Semiparametric multivariate volatility models

Christian M. Hafner[1] Jeroen V.K. Rombouts[2]

Econometric Institute Report EI 2004-21

Abstract

Estimation of multivariate volatility models is usually carried out by quasi maximum likelihood (QMLE), for which consistency and asymptotic normality have been proven under quite general conditions. However, there may be a substantial efficiency loss of QMLE if the true innovation distribution is not multinormal. We suggest a nonparametric estimation of the multivariate innovation distribution, based on consistent parameter estimates obtained by QMLE. We show that under standard regularity conditions the semiparametric efficiency bound can be attained. Without reparametrizing the conditional covariance matrix (which depends on the particular model used), adaptive estimation is not possible. However, in some cases the efficiency loss of semiparametric estimation with respect to full information maximum likelihood decreases as the dimension increases. In practice, one would like to restrict the class of possible density functions to avoid the curse of dimensionality. One way of doing so is to impose the constraint that the density belongs to the class of spherical distributions, for which we also derive the semiparametric efficiency bound and an estimator that attains this bound. A simulation experiment demonstrates the efficiency gain of the proposed estimator compared with QMLE.

Keywords: Multivariate volatility, GARCH, semiparametric efficiency, adaptivity.
JEL Classification: C14, C22.

[1]Econometric Institute, Erasmus University Rotterdam, P.O.B. 1738, 3000 DR Rotterdam, Netherlands, chafner@few.eur.nl

[2]CORE and Department of Economics, Université catholique de Louvain, Voie du Roman Pays 34, 1348, Louvain-la-Neuve, Belgium, rombouts@core.ucl.ac.be.

1 Introduction

Modelling correlations of multivariate financial time series has attracted considerable interest recently, as computing power more and more enables the researcher to model large covariance matrices in flexible ways. For example, two quite general, nonnested classes of models are the so-called VEC GARCH model described by Engle and Kroner (1995), and the dynamic conditional correlation (DCC) model of Engle (2002), including extensions allowing e.g. for asymmetries analogous to the univariate GARCH literature. For a comparison and other models, see the survey by Bauwens, Laurent and Rombouts (2004).

In this paper we leave the particular form of the conditional covariance matrix unspecified. All we assume is that it depends on a finite dimensional parameter vector, and the objective is to find efficient estimators of this parameter. Estimation by maximum likelihood is straightforward if one supposes a specific parametric distribution of the innovations. However, the choice of this distribution can be quite problematic. Usually one assumes normality, which provides the so-called quasi maximum likelihood estimator (QMLE). It is now well-known that QMLE is consistent under quite general conditions, even if the true underlying distribution is not normal, see e.g. Bollerslev and Wooldridge (1992) and Jeantheau (1998). However, in the case of misspecification there may be a substantial efficiency loss of QMLE compared with the correctly specified maximum likelihood estimator (MLE). On the other hand, assuming a non-normal distribution entails the risk of inconsistent parameter estimation if the distribution is misspecified.

In this paper we follow a nonparametric approach in letting the data determine the distribution of the innovations. With the typically large data sets in finance we would expect to obtain density estimates that are sufficiently close to the true distribution of the innovations. As our model consists of a finite dimensional parameter describing the volatility and correlation dynamics and an infinite dimensional parameter describing the innovation distribution, it may be called semiparametric (SP) as in Engle and Gonzalez-Rivera (1991), who consider univariate GARCH models.

The SP approach is typically more efficient than QMLE. The case where SP estimators are asymptotically as efficient as maximum likelihood estimators where the true distribution is known is usually referred to as adaptivity. For example, in a univariate framework, adaptive estimation of ARMA models has been treated by Kreiss (1987). Often, parameters describing the autoregressive dynamics of a model can be estimated adaptively, whereas scale parameters cannot. For univariate GARCH models, Linton (1993) and Drost and Klaassen (1997) show that adaptive estimators of the autoregressive parameters can be constructed by reparameterizing the volatility process. The same might be possible in the multivariate case, but it will depend on the particular model. As we want this to be sufficiently general, the best we can do is to construct estimators that achieve the semiparametric lower bound. It turns out that this is possible. We show that the semiparametric lower bound is in general different from the parametric lower bound, so that adaptive estimation of the parameter vector is not possible. We characterize some selected distributions and the associated semiparametric lower bound with respect to their distance from the parametric lower bound.

To facilitate the applicability of the model, one can for example assume *a priori* that the innovation distribution belongs to the class of spherical distributions. This bears the important advantage that a nonparametric estimator of the innovation distribution can be constructed in such a way that it has the univariate convergence rate. Hence, there is no 'curse of dimensionality' and the proposed procedure can be applied to highly dimensional systems, if sufficient structure is put on the multivariate volatility model to keep the number of parameters under control. Of course, restricting the class of distributions increases the lower bound of the semiparametric estimator, but it will remain more efficient than QMLE. In a simulation study using a multivariate t distribution, we show that there are substantial efficiency gains of SP over QMLE.

The paper is organized as follows. First, the model framework and the traditional estimation method is introduced. The third section discusses the nonparametric estimation of the innovation distribution, as well as the efficiency of SP estimators. In the fourth section a simulation study is provided. Some Lemmata used in the proofs are given in Appendix B, and the proofs of the propositions in Appendix C. For convenience we summarize some results of matrix algebra and calculus that are used in the paper in Appendix D.

2 The model and assumptions

Consider a vector stochastic process $\{\varepsilon_t\}$ of dimension N with a countable index set and an uncountable state space. We assume that ε_t has the properties of a conditionally heteroskedastic error term, i.e., it has mean zero and is serially uncorrelated. We can write the basic model as

$$\varepsilon_t = H_t(\theta)^{1/2} v_t, \tag{1}$$

where $H_t(\theta)$ is a symmetric and positive definite matrix that may depend on past information up to time $t-1$, and θ is a finite dimensional parameter vector, $\theta \in \Theta \subset R^K$. As usual, we condition on the sigma field generated by all the information (here the ε_t's) until time $t-1$. The σ-field $\mathcal{F}_{t-1}$ contains all this information. Thus, $H_t(\theta)$ is $\mathcal{F}_{t-1}$-measurable.

Occasionally, we will suppress the dependence of H_t on θ for notational convenience. We define the square root of H_t as in (62) so that $H_t^{1/2}$ is also symmetric and positive definite.

In the following we make assumptions about the innovation term v_t, where we denote by I_N the identity matrix of dimension N.

Assumption 1 *The stochastic error $\{v_t\}$ is an i.i.d. sequence with $E[v_t] = 0$, $E[v_t v_t'] = I_N$ and finite fourth moments.*

Because v_t is independent of $\mathcal{F}_{t-1}$, it follows that the conditional covariance matrix of ε_t is H_t. Note that Assumption 1 excludes, for example, a multivariate t distribution with 4 or less degrees of freedom. The assumption of finite fourth moments of v_t does not restrict ε_t to have finite fourth moments. However, to prove consistency and asymptotic normality of estimators such as QML one typically needs higher moments conditions for ε_t as well. Next we make assumptions about the distribution of v_t.

$$\begin{aligned} \mathcal{D} \;=\; & \Big\{ g : \mathbb{R}^N \to \mathbb{R}_{++} \mid \int g(x)dx = 1, \int xg(x)dx = 0, \\ & \int xx'g(x)dx = I_N, \forall i \sup |g^{(i)}(x)| < \infty, M_{\psi\psi} < \infty \Big\} \end{aligned} \tag{2}$$

where $\mathbb{R}_{++} = (0, \infty)$, $g^{(i)}(x)$ denotes the ith partial derivative of $g(\cdot)$, and $M_{\psi\psi}$ is the Fischer information for scale, i.e. $M_{\psi\psi} = \mathrm{E}[\psi_t \psi_t'] < \infty$, where

$$\psi_t(v_t(\theta)) = -\mathrm{vec}\left(I_N + \frac{\partial \log g(v_t)}{\partial v_t} v_t' \right) \tag{3}$$

is the score vector with respect to the scale parameters.

Assumption 2 *Let v_t have density function $g(v_t) \in \mathcal{D}$.*

The assumption of a finite Fischer information for scale is standard in the literature on semiparametric scale models, see e.g., Linton (1993) and Drost and Klaassen (1997). Note that g is not required to be in a parametric class of densities, so that it can depend on a possibly infinite dimensional vector η. The vector η can be regarded as a nuisance parameter in our framework, since we are primarily interested in the estimation of θ.

Next, we summarize regularity conditions that are used by Jeantheau (1998) and Comte and Lieberman (2003) to show consistency of quasi maximum likelihood estimators.

Assumption 3 *We assume the following conditions.*

1. *Θ is compact.*

2. *$\forall\theta_0 \in \Theta$, model (1) admits a unique strictly stationary and ergodic solution ε_t.*

3. *There exists a deterministic constant $c > 0$ such that $\forall t$, $\forall\theta \in \Theta$, $|H_t(\theta)| \geq c$.*

4. *$\forall\theta_0 \in \Theta$, $E[|\log(|H_t(\theta)|)|] < \infty$*

5. *The model is identifiable.*

6. *$H_t(\theta)$ is a continuous function of θ.*

Several specifications have been proposed for H_t, see Bauwens, Laurent and Rombouts (2003) for a survey. Our results concerning semiparametric estimation of θ are sufficiently general to be applicable to any such specification as long as the regularity conditions of Assumption 3 hold. However, in our simulation study of Section 4 we will work with the so–called VEC representation of a multivariate GARCH(p, q) model, which is given by

$$h_t = \text{vech}(H_t) = \omega + \sum_{i=1}^{q} A_i \text{vech}(\varepsilon_{t-i}\varepsilon'_{t-i}) + \sum_{j=1}^{p} B_j \text{vech}(H_{t-j}), \tag{4}$$

where A_i and B_i are $N^* \times N^*$ parameter matrices, and ω is an N^* parameter vector with $N^* = N(N+1)/2$. If one imposes restrictions on the parameters ω, A_i and B_j that guarantee positive definite H_t, such as the so-called BEKK model of Engle and Kroner (1995), then a sufficient condition for stationarity of ε_t is that the eigenvalues of $\sum_{i=1}^{q} A_i + \sum_{j=1}^{p} B_j$ have modulus smaller than one.

We now turn to the problem of estimating θ. If one supposes that the distribution of v_t is known then maximum likelihood estimation (MLE) is in principle straightforward. Nevertheless, because the number of parameters is often large, estimation can become a tedious exercise. If v_t is assumed to be normally distributed with zero mean vector and I_N variance matrix then ε_t will be conditionally normally distributed with zero mean vector and H_t as covariance matrix. The likelihood, up to an additive constant, for a sample of n observations then takes the form

$$L^{qml}(\theta) = -\sum_{t=1}^{n} \frac{1}{2} \log |H_t| - \frac{1}{2} \varepsilon'_t H_t^{-1} \varepsilon_t. \tag{5}$$

Defining $l_t^{qml}(\theta) = -\frac{1}{2}\log|H_t| - \frac{1}{2}\varepsilon'_t H_t^{-1}\varepsilon_t$, we can write $L^{qml}(\theta) = \sum_{t=1}^{n} l_t^{qml}(\theta)$. As shown by Bollerslev and Wooldridge (1992) in a general conditional heteroskedasticity framework, maximizing (5) provides consistent estimates even if the likelihood is misspecified under fairly general conditions. Therefore this method has been termed Quasi Maximum Likelihood (QML) estimation. We next assume finiteness of expectations of the Hessian and the outer product of the gradients.

Assumption 4 *We assume that*

1. $\mathcal{I} = E\left[\frac{\partial l_t^{qml}}{\partial \theta}\frac{\partial l_t^{qml}}{\partial \theta'}\right] < \infty$

2. $\mathcal{J} = -E\left[\frac{\partial^2 l_t^{qml}}{\partial \theta \partial \theta'}\right] < \infty$

3. *For all* i, j, k, $E\left[\sup_{\|\theta-\theta_0\|\leq\delta}\left|\frac{\partial^3 l_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right|\right] < \infty$ *for all* $\delta > 0$,

where expectations are taken with respect to the true distribution and are evaluated at the true parameter vector θ_0.

