frequency data are just not available. On the contrary, exact relationships allow to infer the diffusion parameters free of approximation error even when using data sampled at a rather low frequency. Finally, it would be useful to extend the results of this chapter to jump-diffusion processes. Results on jump-diffusion approximations can be fount in Ethier and Kurtz (1986b), though a generalization to multivariate systems has not been addressed yet.

4.6 Appendix: Proofs

Proof of Proposition 4.1. The first step is to compute the increments of the process (4.9)-(4.11), that is

$$\begin{aligned}
{}_hY_{kh} - {}_hY_{(k-1)h} &= {}_hS_{kh}\,{}_h\eta_{kh} \\
{}_hV_{(k+1)h} - {}_hV_{kh} &= c_h + A_h\,{}_hS_{kh}^2 h^{-1}({}_h\eta_{kh} \odot {}_h\eta_{kh}) + (B_h - I_N)\,{}_hV_{kh} \\
{}_hQ_{(k+1)h} - {}_hQ_{kh} &= (1-\vartheta_h-\gamma_h)\bar{Q} + \vartheta_h h^{-1}\text{vech}({}_hP_{khh}\eta_{kh}\,{}_h\eta'_{khh}P_{kh}) + (\gamma_h - 1)_hQ_{kh}
\end{aligned}$$

Second we compute the moment conditions (conditioned on $M_{kh} = \{kh,\ {}_hY_0,...,{}_hY_{(k-1)h},\ {}_hV_0,...,{}_hV_{kh},\ {}_hQ_0,...,{}_hQ_{kh}\}$) to define suitable convergence conditions as required by Assumption 4.1. To simplify the notation, let us define the difference operator over an interval of size h as $\Delta : \Delta\,{}_hX_{kh} =_h X_{kh} -_h X_{(k-1)h}$. The first conditional moment per unit of time of the increments of (4.9)-(4.11) is given by

$$h^{-1}\text{E}[\Delta\,{}_hY_{kh}|M_{kh}] = {}_hS_{kh}\,\text{E}[\,{}_h\eta_{kh}] = 0 \tag{4.46}$$

$$\begin{aligned}
h^{-1}\text{E}[\Delta\,{}_hV_{(k+1)h}|M_{kh}] &= h^{-1}c_h + A_h\,{}_hh^{-2}S_{kh}^2\text{E}[\,{}_h\eta_{kh} \odot {}_h\eta_{kh}] + h^{-1}(B_h - I_N)\,{}_hV_{kh} \\
&= h^{-1}c_h + h^{-1}(A_h + B_h - I_N)\,{}_hV_{kh}
\end{aligned} \tag{4.47}$$

$$\begin{aligned}
h^{-1}\text{E}[\Delta\,{}_hQ_{(k+1)h}|M_{kh}] &= h^{-1}(1-\vartheta_h-\gamma_h)\bar{Q} + h^{-2}\vartheta_h\text{vech}(\text{E}[\,{}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh}]) \\
&\quad + h^{-1}(\gamma_h - 1)\,{}_hQ_{kh} \\
&= h^{-1}(1-\vartheta_h-\gamma_h)\bar{Q} + h^{-1}(\vartheta_h + \gamma_h - 1)\,{}_hQ_{kh}
\end{aligned} \tag{4.48}$$

where $\text{E}[\,{}_h\eta_{kh} \odot {}_h\eta_{kh}] = h\,1_N$ and $\text{vech}(\,{}_hP_{kh}\text{E}[\,{}_h\eta_{kh}\,{}_h\eta'_{kh}]\,{}_hP_{kh}) = h\,\text{vech}(\,{}_hP_{kh}\,{}_hR_{kh}\,{}_hP_{kh}) = h\,{}_hQ_{kh}$.

The computation of the second moments per unit of time require some more algebra.

Let us consider the following partition

$$\text{vech}\left(\text{Var}([\Delta Y_h, \Delta_h V_{(k+1)h}, \Delta_h Q_{(k+1)h}]'|M_{kh})\right) = \begin{bmatrix} \text{Var}(\Delta_h Y_{kh}|M_{kh}) \\ \text{Cov}(\Delta_h Y_{kh}, \Delta_h V_{(k+1)h}|M_{kh})' \\ \text{Cov}(\Delta_h Y_{kh}, \Delta_h Q_{(k+1)h}|M_{kh})' \\ \text{Var}(\Delta_h V_{(k+1)h}|M_{kh}) \\ \text{Cov}(\Delta_h V_{(k+1)h}, \Delta_h Q_{(k+1)h}|M_{kh})' \\ \text{Var}(\Delta_h V_{(k+1)h}|M_{kh}) \end{bmatrix}.$$

The conditional variance of $\Delta_h Y_{kh}$ standardized by h is given by

$$h^{-1}\text{Var}[\Delta_h Y_{kh}|M_{kh}] = h^{-1}\,{}_hS_{kh}\text{E}({}_h\eta_{kh}\,{}_h\eta'_{kh})\,{}_hS'_{kh} = {}_hS_{kh}\,{}_hR_{kh}\,{}_hS_{kh}. \tag{4.49}$$

Similarly the conditional variance of $\Delta_h V_{(k+1)h}$ is given by

$$\begin{aligned} h^{-1}\text{Var}[\Delta_h V_{(k+1)h}|M_{kh}] &= A_h\,{}_hS^2_{kh}h^{-3}\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})({}_h\eta_{kh}\odot\,{}_h\eta_{kh})']\,{}_hS^2_{kh}A'_h \\ &\quad - A_h\,{}_hS^2_{kh}h^{-3}\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})]\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})]'\,{}_hS^2_{kh}A'_h \\ &= A_h\,{}_hS^2_{kh}h^{-3}\left[\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})({}_h\eta_{kh}\odot\,{}_h\eta_{kh})']\right. \\ &\quad \left.-\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})]\text{E}[({}_h\eta_{kh}\odot\,{}_h\eta_{kh})]'\right]\,{}_hS^2_{kh}A'_h \end{aligned} \tag{4.50}$$

