NYU STERN

Salomon Center for the Study of Financial Institutions

Working Paper Series **CENTER FOR FINANCIAL ECONOMETRICS**

VECTOR MULTIPLICATIVE ERROR MODELS: REPRESENTATION AND INFERENCE

Fabrizio Cipollini
Robert F. Engle
Giampiero M. Gallo

CFE-06-01

Vector Multiplicative Error Models: Representation and Inference*

Fabrizio Cipollini† Robert F. Engle‡ Giampiero M. Gallo§

This version: October 2006

Abstract

The Multiplicative Error Model introduced by Engle (2002) for positive valued processes is specified as the product of a (conditionally autoregressive) scale factor and an innovation process with positive support. In this paper we propose a multivariate extension of such a model, by taking into consideration the possibility that the vector innovation process be contemporaneously correlated. The estimation procedure is hindered by the lack of probability density functions for multivariate positive valued random variables. We suggest the use of copula functions and of estimating equations to jointly estimate the parameters of the scale factors and of the correlations of the innovation processes. Empirical applications on volatility indicators are used to illustrate the gains over the equation by equation procedure.

*We thank Christian T. Brownlees, Marco J. Lombardi and Margherita Velucchi for many discussions on MEMs and multivariate extensions, as well as participants in seminars at CORE and IGIER–Bocconi for helpful comments. The usual disclaimer applies.

†Department of Statistics "G. Parenti", University of Firenze, Italy. e-mail: cipollin@ds.unifi.it

‡Department of Finance, Stern School of Business, New York University; email rengle@stern.nyu.edu

§Corresponding author: Department of Statistics "G. Parenti", University of Firenze, Viale G.B. Morgagni, 59 - 50134 Firenze – Italy. e-mail: gallog@ds.unifi.it

Contents

1 Introduction

The study of financial market behavior is increasingly based on the analysis of the dynamics of nonnegative valued processes, such as exchanged volume, high–low range, absolute returns, financial durations, number of trades, and so on. Generalizing the GARCH (Bollerslev (1986)) and ACD (Engle and Russell (1998)) approaches, Engle (2002) reckons that one striking regularity of financial time series is that persistence and clustering characterizes the evolution of such processes. As a result, the process describing the dynamics of such variables can be specified as the product of a conditionally deterministic scale factor which evolves according to a GARCH–type equation and an innovation term which is i.i.d. with unit mean. Empirical results (e.g. Chou (2005); Manganelli (2002)) show a good performance of these types of models in capturing the stylized facts of the observed series.

More recently, Engle and Gallo (2006) have investigated three different indicators of volatility, namely absolute returns, daily range and realized volatility in a multivariate context in which each lagged indicator was allowed to enter the equation of the scale factor of the other indicators. Model selection techniques were adopted to ascertain the statistical relevance of such variables in explaining the dynamic behavior of each indicator. The model was estimated assuming a diagonal variance covariance matrix of the innovation terms. Further applications along the same lines may be found in Gallo and Velucchi (2005) for different measures of volatility based on 5-minute returns, Brownlees and Gallo (2005) for hourly returns, volumes and number of trades, Engle *et al.* (2005) for volatility transmission across financial markets.

Estimation equation by equation ensures consistency of the estimators in a quasi-maximum likelihood context, given the stationarity conditions discussed by Engle (2002). This simple procedure is obviously not efficient, since correlation among the innovation terms is not taken into account: in several cases, especially when predetermined variables are inserted in the specification of the conditional expectation of the variables, it would be advisable to work with estimators with better statistical properties, since model selection and ensuing interpretation of the specification is crucial in the analysis.

In this paper we investigate the problems connected to a multivariate specification and estimation of the MEM. Since joint probability distributions for nonnegative–valued random variables are not available except in very special cases, we resort to two different strategies in order to manage vector-MEM: the first is to adopt copula functions to link together marginal probability density functions specified as Gamma as in Engle and Gallo (2006); the second is to adopt an Estimating Equation approach (Heyde (1997), Bibby *et al.* (2004)). The empirical applications performed on the General Electric stock data show that there are some numerical differences in the estimates obtained by the three methods: copula and estimating equations results are fairly similar to one another while the equation–by–equation procedure provides estimates which depart from the system–based ones the more correlated the variables inserted in the analysis.

2 The Univariate MEM Reconsidered

Let us start by recalling the main features of the univariate Multiplicative Error Model (MEM) introduced by Engle (2002) and extended by Engle and Gallo (2006). For ease of reference, some characteristics and properties of the statistical distributions involved are detailed in appendix A.