Under Assumptions 1 to 4, Comte and Lieberman (2003) prove that the asymptotic distribution of QML parameter estimates $\tilde{\theta}$ is given by

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, V_{qml})$$

with $V_{qml} = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$. In the case of correct specification, i.e., the distribution of v_t is indeed multinormal, $\mathcal{J} = \mathcal{I}$, and $V_{qml} = \mathcal{I}^{-1}$. Note that our Assumption 1 of v_t being i.i.d. is stronger than Comte and Lieberman's assumption of v_t being a martingale difference sequence. For much of our paper, Assumption 1 can probably relaxed to v_t being a martingale difference, but we prefer to follow most of the semiparametrics literature, which usually assumes i.i.d. errors, because we rely on many results of this literature. On the other hand, Comte and Lieberman (2003) assume that the components of v_t for a given t are independent, which seems to be a rather strong assumption. We believe that it should be possible to prove their asymptotic result without this independence assumption but leave this open to debate.

While the QML estimator is consistent, it is inefficient if the likelihood is misspecified. Therefore one sometimes considers the multivariate t distribution as an appropriate choice because of potential fat tails in the innovations. The drawback of this assumption is that if the assumption of a specific non-normal distribution is not correct, then in general the estimator may not even be consistent, see e.g. Bollerslev and Wooldridge (1992). Therefore, we will not pursue the assumption of a specific parametric distribution in our paper. In the next section we formalize our motivation for giving all the weight to the data in search for a suitable distribution.

3 Semiparametric estimation

This section describes the methodology used to obtain semiparametric GARCH estimators. We consider two cases. In the first case, no assumption is made about the innovation density besides the regularity conditions of the previous section. In the second case, we assume that the innovation density belongs to the class of spherical densities. We then describe nonparametric density estimators in the alternative cases.

3.1 Semiparametric estimation in the general case

This section describes first a simple iterative estimation procedure and then considers a more efficient semiparametric estimator.

For a general innovation density $g(\cdot)$, the log likelihood may be written as

$$L(\theta) = -1/2 \sum_{t=1}^{n} \log |H_t(\theta)| + \sum_{t=1}^{n} \log g(H_t(\theta)^{-1/2} \varepsilon_t). \tag{6}$$

Note that Assumption 1 requires that $g(\cdot)$ is a density with mean zero and identity covariance matrix. Without this assumption the model would not be identified. In the general case, one idea to estimate $g(\cdot)$ is to first use QMLE (i.e. Gaussian $g(\cdot)$) to obtain standardized residuals, and then estimate the density $g(\cdot)$ nonparametrically. A simple estimation algorithm is the following.

1. Use QMLE to obtain a consistent estimate of θ, $\tilde{\theta}$, say, that gives $\tilde{H}_t = H_t(\tilde{\theta})$.
2. Calculate standardized residuals, $\tilde{v}_t = \tilde{H}_t^{-1/2} \varepsilon_t$. Make sure that they have mean zero and variance I_N.
3. Estimate nonparametrically the density $g(\cdot)$ of $\tilde{v}_t$, giving $\hat{g}(\cdot)$.
4. Maximize L keeping $\hat{g}(\cdot)$ fixed.

This procedure can be viewed as a generalization of the one suggested for univariate GARCH models by Engle and Gonzalez-Rivera (1991). For the estimation of $g(\cdot)$ in step 3, one can use any nonparametric estimation method. For example, we use kernel density estimators as described in Section 3.3. However, as already noticed by Engle and Gonzalez-Rivera (1991), this semiparametric estimator is not likely to achieve the semiparametric lower bound in general.

In the following we propose a semiparametric estimator that attains the semiparametric lower bound and is based on well-known results of the literature on semiparametric estimation. In the context of GARCH models, standard references are Linton (1993), Drost and Klaassen (1997) and Gonzalez-Rivera and Drost (1999). A detailed description of general semiparametric estimation theory and adaptivity is beyond the scope of this paper. We refer to Bickel (1982), Newey (1990), Steigerwald (1992) and Drost, Klaassen and Werker (1997) for details.

One can estimate the semiparametric model using a two-step procedure that uses the so-called influence function to correct an initial consistent estimator such as QML. The correction is essentially a one-step Newton-Raphson algorithm based on the score vector of the likelihood. Assume for now that the unknown density g of v_t is parameterized by some nuisance parameter η, and write this density as $g(v_t, \eta)$. A particular parametrization of g is known as parametric submodel. Let us write the log likelihood as $L(\theta) = \sum_{t=1}^{n} l_t(\theta, \eta)$ with

$$l_t(\theta, \eta) = -\frac{1}{2} \log |H_t(\theta)| + \log g(H_t(\theta)^{-1/2} \varepsilon_t, \eta).$$

and denote by $\dot{l}_t(\theta, \eta) = \partial l_t / \partial \theta$ the score vector w.r.t. the parameter of interest, and by $s_t(\theta, \eta) = \partial l_t / \partial \eta$ the score w.r.t. the nuisance parameter. It is easily seen that $\mathrm{E}[s_t(\theta, \eta)] = 0$. Also, recall that g is restricted to be density of a mean-zero, identity covariance matrix random vector, so that $s_t(\theta, \eta)$ has to be orthogonal to the vector $F_t = (v_t, \mathrm{vech}(v_t v_t' - I_N))$. In the following we suppress the dependence of l_t and $\dot{l}_t$ on η for notational convenience.

In order to obtain efficient estimates, it is required to eliminate the variation of $\dot{l}_t(\theta)$ that is due to the nuisance parameter η. This is achieved by projecting the score on the so-called tangent set, which is the infinite dimensional Hilbert space spanned by all functions with the same characteristics as $s_t(\theta, \eta)$, that is, mean zero and orthogonal to F_t. Thus, the tangent set is defined by

$$\mathcal{T} = \{ f : \mathbb{R}^N \to \mathbb{R}^K \mid \mathrm{E}[f(v_t)] = 0, \mathrm{E}[f(v_t) F_t'] = 0, \mathrm{E}[f(v_t) f(v_t)'] < \infty \}$$

To do the projection against $\mathcal{T}$ it is often crucial to factorize the score $l_t(\theta)$ into a term that only depends on the past and another that depends on the nuisance parameter. The next proposition shows that this is possible for the model under study.

Proposition 1 *For the model (1), the score vector takes the form*

$$\dot{l}_t(\theta) = W_t(\theta) \psi_t(v_t(\theta)) \tag{7}$$

where

$$W_t(\theta) = \frac{\partial vec(H_t)'}{\partial \theta} D_N D_N^+ (I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1}$$

and ψ_t *is given by (3).*

Note that W_t is $\mathcal{F}_{t-1}$-measurable and depends only on the specification of $H_t(\theta)$. The other term ψ_t depends only on the innovation v_t and its density, so that W_t and ψ_t are stochastically independent. As a corollary, note also that the score vector is a martingale difference sequence, which is typically used for deriving the asymptotic distribution of the maximum likelihood estimator. The reason is that $\mathrm{E}[\dot{l}_t(\theta) \mid \mathcal{F}_{t-1}] = W_t \mathrm{E}[\psi_t \mid \mathcal{F}_{t-1}] = W_t \mathrm{E}[\psi_t] = 0$ because $\mathrm{E}[\psi_t] = 0$ as implied by Assumptions 1 and 2.

Let $M_{\psi\psi} = \mathrm{E}[\psi_t \psi_t']$ as in Assumption 2. Furthermore, let $M_{\psi F} = \mathrm{E}[\psi_t F_t']$, $M_{F\psi} = M_{\psi F}'$, and $M_{FF} = \mathrm{E}[F_t F_t']$. We can now derive the projection of $\dot{l}_t(\theta)$ on the tangent set. The orthogonal complement of this projection is the so-called efficient score funtion, that can be used to do a one-step Newton-Raphson improvement of the QML estimator.

Proposition 2 *The projection of $\dot{l}_t(\theta)$ on $\mathcal{T}$ is given by*

$$P_t(\theta) = \mathcal{P}(\dot{l}_t(\theta) \mid \mathcal{T}) = E[W_t(\theta)] \left(\psi_t - M_{\psi F} M_{FF}^{-1} F_t\right). \tag{8}$$

We now propose the following efficient semiparametric estimator,

$$\hat{\theta} = \tilde{\theta} + \left(\sum_{t=1}^{n} \dot{\ell}_t^*(\tilde{\theta}) \dot{\ell}_t^{*\prime}(\tilde{\theta}) \right)^{-1} \sum_{t=1}^{n} \dot{\ell}_t^*(\tilde{\theta}) \tag{9}$$

where $\dot{\ell}_t^*(\theta) = \dot{\ell}_t(\theta) - P_t(\theta)$ is the efficient score function.

Under weak regularity conditions listed by Bickel (1982) and Schick (1986), see also Newey (1990), the asymptotic distribution of the estimator (9) is given by

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, V_{sp})$$

with $V_{sp} = \mathrm{E}[\dot{\ell}_t^* \dot{\ell}_t^{*\prime}]^{-1}$.

By definition, adaptive estimation is possible if and only if $P_t(\theta) = 0$, which means that the semiparametric efficient score, $\dot{\ell}_t^*$, is equal to the parametric score $\dot{\ell}_t$ and, hence, V_{sp} is equal to the parametric lower bound, $V_{ml} = \mathrm{E}[\dot{\ell}_t \dot{\ell}_t']^{-1}$. In the following we characterize the asymptotic covariance matrices of the three estimation methods in terms of W_t and ψ_t.

Proposition 3 *Under Assumptions 1 to 4, the information matrices can be expressed as*

$$V_{ml}^{-1} = E[W_t M_{\psi\psi} W_t'] \tag{10}$$

$$V_{sp}^{-1} = E[W_t M_{\psi\psi} W_t'] - E[W_t] Q E[W_t'] \tag{11}$$

$$V_{qml}^{-1} = E[W_t M_{\psi F} W_t'] E[W_t M_{FF} W_t']^{-1} E[W_t M_{F\psi} W_t'] \tag{12}$$

with

$$Q = M_{\psi\psi} - M_{\psi F} M_{FF}^{-1} M_{F\psi} \tag{13}$$

As a corollary, we obtain that the difference of the information between the MLE and the semiparametric estimator is given by the positive semi-definite matrix

$$V_{ml}^{-1} - V_{sp}^{-1} = \mathrm{E}[W_t]Q\mathrm{E}[W_t'].$$

The matrix Q determines the inefficiency of the SP estimator w.r.t. MLE. Adaptive estimation is possible if and only if $Q = 0$. This would clearly be a special case and we show later that in certain sub-classes of $\mathcal{D}$, $Q = 0$ happens only if v_t is Gaussian distributed. The conclusion is that adaptive estimation of the model in (1) without reparameterization of H_t is not possible. Similar to Gonzalez-Rivera and Drost (1999) it can also be verified that $V_{sp}^{-1} - V_{qml}^{-1}$ is positive semi-definite, meaning that the SP estimator is at least as efficient as the QML estimator.

In practice, the moment matrix M_{FF} can be replaced by its empirical counterpart using the empirical moments of the standardized residuals $\tilde{v}_t$ obtained after the first step. That is, having a consistent estimate $\tilde{\theta}$, one can construct standardized residuals $\tilde{v}_t = H_t(\tilde{\theta})^{-1/2}\varepsilon_t$. Defining further $\widehat{F}_t = (\tilde{v}_t, \mathrm{vech}(\tilde{v}_t\tilde{v}_t' - I_N))$, a consistent estimator M_{FF} is given by $\widehat{M}_{FF} = n^{-1}\sum_{t=1}^{n} \widehat{F}_t\widehat{F}_t'$. The matrix $M_{\psi F}$ does not need to be estimated as the following proposition shows, because an expression is available that holds for any $g \in \mathcal{D}$. To calculate $P_t(\tilde{\theta})$ in (8), one still needs to estimate $\dot{\ell}_t(\tilde{\theta})$ and $\psi_t(\tilde{\theta})$ which both depend on the unknown innovation density g, for which nonparametric methods such as those outlined in Section 3.3 are available.