Since (time subscripts are omitted to simplify the notation) $\text{E}[(\eta\odot\eta)(\eta\odot\eta)']$ is the $(N\times N)$ matrix of fourth moments of η with elements $\text{E}[\eta_i^2\,\eta_j^2]$, for all $i,j=1,...,N$ equal to

$$\begin{aligned} \text{E}[\eta_i^2\,\eta_j^2] &= h^2 3 \quad \text{if } i=j \\ \text{E}[\eta_i^2\,\eta_j^2] &= h^2(1+2\rho_{i,j}^2) \quad \text{if } i\neq j \end{aligned}$$

and $\text{E}[(\eta\odot\eta)]\text{E}[(\eta\odot\eta)]' = h^2(1_N 1'_N)$, then (4.50) simplifies to

$$h^{-1}\text{Var}[\Delta_h V_{(k+1)h}|M_{kh}] = 2h^{-1}A_h\,{}_hS^2_{kh}({}_hR_{kh}\odot\,{}_hR_{kh})_hS^2_{kh}A'_h. \tag{4.51}$$

The variance of $\Delta_h Q_{(k+1)h}$ is given by

$$\begin{aligned} h^{-1}\text{Var}[\Delta_h Q_{(k+1)h}|M_{kh}] &= \vartheta_h^2 h^{-3}\text{E}[\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh})\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh})'] \\ &\quad - \vartheta_h^2 h^{-3}\text{E}[\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh})]\text{E}[\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh})]'. \end{aligned} \tag{4.52}$$

By exploiting the diagonality of ${}_hP_{kh}$ and since $\text{E}[\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta'_{kh}\,{}_hP_{kh})] = h^2\,{}_hQ_{kh}$ we

can rewrite (4.52) as

$$h^{-1}\text{Var}[\Delta_h Q_{(k+1)h}|M_{kh}] = \vartheta_h^2 h^{-3}\left[(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime})\right.$$
$$\left.\text{E}[\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')'](D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}) - h^2\,{}_hQ_{kh}\,{}_hQ_{kh}'\right].$$

Since for any vector a it holds $\text{vech}(aa')\text{vech}(aa') = D_N^+(\text{vec}(aa')\text{vec}(aa'))D_N^{+\prime} = D_N^+(aa'\otimes aa')D_N^{+\prime}$, then

$$h^{-1}\text{Var}[\Delta_h Q_{(k+1)h}|M_{kh}] = \vartheta_h^2 h^{-3}\left[(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime})\right.$$
$$\left.\text{E}[D_N^+({}_h\eta_{kh}\,{}_h\eta_{kh}'\otimes {}_h\eta_{kh}\,{}_h\eta_{kh}')D_N^{+\prime}](D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}) - h^2\,{}_hQ_{kh}\,{}_hQ_{kh}'\right].$$

Let us define ${}_hK_{kh} = h^{-2}\text{E}[{}_h\eta_{kh}\,{}_h\eta_{kh}'\otimes {}_h\eta_{kh}\,{}_h\eta_{kh}']$, the $(N^2\times N^2)$ matrix of fourth moments of ${}_h\eta_{kh}$ such that (time indices are omitted for shortness)

$$\text{E}[\eta\eta'\otimes\eta\eta'] = \begin{bmatrix} \text{E}(\eta_1\eta_1\eta_i\eta_j) \\ \vdots \\ \text{E}(\eta_1\eta_N\eta_i\eta_j) \\ \text{E}(\eta_2\eta_1\eta_i\eta_j) \\ \vdots \\ \text{E}(\eta_2\eta_N\eta_i\eta_j) \\ \vdots \\ \text{E}(\eta_N\eta_1\eta_i\eta_j) \\ \vdots \\ \text{E}(\eta_N\eta_N\eta_i\eta_j) \end{bmatrix}$$

for all $i,j = 1,...,N,\ i\leq j$. Given the gaussianity assumption for the innovations, the elements of K are given by

$$\begin{array}{ll} \text{E}(\eta_j^4) = h^2 3 & j = 1,...,N \\ \text{E}(\eta_i\eta_j^3) = h^2 3\rho_{ij} & i,j = 1,...,N;\ i\neq j \\ \text{E}(\eta_i\eta_l\eta_j^2) = h^2(\rho_{il}+2\rho_{ij}\rho_{lj}) & i,j,l = 1,...,N;\ i\neq j\neq l \\ \text{E}(\eta_i\eta_l\eta_m\eta_j) = h^2(\rho_{il}\rho_{mj}+\rho_{im}\rho_{lj}+\rho_{ij}\rho_{lm}) & i,j,l,m = 1,...,N;\ i\neq j\neq l\neq m \end{array}$$

The second conditional moment of the increments of the correlation driving process simpli-

fies to

$$h^{-1}\text{Var}[\Delta_h Q_{(k+1)h}|M_{kh}] = h^{-1}\vartheta_h^2[(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime})(D_N^+ {}_hK_{kh}D_N^{+\prime}) \\ (D_N^+({}_hP_{kh}\otimes_h P_{kh})D_N^{+\prime}) - {}_hQ_{khh}Q_{kh}']. \quad (4.53)$$

Finally, the conditional covariances are

$$h^{-1}\text{Cov}[\Delta_h Y_{kh}, \Delta_h V_{(k+1)h})|M_{kh}] = h^{-1}\text{E}[({}_hS_{kh}\,{}_h\eta_{kh})(A_h\,{}_hS_{kh}^2 h^{-1}({}_h\eta_{kh}\odot {}_h\eta_{kh}))'] \\ = h^{-2}\,{}_hS_{kh}\text{E}[{}_h\eta_{kh}({}_h\eta_{kh}\odot {}_h\eta_{kh})']\,{}_hS_{kh}^2 A_h = 0 \quad (4.54)$$

where $\text{E}[\eta(\eta\odot\eta)']$ is a matrix that holds the third moments of η, $\text{E}(\eta_i\eta_j^2)\ \forall i\neq j$ all equal to zero given the gaussianity assumption for the innovations. Further, we have

$$h^{-1}\text{Cov}[\Delta_h Y_{kh}, \Delta_h Q_{(k+1)h}|M_{kh}] = h^{-2}\text{E}[({}_hS_{kh}\,{}_h\eta_{kh})(\vartheta_h\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta_{kh}'\,{}_hP_{kh}))'] \\ = h^{-2}\vartheta_h\,{}_hS_{kh}\text{E}[{}_h\eta_{kh}\,\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')'](D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}) = 0 \quad (4.55)$$

and

$$h^{-1}\text{Cov}[\Delta_h V_{(k+1)h}, \Delta_h Q_{(k+1)h}|M_{kh}] = \\ = h^{-3}\text{E}\left[\left(A_h\,{}_hS_{kh}^2({}_h\eta_{kh}\odot {}_h\eta_{kh})\right)\left(\vartheta_h\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta_{kh}'\,{}_hP_{kh})\right)'\right] \\ - h^{-3}\text{E}\left[A_h\,{}_hS_{kh}^2({}_h\eta_{kh}\odot {}_h\eta_{kh})\right]\text{E}\left[\vartheta_h\text{vech}({}_hP_{kh}\,{}_h\eta_{kh}\,{}_h\eta_{kh}'\,{}_hP_{kh})\right]'.$$