2.1 Definition and formulations

Let us consider x_t, a non–negative univariate process, and let $\mathcal{F}_{t-1}$ be the information about the process up to time $t-1$. Then the MEM for x_t is specified as

$$x_t = \mu_t \varepsilon_t, \tag{1}$$

where, conditionally on the information $\mathcal{F}_{t-1}$: μ_t is a nonnegative *conditionally deterministic* (or *predictable*) process, that is the evolution of which depends on a vector of unknown parameters $\boldsymbol{\theta}$,

$$\mu_t = \mu_t(\boldsymbol{\theta}); \tag{2}$$

ε_t is a *conditionally stochastic* i.i.d. process, with density having non–negative support, mean 1 and unknown variance σ^2,

$$\varepsilon_t | \mathcal{F}_{t-1} \sim D(1, \sigma^2). \tag{3}$$

The previous conditions on μ_t and ε_t guarantee that

$$E(x_t | \mathcal{F}_{t-1}) = \mu_t \tag{4}$$
$$V(x_t | \mathcal{F}_{t-1}) = \sigma^2 \mu_t^2. \tag{5}$$

To close the model, we need to adopt a parametric density function for ε_t and to specify an equation for μ_t.

For the former step, we follow Engle and Gallo (2006) in adopting a Gamma distribution which has the usual exponential density function (as in the original Engle and Russell (1998) ACD model) as a special case. We have

$$\varepsilon_t | \mathcal{F}_{t-1} \sim Gamma(\phi, \phi), \tag{6}$$

with $E(\varepsilon_t | \mathcal{F}_{t-1}) = 1$ and $V(\varepsilon_t | \mathcal{F}_{t-1}) = 1/\phi$. Taken jointly, assumptions (1) and (6) can be written compactly as (see appendix A)

$$x_t | \mathcal{F}_{t-1} \sim Gamma(\phi, \phi/\mu_t). \tag{7}$$

Note that a well-known relationship (cf. Appendix A) between the Gamma distribution and the Generalized Error Distribution (GED) suggests an equivalent formulation which

may prove useful for estimation purposes. We have:

$$x_t|\mathcal{F}_{t-1} \sim Gamma(\phi, \phi/\mu_t) \Leftrightarrow x_t^\phi|\mathcal{F}_{t-1} \sim Half - GED(0, \mu_t^\phi, \phi). \tag{8}$$

Thus the conditional density of x_t has a correspondence in a conditional density of x_t^ϕ, that is

$$x_t^\phi = \mu_t^\phi \nu_t \tag{9}$$

where

$$\nu_t|\mathcal{F}_{t-1} \sim Half - GED(0, 1, \phi). \tag{10}$$

This result somewhat generalizes the practice, suggested by Engle and Russell (1998), to estimate an ACD model by estimating the parameters of the second moment of the square root of the durations (imposing mean zero) with a GARCH routine assuming normality of the errors. The result extends to any estimation carried out on observations x_t^ϕ (with ϕ known)[1] assuming a GED distribution with mean zero and dispersion parameter μ_t^ϕ.

As per the specification of μ_t, following, again, Engle (2002) and Engle and Gallo (2006), let us consider the simplest GARCH–type of order (1,1) formulation, and, for the time being, let us not include predetermined variables in the specification.

The *base* $(1, 1)$ specification of μ_t is

$$\mu_t = \omega + \alpha x_{t-1} + \beta \mu_{t-1}, \tag{11}$$

which is appropriate when, say, x_t is a series of positive valued variables such as financial durations or traded volumes. An extended specification is appropriate when one can refer to the sign of financial returns as relevant extra information: well-known empirical evidence suggests a symmetric distribution for the returns but an asymmetric response of conditional variance to negative innovations (so–called leverage effect). Thus, when considering volatility related variables (absolute or squared returns, realized volatility, range, etc.), asymmetric effects can be inserted in the form:

$$\mu_t = \omega^* + \alpha(x_{t-1}^{1/2} \operatorname{sign}(r_{t-1}) + \delta^*)^2 + \gamma x_{t-1} \operatorname{I}(r_{t-1} < 0) + \beta \mu_{t-1}, \tag{12}$$

where δ^* and γ are parameters that capture the asymmetry. Following Engle and Gallo (2006), we have inserted in the model two variants of existing specifications for including asymmetric effects in the GARCH literature: an APARCH-like asymmetry effect (the squared term with δ^* – see Ding *et al.* (1993)) and a GJR-like asymmetry effect (the last term with γ – see Glosten *et al.* (1993)). Computing the square, expression (12) can be rewritten as

$$\mu_t = \omega + \alpha x_{t-1} + \gamma x_{t-1}^{(-)} + \delta x_{t-1}^{(s)} + \beta \mu_{t-1}, \tag{13}$$

where $x_t^{(s)} = x_t^{1/2} \operatorname{sign}(r_t)$, $x_t^{(-)} = x_t \operatorname{I}(r_t < 0)$, $\omega = \omega^* + \alpha\delta^{*2}$ and $\delta = 2\alpha\delta^*$.

[1] In raising the observations x_t to the ϕ makes the ensuing model more similar to the Ding *et al.* (1993) APARCH specification except that the exponent parameter appears also in the error distribution.