To characterize the distance of the SP estimator from adaptivity, it will be necessary to evaluate the matrices $M_{\psi\psi}$ and M_{FF} by numerical integration. However, there are special cases where they take simple forms. In the following we consider two sub-classes of $\mathcal{D}$ for which calculation turns out to be particularly simple. The first one is the class of spherical distributions:

$$\mathcal{D}_s = \left\{ g : \mathbb{R}^N \to \mathbb{R}_{++} \mid g \in \mathcal{D}, \exists f : g(x) = f(x'x) \right\} \tag{14}$$

and the second one is the class where the components are independent with identical and symmetric marginal densities:

$$\mathcal{D}_i = \left\{ g : \mathbb{R}^N \to \mathbb{R}_{++} \mid g \in \mathcal{D}, g(x) = \prod_{i=1}^{N} h(x_i), h(x_i) = h(-x_i) \right\}, \tag{15}$$

where $h(\cdot)$ is the marginal density of any component of v_t.

The following discussion analyzes the properties of the SP estimator w.r.t. MLE in the case that g happens to be in $\mathcal{D}_s$ or in $\mathcal{D}_i$ without having made this assumption to construct the estimator. The situation where one has superior knowledge about g and, for example, knows *a priori* that g is in $\mathcal{D}_s$, is different and is treated in Section 3.2.

The matrix M_{FF} depends on the structure of fourth moments of v_t. Lemma 6 implies that, for spherical distributions, the marginal kurtosis $\kappa = E[v_{ti}^4]$ is linked to any co-kurtosis $c = E[v_{ti}^2 v_{tj}^2], j \neq i$ by $\kappa = 3c$, and M_{FF} depends on only one parameter.

Proposition 4 *1. If $g \in \mathcal{D}$, then*

$$M_{\psi F} = \left[\begin{array}{cc} 0_{N^2 \times N} & 2D_N D_N^+ D_N^{+\prime} \end{array} \right]$$

2. If $g \in \mathcal{D}_s$, then

$$M_{FF} = \left[\begin{array}{cc} I_N & 0_{N \times N^*} \\ 0_{N^* \times N} & 2cD_N^+ D_N^{+\prime} + (c-1)vech(I_N)vech(I_N)' \end{array} \right] \quad (16)$$

$$M_{\psi\psi} = 2\tau D_N D_N^+ + (\tau - 1)vec(I_N)vec(I_N)' \quad (17)$$

where

$$\tau = E\left[\left(\frac{\partial \log g(x)}{\partial x_1} \right)^2 x_1^2 \right] /3. \quad (18)$$

3. If $g \in \mathcal{D}_i$, then

$$M_{FF} = \left[\begin{array}{cc} I_N & 0_{N \times N^*} \\ 0_{N^* \times N} & I_{N^*} + [I_{N^*} \odot \{(\kappa - 2)vech(I_N)vech(I_N)'\}] \end{array} \right] \quad (19)$$

$$M_{\psi\psi} = \tau_2 D_N D_N' + D_N \left[I_{N^*} \odot \{(\tau_1 - 1 - \tau_2)vech(I_N)vech(I_N)'\} \right] D_N' \quad (20)$$

where $\tau_1 = 3\tau$, $\odot$ is the Hadamard product (elementwise multiplication), κ is the marginal kurtosis, and

$$\tau_2 = E\left[\left(\frac{\partial \log g(x)}{\partial x_1} \right)^2 \right]. \quad (21)$$

Note that Assumption 2 implies that τ in (17) and τ_2 in (20) are finite. The scalar τ_2 can be interpreted as the Fischer information for location which for distributions in $\mathcal{D}_i$ is the same for all components.

For illustration consider the bivariate case ($N = 2$). From Proposition 4 it follows immediately that, for $g \in \mathcal{D}_s$,

$$Q = D_N \left[\begin{array}{ccc} 3\tau - 1 - \frac{3c-1}{c(2c-1)} & 0 & \tau - 1 - \frac{1-c}{c(2c-1)} \\ 0 & \tau - \frac{1}{c} & 0 \\ \tau - 1 - \frac{1-c}{c(2c-1)} & 0 & 3\tau - 1 - \frac{3c-1}{c(2c-1)} \end{array} \right] D_N',$$

and for $g \in \mathcal{D}_i$,

$$Q = D_N \left[\begin{array}{ccc} \tau_1 - \frac{4}{\kappa-1} & 0 & 0 \\ 0 & \tau_2 - 1 & 0 \\ 0 & 0 & \tau_1 - \frac{4}{\kappa-1} \end{array} \right] D_N'.$$

Clearly, the parametric and semiparametric lower bounds coincide in the Gaussian case since then $c = 1$, $\tau = 1$ and therefore $Q = 0$. Whether there are other distributions for which this happens is our next concern. In a univariate framework, Gonzalez-Rivera (1997) has shown that a class of symmetric bimodal distributions allows to attain the parametric lower bound, and for $N = 1$ this distribution is in $\mathcal{D}_i$. The following proposition states that, for higher dimensions, the Gaussian distribution is the only one in $\mathcal{D}_i$ and in $\mathcal{D}_s$ for which parametric efficiency can be attained. We do not search here for other, possibly asymmetric, distributions for which the two bounds coincide. It may be possible that such distributions exist, but we leave this as a topic for further research.

Proposition 5 *For the estimation of model (1), the multinormal distribution is the only one in $\mathcal{D}_s$ for which the parametric and semiparametric lower bounds coincide. If $N > 1$, then it is also the only distribution in $\mathcal{D}_i$ for which this occurs.*

Let us now look at three examples of distributions in $\mathcal{D}_s$ and at two examples of distributions in $\mathcal{D}_i$. Table 1 reports the spectral norm of Q for these distributions, which for the case of real, positive semi-definite matrices is equal to the spectral radius $\rho(Q)$, i.e., the largest eigenvalue.

1. The density of a symmetric standardized multivariate t distribution is given by

$$g(v_t) = \frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\{\pi(\nu-2)\}^{N/2}\Gamma(\nu/2)}\left(1+\frac{v_t'v_t}{\nu-2}\right)^{-(\nu+N)/2} \tag{22}$$

   where $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$ is the gamma function. To ensure finite fourth moments of v_t we will assume in the following that $\nu > 4$. Under the density given in (22), $c = (\nu-2)/(\nu-4)$ and $\tau = (\nu+N)/(\nu+N+2)$. Note that for $\nu \longrightarrow \infty$, since the limiting distribution is a Gaussian, $c = 1$, $\tau = 1$, and $M_{\psi\psi} = M_{\psi F} = M_{FF} = 2D_N^+D_N^{+\prime}$. For increasing dimensions N, τ converges to 1 and $M_{\psi\psi}$ converges to $2D_N^+D_N^{+\prime}$.

2. The second example is a multivariate Laplace distribution with density

$$g(v_t) = \frac{(N+1)^{N/2}\Gamma(N/2)}{2\pi^{N/2}(N-1)!}\exp(-\sqrt{(N+1)v_t'v_t}) \tag{23}$$

   For $N = 1, 2, 3$ we find $c = (N+3)/(N+1)$ and $\tau = (N+1)/(N+2)$. Although we do not use it here, we conjecture that these formulae for c and τ hold for any N, which would imply that $c \to 1$ and $\tau \to 1$ for $N \to \infty$, which in turn implies using Proposition 4 that the multivariate Laplace density converges to a multinormal distribution with increasing dimension.

3. The third example is an elliptically symmetric (ES) multivariate logistic distribution with density

$$g(v_t) = c_1 \frac{e^{-c_2 v_t' v_t}}{(1 + e^{-c_2 v_t' v_t})^2} \tag{24}$$

with constants c_1 and c_2 such that (24) integrates to one and $\text{Var}(v_{it}) = 1$. We calculate the values c_1, c_2, the co-kurtosis C, and τ by numerical integration. Note that the univariate distribution ($N = 1$) is different from the distribution usually called logistic. For example, the distribution in (24) is platykurtic, whereas the standard logistic is leptokurtic. The ES logistic distribution is mentioned by Jensen (1985).

4. A product of standardized logistics:

$$g(v_t) = \frac{3^{N/2}}{\pi^N} \prod_{i=1}^{N} \frac{e^{-v_{it}}}{(1 + e^{-v_{it}})^2} \tag{25}$$

In the univariate case, $Q = \tau_1 - 1 - 4/(\kappa - 1) = 0.18$. We find that this value is larger than $\tau_2 - 1 \approx 0.10$, so that the largest eigenvalue remains 0.18 for higher dimensions.

5. A product of bimodals:

$$g(v_t) = \left(\frac{\lambda}{2}\right)^{N\lambda/2} \Gamma(\lambda/2)^{-N} \prod_{i=1}^{N} |v_{it}|^{\lambda-1} e^{-\frac{\lambda}{2} v_{it}^2}, \tag{26}$$

where either $\lambda = 1$ (the multinormal case) or $\lambda > 2$ to ensure continuity and differentiability. For $N = 1$, this density has been shown by Gonzalez-Rivera (1997) to give $Q = 0$. Table 1 reports the value of $\rho(Q)$ for the case $\lambda = 3$.

Note that the densities in (22), (23) and (24) are spherical and those in (25) and (26) are in $\mathcal{D}_i$. Note also that the results reported in Table 1 generalize those for the case $N = 1$ listed by Gonzalez-Rivera (1997). In the univariate case, Q is a positive scalar, so that $\rho(Q)$ is just this scalar itself. Figure 1 displays $\rho(Q)$, viewed as a function of the dimension N, for the Laplace and the t_{12} distribution. We noticed that, for the t_ν distribution, there is a break at $N = \nu - 6$, $\nu \geq 7$, in the sense that $\rho(Q)$ is concave for $N \leq \nu - 6$ and for $N \geq \nu - 6$, but not for all $N \geq 1$. We did not try to prove this result but found it curious enough to mention.

3.2 Semiparametric estimation in $\mathcal{D}_s$

The preceding discussion has not imposed any restriction on the density of v_t other than having mean zero and identity covariance matrix and satisfying some weak regularity conditions. Nonparametric estimators of this density will therefore have to be of full generality and have to ensure consistency for any such density. It is well known that nonparametric estimation in high dimensions suffers from the so-called curse of dimensionality, that is, the data sparseness problem described by Silverman (1986). In high dimensions, convergence rates become very slow and the number of observations required to obtain reasonable estimates goes beyond what is typically available in economics, even in finance.

One way to impose more structure on g that solves the dimensionality problem but still leaves sufficient flexibility is to assume that g belongs to the class of spherical distributions, which we do in the following.

Assumption 5 *For the density g, assume $g \in \mathcal{D}_s$, where $\mathcal{D}_s$ is defined by (14).*

The density $g(v_t)$ is said to be spherical if there exists a positive function f such that $g(v_t) = f(w_t)$ with $w_t = v_t'v_t$. Examples of spherical distributions are the multivariate versions of the normal, the t and Laplace distributions. The relevance of these distributions for empirical work arises from the fact that one often observes fat tails in empirical data, even after correcting for time-varying volatility. They exclude skewness, but in finance this has been of minor interest compared with the leptokurtosis effect. The theoretical relevance of Assumption 5 lies in the fact that the class of spherical distributions (or elliptical, respectively, if the conditional distribution of ε_t is considered) is the most general one that is consistent with the conditional capital asset pricing model, as shown by Berk (1997), see also Hodgson and Vorkink (2003). In our framework, the main advantage of Assumption 5 is that it allows for univariate convergence rates of nonparametric estimators of g.

Proposition 6 *Under Assumption 5, the score vector is given by (7) where W_t simplifies to*

$$W_t(\theta) = \frac{\partial vec(H_t)'}{\partial \theta}(H_t^{-1/2} \otimes H_t^{-1/2})$$

and ψ_t is given by (3).