Using ${}_h\eta_{kh}\odot {}_h\eta_{kh} = \text{diag}({}_h\eta_{kh}\,{}_h\eta_{kh}') = I^*\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')$ and defined ${}_hK_{kh}$ as before

$$h^{-1}\text{Cov}[\Delta_h V_{(k+1)h}, \Delta_h Q_{(k+1)h}|M_{kh}] = \\ = h^{-3}\vartheta_h A_h\,{}_hS_{kh}^2\left(\text{E}[I^*\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')\text{vech}({}_h\eta_{kh}\,{}_h\eta_{kh}')'])(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}\right) \\ -h^2 1_N\,{}_hQ_{kh}'\big) \\ = h^{-1}\vartheta_h A_h\,{}_hS_{kh}^2\left[I^*(D_N^+\,{}_hK_{kh}D_N^{+\prime})(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}) - 1_N\,{}_hQ_{kh}'\right]. \quad (4.56)$$

For the moment conditions (4.46)-(4.48) (drift) and (4.49), (4.51), and (4.53)-(4.56) (second moments) to converge to well behaved functions as $h\to 0$, as required by Assumption 4.1

a) and b), the following limits must exist and be finite

$$\lim_{h\to 0} h^{-1} c_h = c \tag{4.57}$$

$$\lim_{h\to 0} h^{-1}(A_h + B_h - I_N) = -\Lambda \tag{4.58}$$

$$\lim_{h\to 0} h^{-1/2} A_h = A \tag{4.59}$$

$$\lim_{h\to 0} h^{-1}(1 - \vartheta_h - \gamma_h) = \phi \tag{4.60}$$

$$\lim_{h\to 0} h^{-1/2} \vartheta_h = \vartheta, \tag{4.61}$$

where c is a $(N \times 1)$ vector, A and Λ are $(N \times N)$ diagonal matrices and ϕ and ϑ are scalars with all elements positive and finite, such that $c > 0$ (elementwise) ensures positivity of the variance process, $A > 0$ and $\vartheta > 0$ ensure that the rescaled second conditional moment of ${}_hV_{kh}$ and ${}_hQ_{kh}$ does not vanish as $h \to 0$, while $\Lambda > 0$ and $\phi > 0$ ensure covariance stationarity of the return process. Finally, it can be easily shown through tedious computation that, under (4.57)-(4.61), Assumption 4.1 c) holds for $\delta = 2$, i.e.,

$$\begin{aligned}
h^{-1}\lim_{h\to 0} E\left[\left|(\Delta\, {}_hY_{kh})_i\right|^4 |M_{kh}\right] &= 0, \forall i, i = 1, ..., N \\
h^{-1}\lim_{h\to 0} E\left[\left|(\Delta\, {}_hV_{(k+1)h})_i\right|^4 |M_{kh}\right] &= 0, \forall i, i = 1, ..., N \\
h^{-1}\lim_{h\to 0} E\left[\left|(\Delta\, {}_hQ_{(k+1)h})_i\right|^4 |M_{kh}\right] &= 0, \forall i, i = 1, ..., N(N+1)/2.
\end{aligned}$$

which completes the proof. ■

Proof of Proposition 4.2. Substituting (4.57)-(4.61) into (4.46)-(4.48) (first moments) and (4.49), (4.51), and (4.53)-(4.56) (second moments), we obtain

$$\begin{aligned}
h^{-1}\mathrm{E}[\Delta\, {}_hY_{kh}|M_{kh}] &= 0 \\
h^{-1}\mathrm{E}[\Delta\, {}_hV_{(k+1)h}|M_{kh}] &= c - \Lambda\, {}_hV_{kh} + o(1) \\
h^{-1}\mathrm{E}[\Delta\, {}_hQ_{(k+1)h}|M_{kh}] &= \phi(\bar{Q} + {}_hQ_{kh}) + o(1)
\end{aligned}$$

for the drift, while for the second moment

$$\begin{aligned}
h^{-1}\mathrm{Var}[\Delta\, {}_hY_{kh}|M_{kh}] &= {}_hS_{kh}\, {}_hR_{kh}\, {}_hS_{kh} \\
h^{-1}\mathrm{Var}[\Delta\, {}_hV_{(k+1)h}|M_{kh}] &= 2A\, {}_hS_{kh}^2({}_hR_{kh} \odot_h R_{kh})\, {}_hS_{kh}^2\, A + o(1) \\
h^{-1}\mathrm{Var}[\Delta\, {}_hQ_{(k+1)h}|M_{kh}] &= \vartheta^2[(D_N^+({}_hP_{kh} \otimes {}_hP_{kh})D_N^{+\prime})(D_N^+ {}_hK_{kh} D_N^{+\prime}) \\
&\qquad (D_N^+({}_hP_{kh} \otimes_h P_{kh})D_N^{+\prime}) - {}_hQ_{kh}{}_hQ_{kh}'] + o(1)
\end{aligned}$$

$$h^{-1}\text{Cov}[\Delta_h Y_{kh}, \Delta_h V_{(k+1)h}]|M_{kh}) = 0$$
$$h^{-1}\text{Cov}[\Delta_h Y_{kh}, \Delta_h Q_{(k+1)h}]|M_{kh}) = 0$$
$$h^{-1}\text{Cov}[\Delta_h V_{(k+1)h}, \Delta_h Q_{(k+1)h}]|M_{kh}) = \vartheta A_h S_{kh}^2 [I^*(D_N^+ {}_h K_{kh} D_N^{+\prime}) \left(D_N^+ ({}_h P_{kh} \otimes {}_h P_{kh}) D_N^{+\prime}\right) - 1_N {}_h Q'_{kh}] + o(1).$$