Both specifications can be written compactly as

$$\mu_t = \omega + \boldsymbol{x}_{t-1}^{*\prime}\boldsymbol{\alpha}^* + \beta\mu_{t-1}. \tag{14}$$

where, if we assume $\boldsymbol{x}_t^* = (x_t; x_t^{(-)}; x_t^{(s)})$ and $\boldsymbol{\alpha}^* = (\alpha; \gamma; \delta)$ we get the asymmetric specification (13)[2]. The base specification (11) can be retrieved more simply taking $\boldsymbol{x}_t^* = x_t$ and $\boldsymbol{\alpha}^* = \alpha$.

The parameter space for $\boldsymbol{\theta} = (\omega; \boldsymbol{\alpha}^*; \beta)$ must be restricted in order to ensure that $\mu_t \geq 0$ for all t and to ensure stationary distributions for x_t. However restrictions depend on the formulation taken into account: we consider here formulation (13). Sufficient conditions for stationarity can been obtained taking the unconditional expectation of both members of (14) and solving for $E(x_t) = \mu$. This implies x_t stationary if

$$\alpha + \beta + \gamma/2 < 1.$$

Sufficient conditions for non–negativity of μ_t can be obtained taking $\beta \geq 0$ and imposing $\omega + \alpha x_t + \gamma x_t \,\mathrm{I}(r_t < 0) + \delta x_t^{1/2}\,\mathrm{sign}(r_t) \geq 0$ for all x_t's and r_t's. Using proposition 1, one can verify that these conditions are satisfied when

$$\begin{gathered} \beta \geq 0, \qquad \alpha \geq 0, \qquad \alpha + \gamma \geq 0, \\ \text{if } \alpha = 0 \text{ then } \delta \geq 0; \qquad \text{if } \alpha + \delta > 0 \text{ then } \delta \leq 0, \\ \omega - \frac{\delta^2}{4}\left[\frac{\mathrm{I}(\delta < 0)\,\mathrm{I}(\alpha > 0)}{\alpha} + \frac{\mathrm{I}(\delta > 0)\,\mathrm{I}(\alpha + \gamma > 0)}{\alpha + \gamma}\right] \geq 0. \end{gathered} \tag{15}$$

2.2 Estimation and Inference

Let us introduce estimation and inference issues by discussing first the role of a generic observation x_t. From (7), the contribution of x_t to the log–likelihood function is

$$l_t = \ln L_t = \phi \ln \phi - \ln \Gamma(\phi) + (\phi - 1)\ln x_t - \phi(\ln \mu_t + x_t/\mu_t).$$

The contribution of x_t to the score is $\mathbf{s}_t = \begin{pmatrix} \mathbf{s}_{t,\boldsymbol{\theta}} \\ s_{t,\phi} \end{pmatrix}$ with components

$$\begin{aligned} \mathbf{s}_{t,\boldsymbol{\theta}} &= \nabla_{\boldsymbol{\theta}} l_t = \phi \nabla_{\boldsymbol{\theta}} \mu_t \left(\frac{x_t - \mu_t}{\mu_t^2}\right) \\ s_{t,\phi} &= \nabla_{\phi} l_t = \ln \phi + 1 - \psi(\phi) + \ln\left(\frac{x_t}{\mu_t}\right) - \frac{x_t}{\mu_t}, \end{aligned}$$

where $\psi(\phi) = \dfrac{\Gamma'(\phi)}{\Gamma(\phi)}$ is the *digamma* function and the operator $\nabla_{\boldsymbol{\lambda}}$ denotes the derivatives

[2]The following notation applies: by $(\mathbf{x}_1; \ldots; \mathbf{x}_n)$ we mean that the matrices $\mathbf{x}_i$, all with the same number of columns, are stacked vertically; by $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ we indicate that the matrices $\mathbf{x}_i$, all with the same number of rows are stacked horizontally

with respect to (the components of) $\boldsymbol{\lambda}$.

The contribution of x_t to the Hessian is $\mathbf{H}_t = \begin{pmatrix} \mathbf{H}_{t,\boldsymbol{\theta}\boldsymbol{\theta}'} & \mathbf{H}_{t,\boldsymbol{\theta}\phi} \\ \mathbf{H}'_{t,\boldsymbol{\theta}\phi} & \mathbf{H}_{t,\phi\phi} \end{pmatrix}$ with components

$$\begin{aligned}
\mathbf{H}_{t,\boldsymbol{\theta}\boldsymbol{\theta}'} &= \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l_t = \phi \left(\frac{-2x_t + \mu_t}{\mu_t^3} \nabla_{\boldsymbol{\theta}} \mu_t \nabla_{\boldsymbol{\theta}'} \mu_t + \frac{x_t - \mu_t}{\mu_t^2} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \mu_t \right) \\
\mathbf{H}_{t,\boldsymbol{\theta}\phi} &= \nabla_{\boldsymbol{\theta}\phi} l_t = \frac{x_t - \mu_t}{\mu_t^2} \nabla_{\boldsymbol{\theta}} \mu_t \\
H_{t,\phi\phi} &= \nabla_{\phi\phi} l_t = \frac{1}{\phi} - \psi'(\phi),
\end{aligned}$$

where $\psi'(\phi)$ is the *trigamma* function.