Semiparametric estimators of model (1) that are based on Assumption 5 have a lower bound that is larger than the bound in the general case. The score in the nonparametric direction, $s(\theta, \eta)$, is now a function of v_t only through w_t, and therefore, the tangent set contains functions that all depend on w_t only. But still, these functions have to be orthogonal to F_t because g is restricted to have mean zero and identity covariance matrix. Thus, the tangent set under Assumption 5 is defined by

$$\mathcal{T}_s = \{f : \mathbb{R}_+ \to \mathbb{R}^K \mid \mathrm{E}[f(w_t)] = 0, \mathrm{E}[f(w_t)F_t'] = 0, \mathrm{E}[f(w_t)f(w_t)'] < \infty\}, \quad w_t = v_t'v_t.$$

Note also that $\mathcal{T}_s \subset \mathcal{T}$. We can now derive the projection of $\dot{l}_t(\theta)$ on $\mathcal{T}_s$.

Proposition 7 *Under Assumption 5, the projection of $\dot{l}_t(\theta)$ on $\mathcal{T}_s$ is given by*

$$S_t(\theta) = \mathcal{P}(\dot{l}_t(\theta) \mid \mathcal{T}_s) = E[W_t(\theta)] \left(\tilde{\psi}_t - M_{\tilde{\psi}F} M_{\tilde{F}F}^{-1} \tilde{F}_t \right), \tag{27}$$

where $\tilde{\psi}_t = E[\psi_t \mid w_t]$, $\tilde{F}_t = E[F_t \mid w_t]$, $M_{\tilde{\psi}F} = E[\tilde{\psi}_t F_t']$ and $M_{\tilde{F}F} = E[\tilde{F}_t F_t']$

The efficient semiparametric estimator is given by (9) where $\dot{\ell}_t^*(\theta)$ is replaced by $\dot{\ell}_t^*(\theta) = \dot{\ell}_t(\theta) - S_t(\theta)$.

Note that $\tilde{F}_t$ contains the conditional expectations $\mathrm{E}[v_t \mid w_t]$ and $\mathrm{E}[v_t v_t' \mid w_t]$. The first one is zero since for any given w_t, the distribution of v_t conditional on w_t is symmetric in v_t. For the second conditional expectation, $\mathrm{E}[v_t v_t' \mid w_t]$, Hodgson and Vorkink (2003) describe a simple estimation algorithm after the first estimation step.

Under sphericity, estimation of $\tilde{\psi}_t$ reduces to the estimation $\tilde{F}_t$ by noting that there exists a function f such that $g(v_t) = f(w_t)$ and

$$\mathrm{E}\left[\frac{\partial \log g(v_t)}{\partial v_t} v_t' \middle| w_t \right] = \mathrm{E}\left[2f(w_t)^{-1} \frac{\partial f(w_t)}{\partial w_t} v_t v_t' \middle| w_t \right] = 2f(w_t)^{-1} \frac{\partial f(w_t)}{\partial w_t} \mathrm{E}\left[v_t v_t' \mid w_t \right]$$

Another restriction of $\mathcal{D}$ that may be interesting for practical work is to assume that $g \in \mathcal{D}_i$, where $\mathcal{D}_i$ is given in (15). One would again obtain univariate convergence rates and avoid the curse of dimensionality. Another advantage is that under componentwise independence of v_t it is possible to define impulse response functions for volatility that avoid typical orthogonalization and ordering problems, as shown by Hafner and Herwartz (2004). It is also obvious that Proposition 6 holds under the assumption $g \in \mathcal{D}_i$. However, we have not further investigated efficient semiparametric estimation in this class but leave it to future research.

3.3 Nonparametric density estimation

For the nonparametric density estimation, we use kernel estimates. A general multivariate kernel density estimator with bandwidth matrix H and multivariate kernel $\mathcal{K}$ can be written as

$$\hat{g}_H(x) = \frac{1}{n|H|} \sum_{t=1}^{n} \mathcal{K}(H^{-1}(v_t - x))$$

Since the scale of the variables should be the same (same variance in all directions), it is reasonable to use a scalar bandwidth, $H = hI_N$, with $h > 0$. It is well known that by requiring $nh^N \to \infty$ and $h \to 0$ as $n \to \infty$, the multivariate kernel density estimates are consistent and asymptotically normally distributed. The MSE-optimal rate for the bandwidth is $n^{-1/(4+N)}$. We use here a rule of thumb bandwidth as proposed by Silverman (1986). Furthermore, we use a product kernel $\mathcal{K}(x) = \prod_{i=1}^{N} K(x_i)$ and some univariate kernel function K such as Gaussian, quartic or Epanechnikov. Thus, our density estimate becomes

$$\hat{g}_h(x) = \frac{1}{nh^N} \sum_{t=1}^{n} \prod_{i=1}^{N} K(\frac{v_{i,t} - x_i}{h})$$

For details on multivariate kernel density estimation see the excellent survey of Scott (1992).

The multivariate density estimation becomes difficult for high dimenensional cases. When $g(x)$ is spherical then its density must be of the form $f(x'x)$ for some nonnegative function $f(\cdot)$. Fang, Kotz and Ng (1990, p. 36), show that in this case the density of $y = x'x$ can be written as

$$h(y) = \frac{\pi^{N/2}}{\Gamma(N/2)} y^{N/2-1} f(y). \tag{28}$$

Thus one can obtain an estimator of $g(\cdot)$ by estimating $h(\cdot)$ and transforming according to (28).

Since y has a positive support, one faces the problem of estimating its density near the boundary. To solve this problem, Hodgson, Linton and Vorkink (2002) and Hodgson and Vorkink (2003) apply a Box-Cox transformation to y and then use the standard kernel density estimator to the transformed variable. We use an alternative method by applying a gamma kernel estimator, see Chen (2000) who shows that gamma kernels are particularly suited for the estimation of density functions which have bounded support. The gamma kernel estimator can be written as

$$\hat{h}(y) = \frac{1}{n} \sum_{t=1}^{n} G_{\rho_b(y),b}(y_t) \tag{29}$$

where $G_{p,q}$ is the density function of a Gamma(p, q) random variable, and

$$\rho_b(y) = \begin{cases} y/b & \text{if} \quad y \geq 2b \\ \frac{1}{4}(y/b)^2 + 1 & \text{if} \quad y \in [0, 2b). \end{cases} \tag{30}$$

Chen (2000) also provides formula for the bandwidth b that minimize the mean integrated squared error which we will use in our simulation study.

4 Finite sample performance

In this section we are interested in the performance of the proposed SP estimator (9) relative to the QML and ML estimator in finite samples. Intuitively, the semiparametric method should perform better than QML, but worse than ML, when there are strong departures from normality.

For convenience we consider the VEC model in (4) of order $p = q = 1$. To calculate $W_t(\theta)$ in (7) one needs to evaluate $\frac{\partial \text{vec}(H_t)'}{\partial \theta}$. Deriving with respect to the parameters in ω, A and B we get

$$\begin{aligned} \frac{\partial h_t}{\partial \omega'} &= (I_{N^*} - B)^{-1} \\ \frac{\partial h_t}{\partial \text{vec}(A)'} &= \eta'_{t-1} \otimes I_{N^*} + B \frac{\partial h_{t-1}}{\partial \text{vec}(A)'} \\ \frac{\partial h_t}{\partial \text{vec}(B)'} &= h'_{t-1} \otimes I_{N^*} + B \frac{\partial h_{t-1}}{\partial \text{vec}(B)'} \end{aligned}$$

where $\eta_{t-1} = \text{vech}(\varepsilon_{t-1}\varepsilon'_{t-1})$. The bivariate data generating process is given by

$$h_t = \begin{bmatrix} 1 \\ 0.7 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} \begin{bmatrix} \varepsilon^2_{1,t-1} \\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ \varepsilon^2_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.6 \end{bmatrix} \begin{bmatrix} h_{11,t-1} \\ h_{12,t-1} \\ h_{22,t-1} \end{bmatrix}.$$

For the distributional assumption on v_t we take the bivariate t distribution, that is $v_t \sim t_\nu$ with density given in (22). In this exercise we take $\nu = 5$. We assume that it is known that g belongs to the spherical class. For the nonparametric estimation of $g(\cdot)$, this assumption allows us to use the Gamma kernel estimator as explained in Section 3.3. The employed estimator is thus given by (9) where $\dot{\ell}^*_t(\theta) = \dot{\ell}_t(\theta) - S_t(\theta)$ and $S_t(\theta)$ is given by (27). The results are displayed in Table 2.

Concerning the bias, the three estimation procedures perform similarly, perhaps one could notice that SP performs better than QML for most of the parameters. There are, however, substantial differences between the standard deviations. Clearly ML performs best for all the parameters. SP is as expected in between the two other procedures, also for all the parameters. The same holds true for the MSE of $\hat{\theta}$. One can see that a good part of the loss of the inefficient QML (compared to ML) is recaptured by SP. We also compared the performance of the univariate gamma kernel estimator with the bivariate product kernel estimator when the assumption of sphericity is dropped and the projection $P_t(\theta)$ in (8) is used to construct the estimator (9). The latter estimator had higher MSE's for all parameters, but still considerably less than the QML mean squared errors. However, the relative performance of the multivariate kernel estimator for $N > 2$ is likely to become worse if the dimension increases and if the true density is close to the spherical class.

5 Conclusions and outlook

This paper shows that efficiency gains of semiparametric univariate volatility models over QMLE carry over to the multivariate case. We suggest two alternative types of semiparametric estimators. The first one applies to general innovation densities whereas the second one is based on the assumption of sphericity. A practical advantage of the sphericity assumption is that a nonparametric density estimator can be constructed that has one-dimensional convergence rate. Thus it does not run into the 'curse of dimensionality' problem. Both estimators are efficient but the semiparametric and parametric lower bounds are different in general, so that adaptive estimation is not possible. A guideline for future research may be to find, for particular model specifications, reparameterizations of the conditional covariance matrix such that adaptive estimation of a subset of the parameters is possible, analogous to the univariate case. In this paper, this was not our main concern as we wanted to leave the parametric part of the model unspecified as much as possible, because a multitude of possible specifications have been proposed recently. We think that a main drawback of the general approach outlined in Section 3.1 is that, in practice, it will only be feasible in small dimensions. With the typically high dimensions encountered in finance, for example, this is certainly not a nice feature. In high dimensions one would like to reduce the dimensionality and the assumption of sphericity is only one way of doing this. Another one would be to assume that innovations are componentwise i.i.d., a case that we have not looked at further. There are still other ways to restrict the class of distributions to facilitate the problem of nonparametric estimation in high dimensions, and we leave this also as a topic for further research.

Finally, it may be possible to relax the assumption of an i.i.d. innovation term to, say, a martingale difference term, but proofs become considerably more complicated already in the univariate case, and one would need to control the possible temporal dependence by making further assumptions.

Acknowledgements

The authors would like to thank Luc Bauwens, Geert Dhaene, Feico Drost, Wolfgang Härdle, Jens Peter Kreiss, and Bas Werker for helpful discussions. We also thank participants of the CORE econometrics seminar, the York annual meeting in econometrics, the annual econometric study group meeting 2002 in Bristol, and the 2003 workshop 'The art of semiparametrics' in Berlin for valuable comments.

References

[1] Bauwens, L., S. Laurent & J.V.K. Rombouts (2004) Multivariate GARCH models: A survey. CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

[2] Berk, J. (1997) Necessary conditions for the CAPM. *Journal of Economic Theory* **73**, 245–257.

[3] Bickel, P.J. (1982) On adaptive estimation. *Annals of Statistics* **10**, 647-671.

[4] Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1993) *Efficient and Adaptive Statistical Inference for Semiparametric Models.* John Hopkins University Press, Baltimore, MD.

[5] Bollerslev, T. & J. Wooldridge (1992) Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews* **11**,143-72.

[6] Box, G.E.P. & J.S. Hunter (1957) Multi-factor experimental designs for exploring response surfaces. *Annals of Mathematical Statistics* **28**, 195–241.

[7] Chen, S.X. (2000) Probability density function estimation using Gamma kernels. *Annals of the Institute of Statistical Mathematics* **52**, 471–480.

[8] Comte, F. & O. Lieberman (2003) Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis 84*, 61-84.

[9] Drost, F.C., C.A.J. Klaassen & B.J.M. Werker (1997) Adaptive estimation in time series models. *Annals of Statistics* **25**, 786-817.