Hence, as $h \to 0$, the following mappings

$$b(X_t, t) = \begin{bmatrix} 0 \\ c - \Lambda V_t \\ \phi(\bar{Q} - Q_t) \end{bmatrix} \tag{4.62}$$

and

$$\text{vech}(a(X_t, t)) = \begin{bmatrix} S_t R_t S_t \\ 0 \\ 0 \\ 0 \\ 2AS_t^2 (R_t \odot R_t) S_t^2 A \\ \vartheta \left[I^*(D_N^+ K_t D_N^{+\prime}) \left(D_N^+ (P_t \otimes P_t) D_N^{+\prime}\right) - 1_N Q_t'\right]' S_t^2 A \\ \vartheta^2 [(D_N^+ (P_t \otimes P_t) D_N^{+\prime})(D_N^+ K_t D_N^{+\prime})(D_N^+ (P_t \otimes P_t) D_N^{+\prime}) - Q_t Q_t'] \end{bmatrix} \tag{4.63}$$

are solution of (4.1) and (4.2) and represent the drift and the diffusion matrix of the diffusion process $X_t = [Y_t, V_t, Q_t]'$.

We show next that the diffusion matrix (4.63) is singular and that $\text{rank}(a(X_t, t)) = N(N+3)/2 < \dim(a(X_t, t)) = N(N+5)/2$. Let us partition the diffusion matrix as

$$a([Y_t, V_t, Q_t]', t)) = \begin{bmatrix} \Sigma_{YY} & 0 & 0 \\ 0 & \Sigma_{VV} & \Sigma_{VQ} \\ 0 & \Sigma_{QV} & \Sigma_{QQ} \end{bmatrix}, \tag{4.64}$$

where Σ_{YY} and Σ_{VV} are symmetric $(N \times N)$ matrices, Σ_{QQ} is a symmetric $(N(N+1)/2 \times N(N+1)/2)$ matrix and Σ_{VQ} and Σ_{QV} are a $(N \times N(N+1)/2)$ and $(N(N+1)/2 \times N)$ matrices respectively. Since $a([Y_t, V_t, Q_t]', t))$ is symmetric it obviously holds $\Sigma_{VQ} = \Sigma'_{QV}$. Consider the bottom-right block of (4.64), we have

$$\begin{aligned}
\Sigma_{VV} &= 2AS_t^2 (R_t \odot R_t) S_t^2 A \\
\Sigma_{VQ} &= \vartheta \left[I^*(D_N^+ K_t D_N^{+\prime}) \left(D_N^+ (P_t \otimes P_t) D_N^{+\prime}\right) - 1_N Q_t'\right]' S_t^2 A \\
\Sigma_{QV} &= \vartheta A S_t^2 \left[I^*(D_N^+ K_t D_N^{+\prime}) \left(D_N^+ (P_t \otimes P_t) D_N^{+\prime}\right) - 1_N Q_t'\right] \\
\Sigma_{QQ} &= \vartheta^2 \left[(D_N^+ (P_t \otimes P_t) D_N^{+\prime}) D_N^+ K_t D_N^{+\prime} (D_N^+ (P_t \otimes P_t) D_N^{+\prime}) - Q_t Q_t'\right].
\end{aligned}$$

We can now rewrite the bottom-right block of (4.64) as

$$\begin{bmatrix} \Sigma_{VV} & \Sigma_{VQ} \\ \Sigma_{QV} & \Sigma_{QQ} \end{bmatrix} = \begin{bmatrix} AS_t^2 & 0 \\ 0 & \vartheta I_{N(N+1)/2} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} S_t^2 A & 0 \\ 0 & \vartheta I_{N(N+1)/2} \end{bmatrix},$$

where

$$\begin{aligned} \underset{(N(N+3)/2\times N)}{C_1} &= \begin{bmatrix} 2(R_t \odot R_t) \\ \left[\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right)(D_N^+ K_t D_N^{+\prime})I^{*\prime} - Q_t 1_N'\right] \end{bmatrix} \\ \underset{(N(N+3)/2\times N(N+1)/2)}{C_2} &= \begin{bmatrix} I^*(D_N^+ K_t D_N^{+\prime})\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) - 1_N Q_t' \\ (D_N^+(P_t \otimes P_t)D_N^{+\prime})D_N^+ K_t D_N^{+\prime}(D_N^+(P_t \otimes P_t)D_N^{+\prime}) - Q_t Q_t' \end{bmatrix}. \end{aligned}$$

Let now select from C_2 the columns corresponding to the diagonal elements of the Q_t^* matrix. Since $Q_t = \text{vech}(Q_t^*)$ this selection can be done by multiplying C_2 by $I^{*\prime}$- i.e., selecting one column every $N - i$ $(i = 1, ..., N-1)$ starting from the first. The resulting $(N^2 + 3N)/2 \times N$ matrix

$$\begin{aligned} C_2 I^{*\prime} &= \begin{bmatrix} \left[I^*(D_N^+ K_t D_N^{+\prime})\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) - 1_N Q_t'\right] I^{*\prime} \\ \left[(D_N^+(P_t \otimes P_t)D_N^{+\prime})D_N^+ K_t D_N^{+\prime}(D_N^+(P_t \otimes P_t)D_N^{+\prime}) - Q_t Q_t'\right] I^{*\prime} \end{bmatrix} \\ &= \begin{bmatrix} I^*(D_N^+ K_t D_N^{+\prime})\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) I^{*\prime} - 1_N \text{vech}(I_N)'\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) I^{*\prime} \\ (D_N^+(P_t \otimes P_t)D_N^{+\prime})D_N^+ K_t D_N^{+\prime}(D_N^+(P_t \otimes P_t)D_N^{+\prime}) I^{*\prime} - Q_t \text{vech}(I_N)'\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) I^{*\prime} \end{bmatrix} \\ &= \begin{bmatrix} \left(I^*(D_N^+ K_t D_N^{+\prime}) - 1_N \text{vech}(I_N)'\right)\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) I^{*\prime} \\ \left[(D_N^+(P_t \otimes P_t)D_N^{+\prime})(D_N^+ K_t D_N^{+\prime}) - Q_t \text{vech}(I_N)'\right]\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right) I^{*\prime} \end{bmatrix} \\ &= \begin{bmatrix} 2(R_t \odot R_t)(P_t P_t) \\ \left[\left(D_N^+(P_t \otimes P_t)D_N^{+\prime}\right)(D_N^+ K_t D_N^{+\prime})I^{*\prime} - Q_t 1_N'\right](P_t P_t) \end{bmatrix} \end{aligned}$$