Exploiting the previous results we obtain the following first order conditions for $\boldsymbol{\theta}$ and ϕ:

$$\frac{1}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \mu_t \frac{x_t - \mu_t}{\mu_t^2} = 0 \tag{17}$$

$$\ln \phi + 1 - \psi(\phi) + \frac{1}{T} \sum_{t=1}^{T} \left[\ln \left(\frac{x_t}{\mu_t} \right) - \frac{x_t}{\mu_t} \right] = 0 \tag{18}$$

As noted by Engle and Gallo (2006), first–order conditions for $\boldsymbol{\theta}$ do not depend on ϕ. As a consequence, whatever value ϕ may take, any Gamma-based MEM or any equivalent GED-based power formulation will provide the same point estimates for $\boldsymbol{\theta}$. The ML estimation of ϕ can then be performed after $\boldsymbol{\theta}$ has been estimated.

Furthermore, so long as $\mu_t = E(x_t|\mathcal{F}_{t-1})$, the expected value of the score of $\boldsymbol{\theta}$ evaluated at the true parameters is a vector of zeroes even if the density of $\varepsilon_t|\mathcal{F}_{t-1}$ does not belong to the $Gamma(\phi, \phi)$ family.

Altogether, these considerations strengthen the claim (e.g. Engle (2002) for the case $\phi = 1$) that, whatever the value of ϕ, the log–likelihood functions of any one of these formulations (MEM or equivalent power formulation) can be interpreted as Quasi Likelihood functions and the corresponding estimator $\widehat{\boldsymbol{\theta}}$ is a QML estimator.

After some computations we obtain the asymptotic variance–covariance matrix of the ML estimator

$$V_\infty = \begin{pmatrix} \phi \lim_{t\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\mu_t^2} \nabla_{\boldsymbol{\theta}} \mu_t \nabla_{\boldsymbol{\theta}'} \mu_t & \mathbf{0} \\ \mathbf{0} & \psi'(\phi) - \frac{1}{\phi} \end{pmatrix}^{-1}.$$

Note that even if ϕ is not involved in the estimate of $\boldsymbol{\theta}$ the variance of $\widehat{\boldsymbol{\theta}}$ is proportional to $1/\phi$. We note also that the ML estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\phi}$ are asymptotically uncorrelated.

Alternative estimators of ϕ are available, which exploit the orthogonal nature of the pa-

rameters $\boldsymbol{\theta}$ and ϕ. As an example, by defining $u_t = x_t/\mu_t - 1$, we note that[3] $V(u_t|\mathcal{F}_{t-1}) = \phi^{-1}$. Therefore, a simple method of moment estimator is

$$\widehat{\phi^{-1}} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t^2. \tag{19}$$

In view of (18), this expression has the advantage of not being affected by the presence of x_t's equal to zero (cf. also the comments below).

Obviously, the inference about $(\boldsymbol{\theta}, \phi)$ must be based on estimates of V_∞, that can be obtained evaluating the average Hessian or the average outer product of the gradients evaluated at the estimates $(\widehat{\boldsymbol{\theta}}, \widehat{\phi})$. The *sandwich* estimator

$$\widehat{V}_\infty = \widehat{\overline{\mathbf{H}}}_T^{-1} \widehat{\overline{\mathbf{OPG}}}_T \widehat{\overline{\mathbf{H}}}_T^{-1},$$

where $\widehat{\overline{\mathbf{H}}}_T = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{H}}_t$ and $\widehat{\overline{\mathbf{OPG}}}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{s}}_t \hat{\mathbf{s}}_t'$, get rid of the dependence of the submatrix relative to $\boldsymbol{\theta}$ on ϕ altogether.

2.3 Some Further Comments

2.3.1 MEM as member of a more general family of models

The MEM defined in section 2.1 belongs to the Generalized Linear Autoregressive Moving Average (GLARMA) family of models introduced by Shephard (1995) and extended by Benjamin *et al.* (2003)[4]. In the GLARMA model, the conditional density of x_t is assumed to belong to the same exponential family, with density given by[5]

$$f(x_t|\mathcal{F}_{t-1}) = \exp\left[\phi\left(x_t\vartheta_t - b(\vartheta_t)\right) + d(x_t, \phi)\right]. \tag{20}$$

ϑ_t and ϕ are the canonical and precision parameters respectively, whereas $b(\cdot)$ and $d(\cdot)$ are specific functions that define the particular distribution into the family. The main moments are

$$\begin{aligned} E(x_t|\mathcal{F}_{t-1}) &= b'(\vartheta_t) = \mu_t \\ V(x_t|\mathcal{F}_{t-1}) &= b''(\vartheta_t)/\phi = v(\mu_t)/\phi. \end{aligned}$$

Since μ_t may have a bounded domain, it is useful to define its (p, q) dynamics through a twice differentiable and monotonic *link function* $g(\cdot)$, introduced in general terms by

[3] We thank Neil Shephard for pointing this out to us.