[10] Drost, C.D. & C.A.J. Klaassen (1997) Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* **81**, 193-221.

[11] Engle, R.F. (2002) Dynamic conditional correlation - A simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20**, 339–350.

[12] Engle, R. & K. Kroner (1995) Multivariate simultaneous GARCH. *Econometric Theory* **11**, 122-150.

[13] Engle, R. & Gonzalez-Rivera, G. (1991), Semiparametric ARCH models. *Journal of Business and Economic Statistics* **9**, 345-359.

[14] Fang, K.T., S. Kotz & K.W. Ng (1990) *Symmetric Multivariate and Related Distributions*. Chapman and Hall: London.

[15] Gonzalez-Rivera, G. (1997) A note on adaptation in ARCH models. *Econometric Reviews* **16**, 55-68.

[16] Gonzalez-Rivera, G. & C.D. Drost (1999) Efficiency comparisons of maximum-likelihood-based estimators in GARCH models. *Journal of Econometrics* **93**, 93-111.

[17] Hafner, C.M. & H. Herwartz (2004) Volatility impulse response functions for multivariate GARCH models: An exchange rate illustration. *Journal of International Money and Finance*, forthcoming.

[18] Hodgson, D.J. & K.P. Vorkink (2003) Efficient estimation of conditional asset-pricing models. *Journal of Business and Economic Statistics* **52**, 471-480.

[19] Hodgson, D.J., O. Linton & K.P. Vorkink (2002) Testing the capital asset pricing model efficiently under elliptical symmetry: A semiparametric approach. *Journal of Applied Econometrics* **17**, 617-639.

[20] Jeantheau, T. (1998) Strong consistency of estimators for multivariate ARCH models. *Econometric Theory* **14**, 70-86.

[21] Jensen, D.R. (1985) *Multivariate Distributions*. In Kotz, S., N.L. Johnson & C.B. Read (ed.) Encyclopedia of Statistical Sciences, 6, Wiley, pp. 43-55.

[22] Linton, O. (1993) Adaptive estimation of ARCH models. *Econometric Theory* **9**, 539-569.

[23] Kreiss, J.P. (1987) On adaptive estimation in stationary ARMA processes. *The Annals of Statistics* **15**, 112-133.

[24] Lütkepohl, H. (1996) *Handbook of Matrices*. New York: Wiley.

[25] Magnus, J.R. (1988) *Linear Structures*. London: Griffin.

[26] Newey, W.K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99-135.

[27] Schick, A. (1986) On asymptotically efficient estimation in semiparametric models. *Annals of Statistics* **14**, 1139-1151.

[28] Scott, D. (1992) *Multivariate Density Estimation : Theory, Practice, and Visualisation*. New York: Wiley.

[29] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.

[30] Steigerwald, D.G. (1992) Adaptive estimation in time series regression models. *Journal of Econometrics* **54**, 251-275.

Appendix A: Index Set Definitions

Define the index

$$k_{ij}^N = i + (j-1)(N - \frac{j}{2}) \tag{31}$$

and the index sets

$$\mathcal{K}_{ij}^N = \begin{cases} \emptyset & N = 1 \\ \{k_{ij}^N \mid j = 1, \ldots, N-1; \quad i = j+1, \ldots, N\} & N \geq 2 \end{cases} \tag{32}$$

and

$$\mathcal{K}_{ii}^N = \{k_{ii}^N \mid i = 1, \ldots, N\} \tag{33}$$

The index k_{ij}^N is the position of the (i,j)-th element of an $(N \times N)$ symmmetric matrix A in the vector $\text{vech}(A)$. Remember that $\text{vech}(A)$ contains $N^* = N(N+1)/2$ elements. $\mathcal{K}_{ij}^N$ contains all indices of the elements below the diagonal of A in the vector $\text{vech}(A)$, this set contains $N(N-1)/2$ elements. The set $\mathcal{K}_{ii}^N$ contains all indices of the N diagonal elements. For example, for $N = 2$, $\mathcal{K}_{ij}^2 = \{2\}$ and $\mathcal{K}_{ii}^2 = \{1,3\}$, and for $N = 3$, $\mathcal{K}_{ij}^3 = \{2,3,5\}$ and $\mathcal{K}_{ii}^3 = \{1,4,6\}$. Note that $\mathcal{K}_{ij}^N \cup \mathcal{K}_{ii}^N = \{1, \ldots, N^*\}$ and $\mathcal{K}_{ij}^N \cap \mathcal{K}_{ii}^N = \emptyset$.

Appendix B: Lemmata

Lemma 1 *For given nonsingular matrices $A, B(m \times m)$, let $Z = B \otimes A + A \otimes B$. Then*

$$D_m D_m^+ Z^{-1} D_m D_m^+ = D_m D_m^+ Z^{-1} \tag{34}$$

Proof: Assume the contrary. Then, multiplying both sides of (34) from the right by Z gives $D_m D_m^+ Z^{-1} D_m D_m^+ Z \neq D_m D_m^+$. Using (72), one can write the left hand side of this inequality as $D_m D_m^+ Z^{-1} Z D_m D_m^+ = D_m D_m^+ D_m D_m^+ = D_m D_m^+$, making use of the fact that $D_m^+ D_m = I_{m(m+1)/2}$. But this yields a contradiction, so that the original equality in (34) must hold. □

Lemma 2 *For given matrices $A, B(m \times m)$,*

$$D_m'(A \otimes B) D_m D_m^+ = \frac{1}{2} D_m'(A \otimes B + B \otimes A) \tag{35}$$

Proof: Assume the contrary. Vectorizing the left hand side of (35), one obtains $(D_m D_m^+ \otimes D_m')\text{vec}(A \otimes B)$. Using (67) and vectorizing the right hand side of (35), one obtains $\frac{1}{2}(I_{m^2} \otimes D_m' + C_{mm} \otimes D_m' C_{mm})\text{vec}(A \otimes B)$. But using (70), this is equal to $\frac{1}{2}(I_{m^2} \otimes D_m' + C_{mm} \otimes D_m')\text{vec}(A \otimes B) = \frac{1}{2}[(I_{m^2} + C_{mm}) \otimes D_m']\text{vec}(A \otimes B)$. Finally, using (69), this is equal to $(D_m D_m^+ \otimes D_m')\text{vec}(A \otimes B)$, which yields to a contradiction. Therefore, equality must hold in (35). □

Lemma 3 *For a given symmetric, positive definite matrix $A(m \times m)$,*

$$2D_m D_m^+ (I_m \otimes A + A^{1/2} \otimes A^{1/2})^{-1} vec(I_m) = vec(A^{-1}). \tag{36}$$

Proof: Assume the contrary. The left hand side of (36) can be written as

$$2D_m D_m^+ \left[(I_m \otimes A^{1/2})(I_m \otimes A^{1/2} + A^{1/2} \otimes I_m) \right]^{-1} \text{vec}(I_m)$$

or, using (63), as

$$2D_m D_m^+ (I_m \otimes A^{1/2} + A^{1/2} \otimes I_m)^{-1} (I_m \otimes A^{-1/2}) \text{vec}(I_m).$$

Applying Lemma 1, this is equivalent to

$$2(I_m \otimes A^{1/2} + A^{1/2} \otimes I_m)^{-1} D_m D_m^+ (I_m \otimes A^{-1/2}) \text{vec}(I_m).$$

As we assumed inequality in (36), we have that

$$2D_m D_m^+ (I_m \otimes A^{-1/2}) \text{vec}(I_m) \neq (I_m \otimes A^{1/2} + A^{1/2} \otimes I_m) \text{vec}(A^{-1}).$$

or, using (59), $D_m D_m^+ \text{vec}(A^{-1/2}) \neq \text{vec}(A^{-1/2})$. However, since $A^{-1/2}$ is symmetric because of our definition of a matrix square root in (62), it holds that $\text{vec}(A^{-1/2}) = D_m \text{vech}(A^{-1/2})$, and because $D_m^+ D_m = I_{m(m+1)/2}$ this inequality leads to a contradiction, so that equality must hold in (36). Q.E.D.

Lemma 4 *The matrix $D_N^+ D_N^{+\prime}$ is a $(N^* \times N^*)$ diagonal matrix with 1 at the (i,i)-th position, $i \in \mathcal{K}_{ii}^N$, and 1/2 at the (j,j)-th position, $j \in \mathcal{K}_{ij}^N$, where $\mathcal{K}_{ij}^N$ is defined in (32) and $\mathcal{K}_{ii}^N$ in (33).*

Proof: The statement holds for $D_1^+ D_1^{+\prime} = 1$. Noting that $k_{ii}^{N+1} = k_{i-1,i-1}^N + N + 1$, where k_{ij}^N is defined in (31), and using the recursive equation for $D_{N+1}^+ D_{N+1}^{+\prime}$ in (74), the statement follows by induction. Q.E.D.

Lemma 5 *The matrix $vech(I_N) vech(I_N)'$ is a $(N^* \times N^*)$ matrix with 1 at the (i,j)-th position, $i,j \in \mathcal{K}_{ii}^N$, and 0 elsewhere, where $\mathcal{K}_{ii}^N$ is defined in (33).*

Proof: By definition, $\mathcal{K}_{ii}^N$ contains the positions of the diagonal elements of a $(N \times N)$ matrix A in the vector $\text{vech}(A)$. Therefore, $\mathcal{K}_{ii}^N$ contains the positions of the ones in the vector $\text{vech}(I_N)$. The matrix $\text{vech}(I_N)\text{vech}(I_N)'$ then contains ones at pairs of any permutations of these positions, and zeros elsewhere. So there is a total of N^2 ones in the matrix. Q.E.D.

Lemma 6 *For any spherical distribution,*

$$E\left[\prod_{j=1}^{N} X_j^{\alpha_j}\right] = \begin{cases} 0 & \text{if one (or more) } \alpha_j \text{ is odd} \\ K_\alpha \prod_{j=1}^{N} \frac{\alpha_j!}{(\alpha_j/2)!} & \text{if all } \alpha_j \text{ are even} \end{cases}$$

where $\alpha = \sum_{j=1}^{N} \alpha_j$ *and* K_α *depends on* α *only.*

Proof: see Box and Hunter (1957).

Appendix C: Proofs

Proof of Proposition 1:

Let us write the likelihood as $L(\theta) = \sum_{t=1}^{n} l_t(\theta)$ with

$$l_t = -\frac{1}{2}\log|H_t| + \log g(H_t^{-1/2}\varepsilon_t).$$

The score vector is given by

$$\frac{\partial l_t(\theta)}{\partial\theta} = -\frac{1}{2}\frac{\partial \log|H_t|}{\partial\theta} + \frac{\partial \log g(H_t^{-1/2}\varepsilon_t)}{\partial\theta}$$

where the first term has components

$$\frac{\partial \log|H_t|}{\partial\theta_i} = \text{vec}(H_t^{-1})'\frac{\partial \text{vec}(H_t)}{\partial\theta_i} = \text{Tr}\left(H_t^{-1}\frac{\partial H_t}{\partial\theta_i}\right) = \text{Tr}\left(H_t^{-1/2}\frac{\partial H_t}{\partial\theta_i}H_t^{-1/2}\right)$$

using (80). With the chain rule for matrix differentiation (75), we can write

$$\frac{\partial \log g(H_t^{-1/2}\varepsilon_t)}{\partial\theta_i} = \frac{\partial \log g(x)}{\partial x'}\frac{\partial (H_t^{-1/2}\varepsilon_t)}{\partial \text{vec}(H_t)'}\frac{\partial \text{vec}(H_t)}{\partial\theta_i}$$

Applying (77) and (79) we can further write

$$\frac{\partial (H_t^{-1/2}\varepsilon_t)}{\partial \text{vec}(H_t)'} = -(\varepsilon_t' \otimes I_N)(H_t^{-1/2} \otimes H_t^{-1/2})\frac{\partial \text{vec}(H_t^{1/2})}{\partial \text{vec}(H_t)'} \tag{37}$$

Then, $\frac{\partial \text{vec}(H_t^{1/2})}{\partial \text{vec}(H_t)'}$ can be appropriately defined by noting that H_t is symmetric and by the definition of the matrix square root (62) $H_t^{1/2}$ is symmetric as well. By (78) we know that in this case

$$\frac{\partial \text{vech}(H_t)}{\partial \text{vech}(H_t^{1/2})'} = D_N^+(H_t^{1/2} \otimes I_N + I_N \otimes H_t^{1/2})D_N \tag{38}$$

$$= D_N^+ Z D_N \tag{39}$$

with $Z = (H_t^{1/2} \otimes I_N + I_N \otimes H_t^{1/2})$, where D_N^+ denotes the generalized inverse (64) of the duplication matrix D_N. If the matrix Z is invertible, a natural definition for the derivative of the matrix square root is

$$\begin{aligned} \frac{\partial \text{vech}(H_t^{1/2})}{\partial \text{vech}(H_t)'} &= (D_N^+ Z D_N)^{-1} && (40) \\ &= D_N^+ Z^{-1} D_N, && (41) \end{aligned}$$

where (41) uses (73). Using (76) we then obtain

$$\begin{aligned} \frac{\partial \text{vec}(H_t^{1/2})}{\partial \text{vec}(H_t)'} &= D_N D_N^+ Z^{-1} D_N D_N^+ && (42) \\ &= D_N D_N^+ Z^{-1} && (43) \end{aligned}$$

by Lemma 1.