satisfies $C_2 I^{*\prime} = C_1(P_t P_t)$ (where $P_t P_t$ is a diagonal matrix of dimension N holding on the diagonal p_i^2, $i = 1, ..., N$). Thus, the diffusion matrix is characterized by N columns linearly dependent and therefore it is singular by construction with $\text{rank}(a([Y_t, V_t, Q_t]')) = N(N+3)/2 < \dim([a(Y_t, V_t, Q_t]')) = N(N+5)/2$. ■

Proof of Theorem 4.1. The process (4.9)-(4.14) is clearly Markovian with drift and second moment per unit of time given by (4.46)-(4.48) (drift) and (4.49), (4.51), and (4.53)-(4.56) (second moments), respectively. The Theorem of weak convergence applies if Assumption 4.1 to 4.4 hold. Proposition 4.1 and 4.2 provide convergence conditions and suitable solutions of the moment conditions in Assumption 4.1. Thus Assumption 4.1 holds and drift and diffusion matrix for the system of stochastic differential equations $\mathrm{d}X_t$ are defined. The

scale matrix $\sigma(X_t, t)$ can be obtained by Cholesky or Spectral decomposition of (4.21) so that Assumption 4.2 holds. We assume that (4.14) satisfies Assumption 4.3 and that for each $h \geq 0$, $\nu_h\left([Y_0, V_0, Q_0]' : V_0 > 0 \text{ (elementwise) and } \epsilon' \mathrm{Q}_0^* \epsilon > 0, \forall \epsilon \in \mathbb{R}^{\mathrm{N}} \backslash \{0\}\right) = 1$. Condition 1 is satisfied given Assumption 4.1 c) which ensures continuity of the sample paths of the limit process X_t with probability one. Condition 2 holds since the diffusion matrix and the inner product of the drift and the state variable X are at most of order two in X. Thus Assumption 4.4 holds and the Theorem of weak convergence applies. The rank condition of the scale matrix $s(X_t, t)$ follows directly from Proposition 4.2 and completes the proof. ■

Proof of Proposition 4.3. The proof follows directly from the proof of Proposition 4.1 under the parameter restriction $\vartheta_h = \gamma_h = 0 \ \forall h$, i.e., ${}_hR_{kh} = (\bar{Q}^* \odot I_N)^{-1/2} \bar{Q}^* (\bar{Q}^* \odot I_N)^{-1/2} = R \ \forall kh, \ k \in \mathbb{N}, \ h > 0$. ■

Proof of Theorem 4.2. The theorem of weak convergence applies by symmetry with the unrestricted model (see Theorem 4.1). In particular Assumption 4.1 holds by symmetry with Proposition 4.2 under the given parameter constraint. Hence substituting (4.57)-(4.59) into the two sets of equations (4.46)-(4.47) and (4.49), (4.51), and (4.54), as $h \to 0$, we obtain the following mappings

$$b(X_t, t) = \begin{bmatrix} 0 \\ c - \Lambda V_t \end{bmatrix} \tag{4.65}$$

and

$$a(X_t, t) = \begin{bmatrix} S_t R S_t & 0 \\ 0 & 2AS_t^2 (R \odot R) S_t^2 A \end{bmatrix} \tag{4.66}$$

which are solution of (4.1) and (4.2) and represent the drift and the diffusion matrix of the diffusion process $X_t = [Y_t, V_t, Q_t]'$. Under the assumption that for each $h \geq 0$, $\nu_h([Y_0, V_0]' : V_0 > 0 \text{ (elementwise)}) = 1$ and $\det(R) > 0$, the diffusion matrix in (4.66) is full rank with $\mathrm{rank}((a(X_t, t)) = N^2$. ■

Proof of Proposition 4.4. Consider the limit, as $h \to 0$, of the moments of interests (4.51), (4.53)and (4.56). Then as required by Assumption 4.1 b)

$$\lim_{h \to 0} h^{-1} \mathrm{Var}[\Delta \, {}_hV_{(k+1)h} | M_{kh}] = \lim_{h \to 0} 2h^{-1} A_h \, {}_hS_{kh}^2 ({}_hR_{kh} \odot \, {}_hR_{kh}) {}_hS_{kh}^2 A_h' \geq 0 \ \text{ finite}$$

which hold with equality if

$$\lim_{h \to 0} h^{-1/2} A_h = 0 \tag{4.67}$$

that is A_h is of order $h^{1/2+\delta}$, $\delta > 0$. Clearly, the rate of $h^{1/2}$, which is the case discussed

in Proposition 4.1, is the slowest possible convergence rate which avoids divergence of the second conditional moment of ${}_hV_{kh}$ as h $\to 0$. Similarly,

$$\lim_{h\to 0} h^{-1}\mathrm{Var}[\Delta\, {}_hQ_{(k+1)h}|M_{kh}] = \lim_{h\to 0} h^{-1}\vartheta_h^2[(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime})(D_N^+\, {}_hK_{kh}D_N^{+\prime}) \\ (D_N^+({}_hP_{kh}\otimes_h P_{kh})D_N^{+\prime}) - {}_hQ_{khh}Q_{kh}'] \geq 0 \text{ finite}$$

holds with equality if

$$\lim_{h\to 0} h^{-1/2}\vartheta_h = 0 \tag{4.68}$$

that is, ϑ_h is of order $h^{1/2+\delta}$, $\delta > 0$. The two conditions above can be written as

$$\lim_{h\to 0} h^{-(1/2+\delta)}A_h = \tilde{A} > 0 \text{ finite}$$
$$\lim_{h\to 0} h^{-(1/2+\delta)}\vartheta_h = \tilde{\vartheta} > 0 \text{ finite.}$$