[4] For other works related to GLARMA models see, among others, Li (1994), Davis *et al.* (2002).

[5] We adapt the formulation of Benjamin *et al.* (2003) to this work.

Benjamin *et al.* (2003) as

$$g(\mu_t) = \mathbf{z}_t'\boldsymbol{\gamma} + \sum_{i=1}^{p} \alpha_i \mathcal{A}(x_{t-i}, \mathbf{z}_{t-i}, \boldsymbol{\gamma}) + \sum_{j=1}^{q} \beta_j \mathcal{M}(x_{t-j}, \mu_{t-j}),$$

where $\mathcal{A}(\cdot)$ and $\mathcal{M}(\cdot)$ are functions representing the autoregressive and moving average terms. Such a formulation, admittedly too general for practical purposes, can be replaced by a more *practical version*:

$$g(\mu_t) = \mathbf{z}_t'\boldsymbol{\gamma} + \sum_{i=1}^{p} \alpha_i \left[g(x_{t-i}) - \mathbf{z}_{t-i}'\boldsymbol{\gamma}\right] + \sum_{j=1}^{q} \beta_j \left[g(x_{t-j}) - g(\mu_{t-j})\right]. \qquad (21)$$

In this formulation, for certain functions g it may be necessary to replace some x_{t-i}'s by x^*_{t-i} to avoid the nonexistence of $g(x_{t-i})$ for certain values of the argument (e.g. zeros, see Benjamin *et al.* (2003) for some examples).

Specifying (20) as a $Gamma(\phi, \phi/\mu_t)$ density (in this case $v(\mu_t) = \mu_t^2$) and $g(\cdot)$ as the identity function, after some simple arrangements of (21) it is easily verified that the MEM is a particular GLARMA model.

2.3.2 Exact zero values in the data

In introducing the MEM for modelling non–negative time series, Engle (2002) states that the structure of the model avoids problems caused by exact zeros in $\{x_t\}$, problems that, instead, are typical of $\log(x_t)$ formulations. This needs to be further clarified here.

Considering the MEM as defined by formula (1), in principle an $x_t = 0$ can be caused by $\varepsilon_t = 0$, by $\mu_t = 0$ or by both. We discuss these possibilities in turn.

- As the process μ_t is supposed deterministic conditionally to the information $\mathcal{F}_{t-1}$ (section 2.1), the value of μ_t cannot be 'altered' from any of the observations after time $t-1$: in practice it is not possible to take $\mu_t = 0$ if $x_t = 0$. Hence: or μ_t is really 0 (very unlikely!) or an $x_t = 0$ cannot be a consequence of $\mu_t = 0$. As a second observation, a possible $\mu_t = 0$ causes serious problems to the inference, as evidenced by the first–order conditions (17).

- The only possibility is then $\varepsilon_t = 0$. However, as we assumed $\varepsilon_t|\mathcal{F}_{t-1} \sim Gamma(\phi, \phi)$, this distribution cannot be defined for 0 values when $\phi < 1$. Furthermore, even restricting the parameter space to $\phi \geq 1$, the first order condition for ϕ in (18) requires the computation of $\ln(x_t)$, not defined for exact zero values. In practice the ML estimation of ϕ is not feasible in presence of zeros which enforces the usefulness of the method of moments estimator introduced above.

The formulation of the MEM considered by Engle (2002) avoid problems caused by exact zeros because it assume $\varepsilon_t|\mathcal{F}_{t-1} \sim Exponential(1)$, the density of which, $f(\varepsilon_t|\mathcal{F}_{t-1}) =$

$e^{-\varepsilon_t}$, can be defined for $\varepsilon_t \geq 0$. Obviously, any MEM with a *fixed* $\phi \geq 1$ shares this characteristic. We guess that, despite the QML nature of the $\widehat{\boldsymbol{\theta}}$ estimator in section 2.2, when there are many zeros in the data, more appropriate inference can be obtained if we structure the conditional distribution of ε_t as a mixture between a discrete component (at 0) and an absolutely continuous component (on $(0, \infty)$).

3 The vector MEM

As introduced in Engle (2002), a vector MEM is a simple generalization of the univariate MEM. It is a suitable representation of non–negative valued processes which have dynamic interactions with one another. The application in Engle and Gallo (2006) envisages three indicators of volatility (absolute returns, daily range and realized volatility), while the study in Engle *et al.* (2005) takes daily absolute returns on seven East Asian markets to study contagion. In all cases, the model is estimated equation by equation.