Plugging (43) into (37), we obtain

$$\begin{aligned} \frac{\partial (H_t^{-1/2}\varepsilon_t)}{\partial \text{vec}(H_t)'} &= -(\varepsilon_t' \otimes I_N)(H_t^{-1/2} \otimes H_t^{-1/2}) D_N D_N^+ Z^{-1} && (44) \\ &= -(v_t' \otimes I_N)(H_t^{1/2} \otimes I_N)(H_t^{-1/2} \otimes H_t^{-1/2}) Z^{-1} D_N D_N^+ && (45) \\ &= -(v_t' \otimes I_N)(I_N \otimes H_t^{-1/2}) Z^{-1} D_N D_N^+ && (46) \\ &= -(v_t' \otimes I_N)\{Z(I_N \otimes H_t^{1/2})\}^{-1} D_N D_N^+ && (47) \\ &= -(v_t' \otimes I_N)(I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1} D_N D_N^+ && (48) \end{aligned}$$

where (45) uses the fact that both Z^{-1} and $D_N D_N^+$ are symmetric and (47) uses (63). Thus,

$$\begin{aligned} \frac{\partial \log g(v_t)}{\partial \theta_i} &= -\frac{\partial \log g(v_t)}{\partial v_t'}(v_t' \otimes I_N)(I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1} D_N D_N^+ \frac{\partial \text{vec}(H_t)}{\partial \theta_i} \\ &= -\text{vec}\left(\frac{\partial \log g(v_t)}{\partial v_t} v_t'\right)' (I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1} D_N D_N^+ \frac{\partial \text{vec}(H_t)}{\partial \theta_i} \\ &= -\text{vec}\left(\frac{\partial \log g(v_t)}{\partial v_t} v_t'\right)' W_t' \end{aligned}$$

with $W_t = \frac{\partial \text{vec}(H_t)'}{\partial \theta} D_N D_N^+ (I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1}$. Defining $\psi_t = -\text{vec}(I_N + (\partial \log g(v_t)/\partial v_t) v_t')$, this can be written as

$$\frac{\partial \log g(v_t)}{\partial \theta_i} = \psi_t' W_t' + \text{vec}(I_N)' W_t'.$$

However, by Lemma 3 the term $\text{vec}(I_N)' W_t'$ is equal to $\frac{1}{2}\text{vec}(H_t^{-1})' \frac{\partial \text{vec}(H_t)}{\partial \theta'}$. Thus, the score vector can be written as

$$\frac{\partial l_t(\theta)}{\partial \theta} = \frac{\partial \text{vec}(H_t)'}{\partial \theta} \left\{ -\frac{1}{2}\text{vec}(H_t^{-1}) - D_N D_N^+ (I_N \otimes H_t + H_t^{1/2} \otimes H_t^{1/2})^{-1} \text{vec}\left(\frac{\partial \log g(v_t)}{\partial v_t} v_t'\right) \right\}$$

$$
\begin{aligned}
&= -\frac{1}{2}\frac{\partial \mathrm{vec}(H_t)'}{\partial \theta}\mathrm{vec}(H_t^{-1}) + W_t\psi_t + \frac{1}{2}\frac{\partial \mathrm{vec}(H_t)'}{\partial \theta}\mathrm{vec}(H_t^{-1}) \\
&= W_t\psi_t,
\end{aligned}
$$

as stated. Q.E.D.

Proof of Proposition 2

First note that $P_t \in \mathcal{T}$ because $\mathrm{E}[P_t] = 0$, $\mathrm{E}[P_t F_t'] = 0$, and $\mathrm{E}[P_t P_t']$ is finite since $\mathrm{E}[\psi_t \psi_t'] < \infty$ by Assumption 2 and $\mathrm{E}[F_t F_t'] < \infty$ by Assumption 1 (finite fourth moments).

Next, we show that the orthogonal complement of the projection is orthogonal to $\mathcal{T}$. It can be written as

$$\dot{\ell}_t^*(\theta) = \dot{\ell}_t(\theta) - P_t = (W_t - \mathrm{E}[W_t])\psi_t + \mathrm{E}[W_t] M_{\psi F} M_{FF}^{-1} F_t \tag{49}$$

The first term on the right hand side of (49) is orthogonal to $\mathcal{T}$ since $(W_t - \mathrm{E}[W_t])$ has mean zero and is independent of v_t and, hence, independent of all elements in $\mathcal{T}$. The second term on the right hand side of (49) is orthogonal to $\mathcal{T}$ because it consists of linear combinations of F_t and F_t is, by definition, orthogonal to $\mathcal{T}$. Q.E.D.

Proof of Proposition 3

1. $V_{ml}^{-1} = \mathrm{E}[\dot{l}_t \dot{l}_t'] = \mathrm{E}[W_t \psi_t \psi_t' W_t'] = \mathrm{E}[W_t M_{\psi\psi} W_t']$

2. $$V_{sp}^{-1} = \mathrm{E}[\dot{l}_t \dot{l}_t'] - \mathrm{E}[\dot{l}_t P_t'] - \mathrm{E}[P_t \dot{l}_t'] + \mathrm{E}[P_t P_t'].$$

   The second term is

   $$
   \begin{aligned}
   \mathrm{E}[\dot{l}_t P_t'] &= \mathrm{E}\left[W_t\psi_t(\psi_t' - F_t' M_{FF}^{-1} M_{F\psi})\mathrm{E}(W_t')\right] \\
   &= \mathrm{E}[W_t]\mathrm{E}[\psi_t\psi_t' - \psi_t F_t' M_{FF}^{-1} M_{F\psi}]\mathrm{E}[W_t'] \\
   &= \mathrm{E}[W_t]\left(M_{\psi\psi} - M_{\psi F} M_{FF}^{-1} M_{F\psi}\right)\mathrm{E}[W_t']
   \end{aligned}
   $$

   Similar calculations show that $\mathrm{E}[\dot{l}_t P_t'] = \mathrm{E}[P_t \dot{l}_t'] = \mathrm{E}[P_t P_t']$, which then gives the stated result.

3. Recall that $V_{qml} = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$, and thus $V_{qml}^{-1} = \mathcal{J}\mathcal{I}^{-1}\mathcal{J}$ with

   $$\mathcal{J} = -\mathrm{E}\left[\frac{\partial^2 l_t^{qml}}{\partial\theta\partial\theta'}\right], \quad \mathcal{I} = \mathrm{E}\left[\frac{\partial l_t^{qml}}{\partial\theta}\frac{\partial l_t^{qml}}{\partial\theta'}\right],$$

   and $\frac{\partial l_t^{qml}}{\partial\theta} = W_t K_N F_t$, where $K_N = \left[\begin{array}{cc} 0_{N^2\times N} & D_N \end{array}\right]$. We have

   $$\mathcal{I} = \mathrm{E}[W_t K_N F_t F_t' K_N' W_t'] = \mathrm{E}[W_t K_N M_{FF} K_N' W_t']$$

and

$$\mathcal{J} = -\int_{I\!R^N} g(x)\frac{\partial^2 l_t^{qml}}{\partial\theta\partial\theta'}dx = \int_{I\!R^N} \frac{\partial g(x)}{\partial\theta}\frac{\partial l_t^{qml}}{\partial\theta'}dx \tag{50}$$

$$= \int_{I\!R^N} g(x)\frac{\partial \log g(x)}{\partial\theta}\frac{\partial l_t^{qml}}{\partial\theta'}dx = \mathrm{E}\left[\frac{\partial \log g(x)}{\partial\theta}\frac{\partial l_t^{qml}}{\partial\theta'}\right] \tag{51}$$

$$= \mathrm{E}\left[W_t\psi_t F_t K_N' W_t'\right] = \mathrm{E}\left[W_t M_{\psi F} K_N' W_t'\right] \tag{52}$$

Q.E.D.

Proof of Proposition 4

1. $M_{\psi F}$: Note first that ψ_t is orthogonal to v_t: $\mathrm{E}[\psi_t v_t'] = -\mathrm{E}[\mathrm{vec}\{(\partial \log g(v_t)/\partial v_t)v_t'\}v_t']$ and writing this expectation elementwise one obtains $-\int_{I\!R^N}(\partial g(v_t)/\partial v_{it})v_{jt}v_{kt}dv_t$. Using integration by parts, for $i = j = k$ this is just equal to $2\mathrm{E}[v_{it}] = 0$, and for $i = j \neq k$ it is equal to $\mathrm{E}[v_{kt}] = 0$, and the same holds for $i \neq j \neq k$ and $i \neq j = k$. Hence, the left block of $M_{\psi F}$ is equal to $0_{N^2\times N}$.

   The right block can be written as

   $$-\mathrm{E}[\mathrm{vec}(\frac{\partial \log g(v_t)}{\partial v_t}v_t')\mathrm{vech}(v_t v_t')'] - \mathrm{vec}(I_N)\mathrm{vech}(I_N)'. \tag{53}$$

   Writing the expectation term in (53) elementwise, one obtains $-\int_{I\!R^N}\frac{\partial g(v_t)}{\partial v_{it}}v_{jt}v_{kt}v_{lt}dv_t$ which for $i = j = k = l$ is equal to 3, for $i = j \neq k = l$ is equal to 1, and zero otherwise. Therefore, we have the following symmetry relation,

   $$\int_{I\!R^N}\frac{\partial g(v_t)}{\partial v_{it}}v_{jt}v_{kt}v_{lt}dv_t = \int_{I\!R^N}\frac{\partial g(v_t)}{\partial v_{jt}}v_{it}v_{kt}v_{lt}dv_t,$$

   and, as a consequence, the first term in (53) can be written as $D_N J_t$ where

   $$J_t = -\mathrm{E}\left[\mathrm{vech}\left(\frac{\partial \log g}{v_t}v_t'\right)\mathrm{vech}(v_t v_t')'\right]$$

   is, due to the above elementwise calculations, a $(N^* \times N^*)$ matrix with 3 at the (i,i)th position, $i \in \mathcal{K}_{ii}^N$; a 1 at the positions (i,j), $i,j \in \mathcal{K}_{ii}^N, i \neq j$; a 1 at the positions (i,i), $i \in \mathcal{K}_{ij}^N$; and zeros elsewhere. Thus, using Lemma 4 and 5, $J_t = \mathrm{vech}(I_N)\mathrm{vech}(I_N)' + 2D_N^+ D_N^{+\prime}$. Rearranging and multiplying from the left by D_N we obtain $D_N J_t - \mathrm{vec}(I_N)\mathrm{vech}(I_N)' = 2D_N D_N^+ D_N^{+\prime}$, the stated expression for the right block of $M_{\psi F}$ in (53).