Finally, either condition (4.67) or (4.68) or both, also ensure that the cross conditional moment of ${}_hV_{kh}$ and ${}_hQ_{kh}$, that is

$$\lim_{h\to 0} h^{-1}\mathrm{Cov}[\Delta\, {}_hV_{(k+1)h}, \Delta\, {}_hQ_{(k+1)h}|M_{kh}] = \\ = \lim_{h\to 0} h^{-1}\vartheta_h A_h\, {}_hS_{kh}^2\left[I^*(D_N^+\, {}_hK_{kh}D_N^{+\prime})(D_N^+({}_hP_{kh}\otimes {}_hP_{kh})D_N^{+\prime}) - 1_N\, {}_hQ_{kh}'\right] \geq 0 \text{ finite}$$

holds with equality. Hence, under (4.67) and (4.68) Assumption 4.1 holds. ■

Proof of Theorem 4.3. Under Proposition 4.4, the theorem of weak convergence applies by symmetry with Theorem 4.1. Furthermore, depending on the combination of condition we impose, we obtain

i) Diffusion with deterministic variances and stochastic correlations
Under (4.57), (4.58), (4.60), (4.61) and (4.67), the second conditional moments scaled by h^{-1} (4.51) and (4.56) vanish as $h \to 0$. The diffusion process $X_t = [Y_t\, V_t\, Q_t]'$ is characterized by a drift given in (4.62) and diffusion matrix given by

$$a(X_t,t) = \begin{bmatrix} S_tR_tS_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \vartheta^2[(D_N^+(P_t\otimes P_t)D_N^{+\prime})(D_N^+K_tD_N^{+\prime}) \\ & & (D_N^+(P_t\otimes P_t)D_N^{+\prime}) - Q_tQ_t'] \end{bmatrix}$$

that is, a process with time varying but deterministic variances of the marginal processes. The diffusion is driven by $N(N+3)/2 = \mathrm{rank}(a(X_t,t)) < \dim(a(X_t,t))$ independent Brow-

nian motions, independent from the initial value $X_0 = [Y_0\, V_0\, Q_0]'$.

ii) Diffusion with stochastic variances and deterministic correlations
Under (4.57), (4.58), (4.59), (4.60) and (4.68), the second conditional moments scaled by h^{-1} (4.53) and (4.56) vanish as $h \to 0$. The diffusion process X_t is characterized by a drift given in (4.62) and diffusion matrix given by

$$a(X_t, t) = \begin{bmatrix} S_t R_t S_t & 0 & 0 \\ 0 & 2AS_t^2(R_t \odot R_r)S_t^2 A & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This process is characterized by a time varying but deterministic correlation driving process. The diffusion is driven by $2N = \text{rank}(a(X_t, t)) < \dim(a(X_t, t))$ independent Brownian motions, independent from the initial value X_0.

iii) Diffusion with deterministic variances and correlations
Under (4.57), (4.58), (4.60) and (4.67)-(4.68), the second conditional moments scaled by h^{-1} (4.51), (4.53) and (4.56) vanish as $h \to 0$. The diffusion process X_t is characterized by a drift given in (4.62) and diffusion matrix given by

$$a(X_t, t) = \begin{bmatrix} S_t R_t S_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This process is characterized by time varying but deterministic variances and correlation driving processes. The diffusion is driven by $N = \text{rank}(a(X_t, t)) < \dim(a(X_t, t))$ independent Brownian motions, independent from the initial value X_0. ■

Bibliography

Aielli, G. P. (2006): "Consistent estimation of large scale dynamic conditional correlations," Department of Statistics, University of Florence.

——— (2009): "Dynamic Conditional Correlations: On Properties and Estimation," Department of Statistics, University of Florence.

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

Alexander, C., and A. Chibumba (1997): "Multivariate Orthogonal Factor GARCH," University of Sussex, Mimeo.

Alexander, C., and E. Lazar (2005): "On the Continuous Limit of GARCH," Discussion Papers in Finance 2005-13.

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, and F. Diebold (2002): *Parametric and non Parametric Volatility Measurement*. Handbook of Financial Econometrics.

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.

——— (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andersen, T., T. Bollerslev, P. Frederiksen, and M. Nielsen (2010): "Continuous-Time Models, Realized Volatilities, and Testable Distributional Implications for Daily Stock Returns," *Journal of Applied Econometrics*, 25(2), 233–261.

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

Ashley, R., C. Granger, and R. Schmelensee (1980): "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149–1168.

Baillie, R., T. Bollerslev, and H. Mikkelsen (1996): "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 74, 3–30.

Bali, T. (2008): "The Intertemporal Relation between Expected Returns and Risk," *Journal of Financial Economics*, 87, 101–131.

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*.

Barndorff-Nielsen, O., and N. Shephard (2004a): "Econometric Analysis of Realized Covariation: High Frequency Covariance, Regression and Correlation in Financial Economics," *Econometrica*, 72.

——— (2004b): "Econometric Analysis of Realized Covariation: High Frequency Covariance, Regression and Correlation in Financial Economics," *Econometrica*, 72, 885–925.

——— (2004c): "Measuring the Impact of Jumps in Multivariate Price Processes Using Bipower Covariation," *DP, Nuffield College, Oxford University*.

——— (2004d): "Power and Bipower Variation with Stochastic Volatility and Jumps," *Journal of Financial Econometrics*, 2, 1–37.

Barndorff-Nielsen, O. E., and N. Shephard (2002): "Estimating Quadratic Variation Using Realised Volatility," *Journal of Applied Econometrics*, 17, 457–477.

Barone-Adesi, G., H. Rasmussen, and C. Ravanelli (2005): "An Option Pricing Formula for the GARCH Diffusion Model," *Computational Statistics and Data Analysis*, 49(2), 287–310.

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

Bauwens, L., M. Lubrano, and F. Richard (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

Becker, R., and A. Clements (2008): "Are Combination Forecasts of S&P 500 Volatility Statistically Superior?," *International Journal of Forecasting*, 24, 122–133.