3.1 Definition and formulations

Let $\mathbf{x}_t$ a K–dimensional process with non–negative components;[6] a vector MEM for $\mathbf{x}_t$ is defined as

$$\mathbf{x}_t = \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t = \operatorname{diag}(\boldsymbol{\mu}_t)\boldsymbol{\varepsilon}_t, \tag{22}$$

where $\odot$ indicates the Hadamard (element–by–element) product. Conditional on the information $\mathcal{F}_{t-1}$, $\boldsymbol{\mu}_t$ is defined as in (2) except that now we are dealing with a $K-$ dimensional vector depending on a (larger) vector of parameters $\boldsymbol{\theta}$. The innovation vector $\boldsymbol{\varepsilon}_t$ is a *conditionally stochastic* K–dimensional i.i.d. process. Its density function is defined over a $[0, +\infty)^K$ support, with unit vector $\mathbb{1}$ as expectation and a general variance–covariance matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1} \sim D(\mathbb{1}, \boldsymbol{\Sigma}). \tag{23}$$

The previous conditions guarantee that

$$E(\mathbf{x}_t | \mathcal{F}_{t-1}) = \boldsymbol{\mu}_t \tag{24}$$

$$V(\mathbf{x}_t | \mathcal{F}_{t-1}) = \boldsymbol{\mu}_t \boldsymbol{\mu}_t' \odot \boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\mu}_t)\boldsymbol{\Sigma}\operatorname{diag}(\boldsymbol{\mu}_t), \tag{25}$$

which is a positive definite matrix by construction, as emphasized by Engle (2002).

3.2 Specifications for $\boldsymbol{\varepsilon}_t$

In this section we consider some alternatives about the specification of the distribution of the error term $\boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1}$ of the vector MEM defined above. The natural extension to

[6] In what follows we will adopt the convention that if $\mathbf{x}$ is a vector or a matrix and a is a scalar, then the expressions $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x}^a$ are meant element by element.

be considered is to limit ourselves to the assumption that $\boldsymbol{\varepsilon}_{i,t}|\mathcal{F}_{t-1} \sim Gamma(\phi_i, \phi_i)$, $i = 1, \ldots, K$.

3.2.1 Multivariate Gamma distributions

Johnson *et al.* (2000, chapter 48) describe a number of multivariate generalizations of the univariate Gamma distribution. However, many of them are bivariate versions, not sufficiently general for our purposes. Among the others, we do not consider here distributions defined via the joint characteristic function, as they require numerical inversion formulas to find their pdf's. Hence, the only useful versions remain the multivariate Gammas of Cheriyan and Ramabhadran (in their more general version, henceforth $GammaCR$, see appendix C.1 for the details), of Kowalckzyk and Trycha and of Mathai and Moschopoulos (Johnson *et al.* (2000, 454–470)). If one wants the domain of $\boldsymbol{\varepsilon}_t$ to be defined on $[0, +\infty)^K$, it can be shown that the three mentioned versions are perfectly equivalent.[7] As a consequence, we will consider the following multivariate Gamma assumption for the conditional distribution of the innovation term $\boldsymbol{\varepsilon}_t$

$$\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1} \sim GammaCR(\phi_0, \boldsymbol{\phi}, \boldsymbol{\phi}),$$

where $\boldsymbol{\phi} = (\phi_1; \ldots; \phi_K)$ and $0 < \phi_0 < \min(\phi_1, \ldots, \phi_K)$. As described in the appendix, all univariate marginal probability functions for $\varepsilon_{i,t}$ are, as required, $Gamma(\phi_i, \phi_i)$, even if the multivariate pdf is expressed in terms of a complicated integral. The conditional variance matrix of $\boldsymbol{\varepsilon}_t$ has elements

$$C(\varepsilon_{i,t}, \varepsilon_{j,t}|\mathcal{F}_{t-1}) = \frac{\phi_0}{\phi_i \phi_j} \tag{26}$$

so that the correlations are

$$\rho(\varepsilon_{i,t}, \varepsilon_{j,t}|\mathcal{F}_{t-1}) = \frac{\phi_0}{\sqrt{\phi_i \phi_j}}. \tag{27}$$

This implies that the $GammaCR$ distribution admits only positive correlation among its components and that the correlation between each couple of elements is strictly linked to the corresponding variances $1/\phi_i$ and $1/\phi_j$. These various drawbacks (the restrictions on the correlation, the very complicated pdf and the constraint $\phi_0 < \min(\phi_1, \ldots, \phi_K)$), suggest to investigate better alternatives.

3.2.2 The Normal copula-Gamma marginals distribution

A different way to define the distribution of $\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1}$ is to start from the assumption that all univariate marginal probability density functions are $Gamma(\phi_i, \phi_i)$, $i = 1, \ldots, K$ and to use *copula functions*.[8] Adopting copulas, the definition of the distribution of a multi-

[7] The proof is tedious and is available upon request.