2. $g \in \mathcal{D}_s$:

To derive M_{FF} for spherical distributions, note that v_t is orthogonal to $\text{vech}(v_t v_t' - I_N)$ due to Lemma 6 (moments containing odd orders are zero). Thus, M_{FF} is block-diagonal. The upper left block of M_{FF} is just $\text{E}[v_t v_t'] = I_N$ by Assumption 1. The lower right block of M_{FF} can be written as

$$\text{E}[\text{vech}(v_t v_t')\text{vech}(v_t v_t')'] - \text{vech}(I_N)\text{vech}(I_N)' \tag{54}$$

Using Lemma 6, the (i,i)-th element, $i \in \mathcal{K}_{ii}^N$, of the first term in (54) is equal to $3c$; the (i,j)-th element, $i,j \in \mathcal{K}_{ii}^N$, $i \neq j$, is equal to c; the (i,i)-th element, $i \in \mathcal{K}_{ij}^N$, is equal to c; and all other elements are zero. Together with Lemma 5, this implies that the (i,i)-th element, $i \in \mathcal{K}_{ii}^N$ of (54) is $3c - 1$; the (i,j)-th element, $i,j \in \mathcal{K}_{ii}^N$, $i \neq j$, is $c - 1$; and the (i,i)-th element, $i \in \mathcal{K}_{ij}^N$ is c. The stated formula is then immediately obtained by applying Lemma 4 and Lemma 5.

Next, we derive $M_{\psi\psi}$ for the case $g \in \mathcal{D}_s$. The idea is to show that $D_N^+ M_{\psi\psi} D_N^{+\prime}$ has the same structure as the lower right block of M_{FF}. We have

$$D_N^+ M_{\psi\psi} D_N^{+\prime} = \text{E}\left[\text{vech}\left(\frac{\partial \log g(x)}{\partial x} x'\right) \text{vech}\left(\frac{\partial \log g(x)}{\partial x} x'\right)'\right] - \text{vech}(I_N)\text{vech}(I_N)'.$$

A typical element of the first term can be written as

$$\text{E}\left[\frac{\partial \log g(x)}{\partial x_i} \frac{\partial \log g(x)}{\partial x_j} x_k x_l\right] = 4\text{E}\left[f^{-2}(x'x)\left(\frac{\partial f(x'x)}{\partial x'x}\right)^2 x_i x_j x_k x_l\right]$$

because $g(x) = f(x'x)$, which is equal to

$$4\int_{I\!R^N} f^{-1}(x'x)\left(\frac{\partial f(x'x)}{\partial x'x}\right)^2 x_i x_j x_k x_l dx \tag{55}$$

Now the function $h(x'x) = 4f^{-1}(x'x)\left(\frac{\partial f(x'x)}{\partial x'x}\right)^2$ depends on x only through $x'x$, is positive and integrable by Assumption 1. Thus, it is itself a spherical density up to some scale and (55) is just the fourth order moment structure with respect to h. Therefore, Lemma 6 applies to h and we obtain the same structure as for the lower right block of M_{FF}. That is, for $i \neq j$,

$$\text{E}\left[\left(\frac{\partial \log g(x)}{\partial x_i}\right)^2 x_i^2\right] = 3\text{E}\left[\left(\frac{\partial \log g(x)}{\partial x_i}\right)^2 x_j^2\right]$$

and

$$\text{E}\left[\frac{\partial \log g(x)}{\partial x_i} \frac{\partial \log g(x)}{\partial x_j} x_i x_j\right] = \text{E}\left[\left(\frac{\partial \log g(x)}{\partial x_i}\right)^2 x_j^2\right]. \tag{56}$$

Thus, $D_N^+ M_{\psi\psi} D_N^{+\prime} = 2\tau D_N^+ D_N^{+\prime} + (\tau - 1)\text{vech}(I_N)\text{vech}(I_N)'$ and, because $D_N D_N^+$ is symmetric and idempotent, we obtain $M_{\psi\psi} = 2\tau D_N D_N^+ + (\tau - 1)\text{vec}(I_N)\text{vec}(I_N)'$.

3. $g \in \mathcal{D}_i$:

Due to the independence and symmetry, M_{FF} is diagonal where the first N diagonal elements are 1 because $\mathrm{E}[v_{it}^2] = 1$, $i = 1, \ldots, N$ by Assumption 1. The next N^* diagonal elements are either equal to $\kappa - 1$, corresponding to $\mathrm{E}[v_{it}^4] - 1$, or equal to 1, corresponding to $\mathrm{E}[v_{it}^2 v_{jt}^2]$, $i \neq j$. The formula given in the statement is easily seen to fulfill this requirement.

Finally, the given formula for $M_{\psi\psi}$ can easily be checked by noting that a typical element can be written as

$$\mathrm{E}\left[\frac{\partial \log g(x)}{\partial x_i}\frac{\partial \log g(x)}{\partial x_j} x_k x_l\right] - I\{(i = k) \wedge (j = l)\},$$

where $I(\cdot)$ denotes the indicator function. For $i = j = k = l$, this is equal to $\tau_1 - 1$ by definition. For $i = k \neq j = l$, we have

$$\mathrm{E}\left[\frac{\partial \log g(x)}{\partial x_i} x_i\right] \mathrm{E}\left[\frac{\partial \log g(x)}{\partial x_j} x_j\right] - 1 = 1 - 1 = 0$$

and for $i = j \neq k = l$, due to independence and $\mathrm{E}[x_j^2] = 1$,

$$\mathrm{E}\left[\left(\frac{\partial \log g(x)}{\partial x_i}\right)^2 x_j^2\right] = \mathrm{E}\left[\left(\frac{\partial \log g(x)}{\partial x_i}\right)^2\right],$$

which is equal to τ_2 by definition. All off-diagonal elements of $M_{\psi\psi}$ are zero and the provided formula is easily seen to hold. Q.E.D.

Proof of Proposition 5 Full parametric efficiency is possible if and only if $P_t = 0$. To prove that this occurs only for the multinormal distribution in the class of spherical distributions, consider first the case $N = 2$. Then P_t in (8) is a vector with four components, the second and third of which are equivalent due to the symmetry. Writing the reduced equation system $D_N^+ P_t = 0$ elementwise, the second equation becomes

$$-\frac{\partial \log g(x)}{\partial x_1} x_2 - \frac{1}{c} x_1 x_2 = 0.$$

Using the symmetry of spherical distributions, this yields

$$\frac{\partial g(x)}{\partial x} = -\frac{1}{c} g(x) x \tag{57}$$

whose unique solution is given by $g(x) = const \exp(-\frac{1}{2c}x'x)$, which is the multinormal distribution with covariance matrix cI_N. But since we restricted g to have identity covariance matrix, $c = 1$.

To prove the statement for any dimension N, we have to analyze the structure of the matrix $M_{\psi F}M_{FF}^{-1}$. By Lemma 4, the right block of $D_N^+ M_{\psi F}$ is an $(N^* \times N^*)$ diagonal matrix with 2 at the (i,i)-th position, $i \in \mathcal{K}_{ii}^N$, and 1 at the (j,j)-th position, $j \in \mathcal{K}_{ij}^N$.

Lemma 4 and 5 imply that the lower right block of M_{FF} is an $(N^* \times N^*)$ matrix with $3c-1$ at the (i,i)-th position, $i \in \mathcal{K}_{ii}^N$; $c-1$ at the (i,j)-th position, $i,j \in \mathcal{K}_{ii}^N$, $i \neq j$; c at the (i,i)-th position, $i \in \mathcal{K}_{ij}^N$; and zeros elsewhere. Thus, the i-th row and i-th column, $i \in \mathcal{K}_{ij}^N$, of the lower right block of M_{FF} contains a c at the i-th position and zeros elsewhere. Since M_{FF} is block diagonal, M_{FF}^{-1} is block diagonal as well with the lower right block given by the inverse of the lower right block of M_{FF}. Therefore, the i-th row and i-th column, $i \in \mathcal{K}_{ij}^N$, of the lower right block of M_{FF}^{-1} contains a $1/c$ at the i-th position and zeros elsewhere. This proves that the i-th element, $i \in \mathcal{K}_{ij}^N$, of the vector $D_N^+ M_{\psi F} M_{FF}^{-1} F_t$ is equal to $v_{ti}v_{tj}/c$. One then obtains the same differential equation (57) with unique solution the N-variate normal distribution.

Next we show that the multinormal distribution is also the only one in $\mathcal{D}_i$ for which the parametric lower bound can be attained under the additional constraint that $N \geq 2$. As $D_N^+ Q D_N^{+\prime}$ is diagonal in the case $g \in \mathcal{D}_i$, $Q = 0$ if and only if all diagonal elements of $D_N^+ Q D_N^{+\prime}$ are zero. Since there are only two distinct diagonal elements, $Q = 0$ if and only if two conditions hold: $\tau_1 - 1 = 4/(\kappa - 1)$ and $\tau_2 = 1$. For the case $N = 1$, Gonzalez-Rivera (1997) has shown that the bimodal density (26) fulfills the first of these conditions. We now show that for this density with $N \geq 2$ the second condition does not hold. We have $\partial \log g(v_t)/\partial v_{it} = (\lambda - 1)/v_{it} - \lambda v_{it}$, so that

$$\tau_2 = (\lambda - 1)^2 \mathrm{E}[v_{it}^{-2}] + 2\lambda - \lambda^2,$$

where $\lambda > 2$. However, since $\mathrm{E}[v_{it}^{-2}] > \mathrm{E}[v_{it}^2]^{-1} = 1$ by Jensen's inequality, it follows that $\tau_2 > 1$, violating the second condition for $Q = 0$. In particular, for $\lambda = 3$, it is easy to show that $\mathrm{E}[v_{it}^{-2}] = 3$, so that $\tau_2 = 9$ and the maximum eigenvalue of Q is 8, the value given in Table 1. Note that any other distribution violates the first condition, $\tau_1 - 1 = 4/(\kappa - 1)$, except for the multinormal. This completes the proof. Q.E.D.

Proof of Proposition 6

Under Assumption 5, the matrix $\frac{\partial \log g(v_t)}{\partial v_t} v_t'$ is symmetric, so that $\psi_t = -D_N \mathrm{vech}(I_N + \frac{\partial \log g(v_t)}{\partial v_t} v_t')$. Furthermore, if we define $Z = (H_t^{1/2} \otimes I_N + I_N \otimes H_t^{1/2})$ and using Lemma 2 in Appendix B, we can write,

$$\begin{aligned} Z^{-1} D_N D_N^+ (I_N \otimes H_t^{-1/2}) D_N &= \frac{1}{2} Z^{-1} (I_N \otimes H_t^{-1/2} + H_t^{-1/2} \otimes I_N) D_N \\ &= \frac{1}{2} (H_t^{-1/2} \otimes H_t^{-1/2}) D_N, \end{aligned}$$

by noting that

$$(I_N \otimes H_t^{-1/2} + H_t^{-1/2} \otimes I_N) = (H_t^{-1/2} \otimes H_t^{-1/2})(H_t^{1/2} \otimes I_N + I_N \otimes H_t^{1/2}).$$

So we have that $\dot{l}_t(\theta) = W_t(\theta)\psi_t$, with $W_t = \frac{\partial \text{vec}(H_t)'}{\partial \theta}(H_t^{-1/2} \otimes H_t^{-1/2})$. Q.E.D.

Proof of Proposition 7

First note that $S_t \in \mathcal{T}_s$ because $\text{E}[S_t] = 0$, $\text{E}[S_t F_t'] = 0$, and $\text{E}[S_t S_t']$ is finite since $\text{E}[\tilde{\psi}_t \tilde{\psi}_t'] \leq \text{E}[\text{E}[\psi_t \psi_t' \mid w_t]] = \text{E}[\psi_t \psi_t'] < \infty$ by Assumption 2 and $\text{E}[\tilde{F}_t \tilde{F}_t'] \leq \text{E}[\text{E}[F_t F_t' \mid w_t]] = \text{E}[F_t F_t'] < \infty$ by Assumption 1 (finite fourth moments).