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.

Bollerslev, T., R. Engle, and D. Nelson (1994): *Handbook of Econometrics*vol. IV, chap. Arch Models. North-Holland.

Bollerslev, T., R. Engle, and J. Wooldridge (1988a): "A Capital Asset Pricing Mode1 with Time Varying Covariances," *Journal of Political Economy*, 96, 116–131.

Bollerslev, T., R. Engle, and J. Wooldridge (1988b): "A Capital Asset Pricing Mode1 with Time Varying Covariances," *Journal of Political Economy*, 96, 116–131.

Bollerslev, T., and J. Wright (2001): "High-frequency Data, Frequency Domain Inference and Volatility Forecasting," *Review of Economics and Statistics*, 83, 596–602.

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

Broze, L., O. Scaillet, and J. Zakoian (1998): "Quasi Indirect Inference for Diffusion Processes," *Econometric Theory*, 14, 161–186.

Caporin, M., and M. McAleer (2010): "Ranking Multivariate GARCH Models by Problem Dimension," *SSRN eLibrary*.

Cappiello, L., R. Engle, and K. Sheppard (2006): "Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns," *Journal of Financial Econometrics*, 4(4), 537–572.

Chiriac, R., and V. Voev (2010): "Modelling and Forecasting Multivariate Realized Volatility," *Journal of Applied Econometrics*, forthcoming.

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

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

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

Clements, A., M. Doolan, S. Hurn, and R. Becker (2009): "Evaluating Multivariate Volatility Forecasts," *NCER Working Paper Series 41*.

Corradi, V. (2000): "Reconsidering the Continuous Time Limit of the GARCH(1,1) Process," *Journal of Econometrics*, 96, 145–153.

Corradi, V., N. Swanson, and C. Olivetti (2001): "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315–358.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196.

Davidson, J. (2004): "Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model," *Journal of Business and Economic Statistics*, 22, 16–29.

de Pooter, M., M. Martens, and D. van Dijk (2008): "Predicting the Daily Covariance Matrix for S&P 100 Stocks Using Intraday Data - But Which Frequency to Use?," *Econometric Reviews*, 27(1-3), 199–229.

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

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.

Dovonon, P., N. Meddahi, and S. Goncalves (2009): "Bootstrapping Realized Multivariate Volatility Measures," Working paper.

Drost, C., and T. Nijman (1993): "Temporal Aggregation of GARCH Processes," *Econometrica*, 61, 909–927.

Drost, F., and J. Werker (1996): "Closing the Gap: Continuous Time GARCH Modeling," *Journal of Econometrics*, 74, 31–57.

Duan, J. (1997): "Augmented GARCH(p,q) Process and its Diffusion Limit," *Journal of Econometrics*, 79, 97–127.

Duffie, D., and K. Singleton (1993): "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929–952.

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

Engle, R. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1008.

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

——— (2002b): "Dynamic Conditional Correlation - A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business and Economic Statistics*, 20, 339–350.

——— (2009): *Anticipating correlations*. Princeton Univ. Press.

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

Engle, R., and R. Colacito (2006): "Testing and Valuing Dynamic Correlations for Asset Allocation," *Journal of Business and Economic Statistics*, 24, 238–253.

Engle, R., and B. Kelly (2008): "Dynamic Equicorrelation," Stern School of Business.

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

Engle, R., and G. Lee (1996): "Estimating Diffusion Models of Stochastic Volatility," in *Modeling Stock Market Volatility: Bridging the Gap to Continuous Time*, ed. by P. Rossi, pp. 333–384. Academic Press.

Engle, R., and J. Mezrich (1995): "Grappling with GARCH," *Risk*, pp. 112–117.

Engle, R., V. Ng, and M. Rothschild (1990): "Asset Pricing with a Factor-ARCH Covariance Structure: Empirical Estimates for Treasury Bills," *Journal of Econometrics*, 45, 213–238.

Engle, R., and K. Sheppard (2001): "Theorical and Empirical properties of Dynamic Conditional Correlation Multivariate GARCH," Mimeo, UCSD.

Epps, T. (1979): "Comovements in Stock Prices in the Very Short Run," *Journal of the American Statistical Association*, 74, 291–298.

Ethier, S., and T. Kurtz (1986a): *Markov Processes: Characterization and Convergence*. John Wiley and Sons.

——— (1986b): *Markov Processes: Characterization and Convergence*. John Wiley and Sons.

Fan, J., M. Wang, and Q. Yao (2008): "Modelling Multivariate Volatilities via Conditionally Uncorrelated Components," *Journal of Royal Statistical Society B*, 70, 679–702.

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

Figlewsky, S., and P. Wachtel (1981): "The Formation of Inflationary Expectation," *Review of Economics and Statistics*, 63, 1–10.

Fornari, F., and A. Mele (1997): "Weak Convergence and Distributional Assumptions for a General Class of Nonlinear ARCH Models," *Econometric Reviews*, 16.

——— (2006): "Approximating Volatility Diffusions with CEV-ARCH Models," *Journal of Economic Dynamics & Control*, 30(6), 931–966.

Francq, C., L. Horvath, and J.-M. Zakoian (2009): "Merits and Drawbacks of Variance Targeting in GARCH Models," MPRA Paper n.15143.

Garman, M., and M. Klass (1980): "On the Estimation of Securities Price Volatilities," *Journal of Business*, 53, 67–78.

Ghysels, E., A. Harvey, and E. Renault (1996): *Handbook of Statistics*vol. 14, chap. Stochastic volatility, pp. 119–191. North-Holland.

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., J. Jasiak, and R. Sufana (2009): "The Wishart Autoregressive process of multivariate stochastic volatility," *Journal of Econometrics*, 150(2), 167–181.

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

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods Theory," *Econometrica*, 52, 681–700.

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

Hafner, C. (2008): "Temporal Aggregation of Multivariate GARCH Processes," *Journal of Econometrics*, 142, 467–483.

Hafner, C., and J. Rombouts (2007): "Estimation of Temporally Aggregated Multivariate GARCH Models," *Journal of Statistical Computation and Simulation*, 77, 629–650.

Hansen, P.R., L. J., and A. Lunde (2008): "Moving Average-Based Estimators of Integrated Variance," *Econometric Reviews*, 27, 79–111.