[8] The main characteristics of copulas are summarized in appendix B.

variate r.v. is completed by defining the copula that represents the structure of dependence among the univariate marginal pdf's. Many copula functions have been proposed in theoretical and applied works (see, among others, Embrechts *et al.* (2001) and Bouyé *et al.* (2000)). In particular, the Normal copula possesses many interesting properties: the capability of capture a broad range of dependencies[9], the analytical tractability, the easy of simulation. For simplicity, we will assume

$$\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1} \sim N(\mathbf{R}) - \prod_{i=1}^{K} Gamma(\phi_i, \phi_i). \tag{28}$$

where in the distribution, the first part refers to the copula ($\mathbf{R}$ is a correlation matrix) and the second part to the marginals. As detailed in appendix C.2, this distribution is a special case of multivariate dispersion distributions generated from a Gaussian copula discussed in Song (2000). The conditional variance–covariance matrix of $\boldsymbol{\varepsilon}_t$ has a generic element which is approximately equal to

$$C(\varepsilon_{i,t}, \varepsilon_{j,t}|\mathcal{F}_{t-1}) \simeq \frac{\mathbf{R}_{ij}}{\sqrt{\phi_i \phi_j}}.$$

so that the correlations are, approximately,

$$\rho(\varepsilon_{i,t}, \varepsilon_{j,t}|\mathcal{F}_{t-1}) \simeq \mathbf{R}_{ij}.$$

The advantages of using copulas over a multivariate Gamma specification are apparent and suggest their adoption in this context: the covariance and correlation structures are more flexible (also negative correlations are permitted); the correlations do not depend on the variances of the marginals; there are no complicated constraints on the parameters; the pdf is more easily tractable. Furthermore, by adopting copula functions, the framework considered here can be easily extended to different choices of the distribution of the marginal pdf's (Weibull for instance).

3.3 Specification for $\boldsymbol{\mu}_t$

The base $(1, 1)$ specification for $\boldsymbol{\mu}_t$ can be written as

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\alpha}\mathbf{x}_{t-1} + \boldsymbol{\beta}\boldsymbol{\mu}_{t-1}, \tag{29}$$

where $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have dimensions, respectively, $(K, 1)$, (K, K) and (K, K).

The enlarged $(1, 1)$ multivariate specification will include lagged cross-effects and may include asymmetric effects, defined as before. For example, when the first component of $\mathbf{x}_t$ is $x_{1,t} = r_t^2$ and the conditional distribution of $r_t = x_{1,t}^{1/2} \operatorname{sign}(r_t)$ is symmetric with

[9] The bivariate Normal copula, according to the value of the correlation parameter is capable of attaining the lower Frechèt bound, the product copula and the upper Frechèt bound.

zero mean, generalizing formulation (13) we have

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\alpha}\mathbf{x}_{t-1} + \boldsymbol{\gamma}\mathbf{x}_{t-1}^{(-)} + \boldsymbol{\delta}\mathbf{x}_{t-1}^{(s)} + \boldsymbol{\beta}\boldsymbol{\mu}_{t-1}, \tag{30}$$

where $\mathbf{x}_t^{(-)} = \mathbf{x}_t \, \mathrm{I}(r_t < 0)$, $\mathbf{x}_t^{(s)} = \mathbf{x}_t^{1/2} \operatorname{sign}(r_t)$. Both parameters $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ have dimension (K, K), whence the others are as before.

Another specification can be taken into account when the purpose is to model contagion among volatilities of different markets (Engle *et al.*, 2005). For example, we consider the case in which each component of $\boldsymbol{\mu}_t$ is the conditional variance of the corresponding market index, so that $x_{i,t} = r_{i,t}^2$. If we assume that the conditional distribution of each market index $r_{i,t} = x_{i,t}^{1/2} \operatorname{sign}(r_{i,t})$ is symmetric with zero mean, the 'contagion' (1,1) formulation can be structured exactly as in (30) where $x_{i,t}^{(-)} = x_{i,t} \, \mathrm{I}(r_{it} < 0)$ and $x_{i,t}^{(s)} = x_{i,t}^{1/2} \operatorname{sign}(r_{i,t})$.

The base, the enlarged and the contagion $(1, 1)$ specifications can be written compactly as

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\alpha}^* \mathbf{x}_{t-1}^* + \boldsymbol{\beta}\boldsymbol{\mu}_{t-1}. \tag{31}$$

Taking $\mathbf{x}_t^* = \mathbf{x}_t$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$ we have (29); considering $\mathbf{x}_t^* = (\mathbf{x}_t; \mathbf{x}_t^{(-)}; \mathbf{x}_t^{(s)})$ (of dimension $(3K, 1)$) and $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}; \boldsymbol{\gamma}; \boldsymbol{\delta})$ (of dimension $(K, 3K)$), we obtain (30).