Next, we show that the orthogonal complement of the projection is orthogonal to $\mathcal{T}_s$. It can be written as

$$\dot{\ell}_t^*(\theta) = \dot{\ell}_t(\theta) - S_t(\theta) = (W_t - \text{E}[W_t])\tilde{\psi}_t + \text{E}[W_t] M_{\tilde{\psi}F} M_{\tilde{F}F}^{-1} \tilde{F}_t \tag{58}$$

The first term on the right hand side of (58) is orthogonal to $\mathcal{T}_s$ since $(W_t - \text{E}[W_t])$ has mean zero and is independent of w_t and, hence, independent of all elements in $\mathcal{T}_s$. The second term on the right hand side of (58) is orthogonal to $\mathcal{T}_s$ because $\forall s(w_t) \in \mathcal{T}_s$, $\text{E}[\text{E}[F_t \mid w_t] s(w_t)] = \text{E}[\text{E}[F_t s(w_t) \mid w_t]] = E[F_t s(w_t)]$ by the law of iterated expectations. However, by definition, F_t is orthogonal to $\mathcal{T}_s$. This proves that $\dot{\ell}_t^*(\theta)$ is orthogonal to $\mathcal{T}_s$. Q.E.D.

Appendix D: Some matrix algebra and calculus

The main part of the following results come from Lütkepohl (1996), abbreviated L hereafter.

1. For matrices A, B, C, D of appropriate dimension, we have

$$\text{vec}(ABC) = (C' \otimes A)\text{vec}(B) \tag{59}$$

$$(A \otimes C)(B \otimes D) = (AB) \otimes (CD) \tag{60}$$

$$\text{Tr}(ABCD) = \text{vec}(D')'(C' \otimes A)\text{vec}(B) \tag{61}$$

2. Matrix square root: The square root of a symmetric positive definite matrix X is defined as

$$X^{1/2} = \Gamma \Lambda^{1/2} \Gamma' \tag{62}$$

where the columns of Γ contain the eigenvectors of X and $\Lambda^{1/2}$ is diagonal with the positive square roots of the eigenvalues on its diagonal. Note that $X^{1/2}$ is symmetric and positive definite.

3. L 3.5.1 (1), p.27: $X, Y(m \times m)$ nonsingular:

$$(XY)^{-1} = Y^{-1}X^{-1} \tag{63}$$

4. The (Moore-Penrose) generalized inverse of an $(m \times n)$ matrix X can be defined as

$$X^{+} = (X'X)^{-1}X' \tag{64}$$

if $X'X$ is nonsingular.

5. The $(mn \times mn)$ commutation matrix C_{mn} is defined by

$$C_{mn}\text{vec}(A) = \text{vec}(A') \tag{65}$$

for every $(m \times n)$ matrix A. Let E_{ij}^{mn} be the $(m \times n)$ matrix with 1 in its ij-th position and zeros elsewhere. Then an explicit expression for C_{mn} is given by

$$C_{mn} = \sum_{i=1}^{m}\sum_{j=1}^{n}(E_{ij}^{mn} \otimes E_{ij}^{mn\prime}). \tag{66}$$

For example, C_{22} is given by

$$C_{22} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

L 9.2.2 (5b), p.117: $A(m \times n), B(p \times q)$:

$$B \otimes A = C_{pm}(A \otimes B)C_{nq} \tag{67}$$

6. The $(n^2 \times n(n+1)/2)$ duplication matrix D_n is defined so that

$$D_n\text{vech}(A) = \text{vec}(A) \tag{68}$$

for every symmetric matrix A of order n. For example, D_2 is given by

$$D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

An explicit expression is given by

$$D_n = \sum_{j=1}^{n}(\sum_{i>j}^{n}\text{vec}(E_{ij}^{nn} + E_{ji}^{nn})\text{vech}(E_{ij}^{nn})' + \text{vec}(E_{jj}^{nn})\text{vech}(E_{jj}^{nn})').$$

7. L 9.5.2 (1), p.123: The matrix DD^+ is linked to the commutation matrix by

$$D_m D_m^+ = (I_{m^2} + C_{mm})/2 \tag{69}$$

8. L 9.5.2 (2), p.123:

$$C_{mm} D_m = D_m \tag{70}$$

9. L 9.5.4 (1), p.124: $A(m \times m)$:

$$D_m D_m^+ (A \otimes A) = (A \otimes A) D_m D_m^+ \tag{71}$$

10. Theorem 3.11 (iii) Magnus (1988, p.49): $A, B(m \times m)$:

$$\begin{aligned} D_m D_m^+ (A \otimes B + B \otimes A) D_m D_m^+ &= D_m D_m^+ (A \otimes B + B \otimes A) \\ &= (A \otimes B + B \otimes A) D_m D_m^+ \end{aligned} \tag{72}$$

11. L 9.5.4 (8c), p. 125: $A(m \times m)$:

$$\left(D_m^+ \left(I_m \otimes A + A \otimes I_m\right) D_m\right)^{-1} = D_m^+ \left(I_m \otimes A + A \otimes I_m\right)^{-1} D_m \tag{73}$$

12. L, p. 125:

$$D_{m+1}^+ D_{m+1}^{+\prime} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} I_m & 0 \\ 0 & 0 & D_m^+ D_m^{+\prime} \end{bmatrix} \tag{74}$$

13. Chain rule for matrix differentiation, L 10.7(2), p.203: $X(m \times n)$, $Y(X)(p, \times q)$, $Z(Y)(r \times s)$:

$$\frac{\partial \text{vec}(Z(Y(X)))}{\partial \text{vec}(X)'} = \frac{\partial \text{vec}(Z(Y))}{\partial \text{vec}(Y)'} \frac{\partial \text{vec}(Y(X))}{\partial \text{vec}(X)'} \tag{75}$$

14. Magnus (1988, p.129): $X(m \times m)$ symmetric, $Y(X)$ symmetric matrix function: Using (68), the differential of $\text{vec}(Y)$ can be written as

$$\begin{aligned} \text{dvec}(Y) &= D_m \frac{\partial \text{vech}(Y)}{\partial \text{vech}(X)'} \text{dvech}(X) \\ &= D_m \frac{\partial \text{vech}(Y)}{\partial \text{vech}(X)'} D_m^+ \text{dvec}(X) \end{aligned} \tag{76}$$

15. L 10.4 (3), p.183: $X(m \times n)$, $A(p \times m)$, $B(n \times q)$:

$$\frac{\partial \text{vec}(AXB)}{\partial \text{vec}(X)'} = B' \otimes A \tag{77}$$

16. L 10.5.3 (2), p. 194: $X, A(m \times m)$ symmetric:

$$\frac{\partial \text{vech}(XAX)}{\partial \text{vech}(X)'} = D_m^+(XA \otimes I_m + I_m \otimes XA)D_m \tag{78}$$

17. L 10.6 (1), p.198: $X(m \times m)$ nonsigular:

$$\frac{\partial \text{vec}(X^{-1})}{\partial \text{vec}(X)'} = -X'^{-1} \otimes X^{-1} \tag{79}$$

18. L 10.3.3, p.182: $X(m \times m)$, $|X| > 0$:

$$\frac{\partial \log |X|}{\partial X} = (X')^{-1} \tag{80}$$

g	Class	$N = 1$			$N = 2$			$N = 3$		
		κ	τ	$\rho(Q)$	κ	τ	$\rho(Q)$	κ	τ	$\rho(Q)$
$N(0, I_N)$	$\mathcal{D}_s \cap \mathcal{D}_i$	3	1	0	3	1	0	3	1	0
t ($\nu = 5$)	$\mathcal{D}_s$	9.00	0.75	0.75	9.00	0.78	0.89	9.00	0.80	0.93
t ($\nu = 8$)	$\mathcal{D}_s$	4.50	0.82	0.31	4.50	0.83	0.33	4.50	0.85	0.36
t ($\nu = 12$)	$\mathcal{D}_s$	3.75	0.87	0.15	3.75	0.87	0.17	3.75	0.88	0.18
Laplace	$\mathcal{D}_s$	6.00	0.67	0.20	5.00	0.75	0.30	4.50	0.80	0.27
ES Logistic	$\mathcal{D}_s$	2.38	5.10	11.40	2.57	1.44	0.96	2.69	1.30	0.78
Logistic	$\mathcal{D}_i$	4.20	0.81	0.18	4.20	0.81	0.18	4.20	0.81	0.18
(26), $\lambda = 3$	$\mathcal{D}_i$	1.67	2.33	0	1.67	2.33	8	1.67	2.33	8

Table 1: Marginal kurtosis κ, the value of τ in (18), and the spectral norm of the matrix Q in (13) for alternative distributions g and dimension N. For the spherical distributions, the co-kurtosis is $c = \kappa/3$. For those in $\mathcal{D}_i$, the parameter τ_1 is given by 3τ (τ_2 is not reported).

	ML			*QML*			*SP*		
Population	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$c_{11} = 1$	1.021	0.209	0.0441	1.043	0.286	0.0841	1.0121	0.2394	0.0574
$c_{21} = 0.7$	0.725	0.245	0.0609	0.739	0.309	0.0975	0.7227	0.2964	0.0883
$c_{22} = 1$	1.039	0.218	0.0493	1.060	0.285	0.0853	1.0302	0.2328	0.0551
$\beta_{11} = 0.5$	0.492	0.081	0.0066	0.483	0.113	0.0131	0.4961	0.0930	0.0086
$\beta_{22} = 0.1$	0.070	0.269	0.0735	0.060	0.324	0.1066	0.0824	0.3211	0.1034
$\beta_{33} = 0.6$	0.591	0.062	0.0040	0.581	0.084	0.0075	0.5950	0.0690	0.0047
$\alpha_{11} = 0.2$	0.201	0.038	0.0014	0.201	0.055	0.0030	0.1942	0.0400	0.0016
$\alpha_{22} = 0.1$	0.099	0.032	0.0010	0.102	0.043	0.0018	0.0962	0.0347	0.0012
$\alpha_{33} = 0.2$	0.201	0.035	0.0012	0.208	0.052	0.0028	0.1961	0.0378	0.0014

Table 2: Monte Carlo results based on 500 replications of the diagonal VEC model with $n = 2000$. The innovation density is a bivariate t distribution with 5 degrees of freedom. MSE means mean squared error and SD means standard deviation.

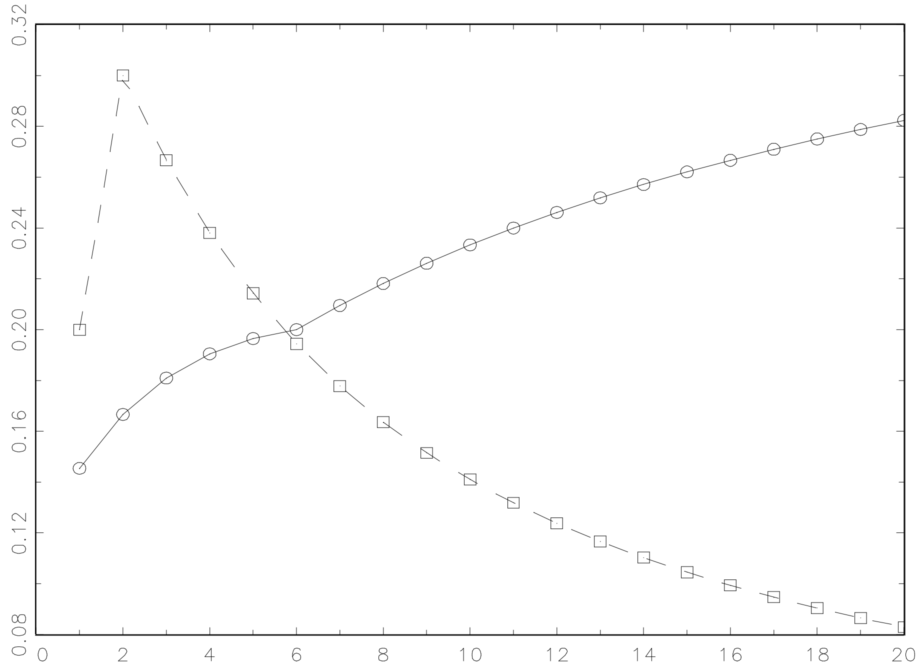

Figure 1: *The spectral norm* $\rho(Q)$ *for the Laplace (dashed) and* t_{12} *distribution (solid), viewed as a function of the dimension* N*.*