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

Hansen, P., Z. Huang, and H. Shek (2010): "Realized GARCH: A Complete Model of Returns and Realized Measures of Volatility," *SSRN*.

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.

Hansen, P., A. Lunde, and J. Nason (2003): "Choosing the Best Volatility Models: The Model Confidence Set Approach," Federal Reserve of Atlanta Working Paper.

——— (2010a): "The Model Confidence Set," *Econometrica*, forthcoming.

——— (2010b): "Model Confidence Sets," *Econometrica*, forthcoming.

Herdin, M., N. Czink, H. Ozcelik, and E. Bonek (2005): "Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-stationary MIMO Channels," IEEE-VCT.

Hobson, D., and L. Rogers (1998): "Complete Models with Stochastic Volatility," *Mathematical Finance*, 8, 27–48.

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

Jeantheau, T. (2004): "A link Between Complete Models with Stochastic Volatility and ARCH Models," *Finance and Stochastics*, 8, 111–131.

Jensen, D. (1984): "Invariant Ordering and Order Preservation," *Inequalities in Statistics and Probability*, 5, 26–34.

Jones, C. (2003): "Temporal Aggregation of Multivariate GARCH Processes," *Journal of Econometrics*, 116, 181–224.

Jorion, P. (1995): "Predicting Volatility in the Foreign Exchange Market," *Journal of Finance*, 50, 507–528.

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

Kariya, T. (1988): "MTV Model and its Application to the Prediction of Stock Prices," in *Proceedings of the Second International Tampere Conference in Statistics*, ed. by T. Pullila, and S. Puntanen, Finland. University of Tampere.

Karolyi, G. (1995): "A Multivariate GARCH Model of International Transmissions of Stock Returns and Volatility: The Case of the United States and Canada," *Journal of Business and Economic Statistics*, vol. 13(1), 11–25.

Keane, M., and D. Runkle (1990): "Testing the Rationality of Price Forecasts: New Evidence from Panel Data," *American Economic Review*, 80, 714–735.

Koch, K. (2007): *Introduction to Bayesian Statistics*. Springer Verlag Berlin.

Kushner, H. (1984): *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*. The MIT Press.

Lanne, M., and P. Saikkonen (2007): "A Multivariate Generalized Orthogonal Factor GARCH Model," *Journal of Business and Economic Statistics*, 25, 61–75.

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

Laurent, S., J. Rombouts, and F. Violante (2009): "On Loss Functions and Ranking Forecasting Performances of Multivariate Volatility Models," Cirano discussion paper 2009-45.

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.

Ledoit, O., and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10.

Lewis, A. (2000): *Option Valuation under Stochastic Volatility*. Finance Press.

Lütkepohl, H. (1996): *Handbook of Matrices*. John Wiley and Sons.

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

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

Moskowitz, T. (2003): "An Analysis of Covariance Risk and Pricing Anomalies," *The Review of Financial Studies*, 16, 301–333.

Mykland, P., and L. Zhang (2006): "Anova for Diffusions and Ito Processes," *The Annals of Statistics*, 34, 1931–1963.

Nelson, D. (1990): "ARCH Models as a Diffusion Approximation," *Journal of Econometrics*, 45, 7–38.

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

——— (1994): "Asymptotic Filtering Theory for Multivariate ARCH Models," NBER Technical Working Papers.

Nelson, D., and D. Foster (1994): "Asymptotic Filtering Theory for Univariate ARCH Models," *Econometrica*, 62, 1–41.

Newey, W., and K. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–08.

Pagan, A., and W. Schwert (1990): "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics*, 45, 267–290.

Palandri, A. (2009): "Sequential Conditional Correlations: Inference and Evaluation," *Journal of Econometrics*, 153(2), 122–132.

Pantula, S. (1989): "Testing for Unit Roots in Time Series Data," *Econometric Theory*, 5, 256–271.

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

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

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

Pesaran, M., and A. Timmerman (2007): "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics*, 137, 134–161.

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 (2009a): "Modeling Multivariate Autoregressive Conditional Heteroskedasticity with the Double Smooth Transition Conditional Correlation GARCH Model," *Journal of Financial Econometrics*, 7, 373–411.

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

Steele, J. (2001): *Stochastic Calculus and Financial Applications*. Springer.

Stentoft, L. (2008): "American Option Pricing When the Underlying Assets Follow Augmented GARCH Processes and their Diffusion Limits," HEC Montreal Working Paper.

Stroock, D., and S. Varadhan (1979): *Multidimensional Diffusion Processes*. Springer-Verlag.

Strook, D., and S. Varadhan (1979): *Multidimensional Diffusion Processes*. Springer-Verlag.

Trifi, A. (2006): "Issues of Aggregation Over Time of Conditional Heteroscedastic Volatility Models: What Kind of Diffusion Do We Recover?," *Studies in Nonlinear Dynamics & Econometrics*, 10(4).

Tse, Y., and A. Tsui (2002): "A Multivariate GARCH Model with Time-Varying Correlations," *Journal of Business and Economic Statistics*, 20, 351–362.

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

Voev, V. (2009): "On the Economic Evaluation of Volatility Forecasts," *CREATES Research Paper 2009-56*.

Voev, V., and A. Lunde (2006): "Integrated Covariance Estimation Using High Frequency Data in Presence of Noise," *Journal of Financial Econometrics*, 5.

Vrontos, I., P. Dellaportas, and D. Politis (2003): "A Full-Factor Multivariate GARCH Model," *Econometrics Journal*, 6, 311–333.

Wang, Y. (2002): "Asymptotic Nonequivalence of GARCH Models and Diffusions," *Annals of Statistics*, 30, 754–783.

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

——— (2006): "Forecasts Evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliot, C. Granger, and A. Timmermann.

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

Zarnowitz, V. (1985): "Rational Expectations and Macroeconomic Forecasts," *Journal of The American Statistical Association*, 100, 1394–1411.

Zhang, L., P. Mykland, and Y. Ait-Sahalia (2004): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data," *Journal of The American Statistical Association*, 100, 1394–1411.

——— (2005): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data," *Journal of The American Statistical Association*, 100 (472), 1394–1411.

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