Considering formulation (31), we have $\boldsymbol{\theta} = (\boldsymbol{\omega}; \boldsymbol{\alpha}^*; \boldsymbol{\beta})$. The parameter space of $\boldsymbol{\theta}$ must be restricted to ensure $\boldsymbol{\mu}_t \geq \mathbf{0}$ for all t and to ensure stationary distributions for $\mathbf{x}_t$. To discuss these, we consider formulation (30).

Sufficient conditions the stationarity of $\boldsymbol{\mu}_t$ are a simple generalization of those showed in section 2.1 for the univariate MEM and can be obtained in an analogous way: $\mathbf{x}_t$ is stationary if all characteristic roots of $\mathbf{A} = \boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}/2$ are smaller than 1 in modulus. We can think of $\mathbf{A}$ as the *impact matrix* in the expression

$$\boldsymbol{\mu}_t = \mathbf{A}\boldsymbol{\mu}_{t-1}. \tag{32}$$

Sufficient conditions for non–negativity of the components of $\boldsymbol{\mu}_t$ are again a generalization of the corresponding conditions of the univariate model, but the derivation is more involved. As proved in appendix E, the vector MEM defined in equation (30) gives $\boldsymbol{\mu}_t \geq 0$ for all t if all the following conditions are satisfied for all $i, j = 1, \ldots, K$:

1. $\beta_{ij} \geq 0$, $\alpha_{ij} \geq 0$, $\alpha_{ij} + \gamma_{2ij} \geq 0$ for all j;
2. if $\alpha_{ij} = 0$ then $\delta_{ij} \geq 0$; if $\alpha_{ij} + \gamma_{ij} = 0$ then $\delta_{ij} \leq 0$;
3. $\omega_i - \dfrac{1}{4} \displaystyle\sum_{j=1}^{K} \delta_{ij}^2 \left[\frac{\mathrm{I}(\delta_{ij} < 0) \, \mathrm{I}(\alpha_{ij} > 0)}{\alpha_{ij}} + \frac{\mathrm{I}(\delta_{ij} > 0) \, \mathrm{I}(\alpha_{ij} + \gamma_{ij} > 0)}{\alpha_{ij} + \gamma_{ij}}. \right] \geq 0$

3.4 Maximum likelihood inference

Given the assumptions about the conditional distribution of $\boldsymbol{\varepsilon}_t$ (section 3.2.2), we have that $\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1}$ has a pdf

$$f(\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1}) = c(F_1(\varepsilon_{1,t}), \ldots, F_K(\varepsilon_{K,t})) \prod_{i=1}^{K} f_i(\varepsilon_{i,t})$$

where $f_i(.)$ and $F_i(.)$ indicate, respectively, the conditional pdf and cdf of the i-th component of $\boldsymbol{\varepsilon}_t$, $c(.)$ is the density function of the chosen copula. In the case considered here,

$$f_i(\varepsilon_{i,t}) = \frac{\phi_i^{\phi_i}}{\Gamma(\phi_i)} \varepsilon_{i,t}^{\phi_i - 1} \exp(-\phi_i \varepsilon_{i,t})$$

$$F_i(\varepsilon_{i,t}) = \Gamma(\phi_i; \phi_i \varepsilon_{i,t})$$

$$c(\mathbf{u}_t) = |\mathbf{R}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}_t'(\mathbf{R}^{-1} - I)\mathbf{q}_t\right]$$

where $\Gamma(\zeta; x)$ indicates the incomplete Gamma function with parameter ζ computed at x (or, in other words, the cdf of a $Gamma(\zeta, 1)$ r.v. computed at x),

$$\mathbf{q}_t = (\Phi^{-1}(F_1(\varepsilon_{1,t})); \ldots; \Phi^{-1}(F_K(\varepsilon_{K,t}))),$$

$\Phi^{-1}(x)$ indicates the quantile function of the standard normal distribution computed at x.

Hence, $\mathbf{x}_t|\mathcal{F}_{t-1}$ has pdf

$$f(\mathbf{x}_t|\mathcal{F}_{t-1}) = c(F_1(x_{1,t}/\mu_{1,t}), \ldots, F_K(x_{K,t}/\mu_{K,t})) \prod_{i=1}^{K} \frac{f_i(x_{i,t}/\mu_{i,t})}{\mu_{i,t}},$$

where

$$\frac{f_i(x_{i,t}/\mu_{i,t})}{\mu_{i,t}}$$

is the pdf of a $Gamma(\phi_i, \phi_i/\mu_{i,t})$.

3.4.1 The log–likelihood

The log–likelihood of the model is then

$$l = \sum_{t=1}^{T} l_t = \sum_{t=1}^{T} \ln f(\mathbf{x}_t | \mathcal{F}_{t-1}),$$

where

$$\begin{aligned} l_t &= \ln c(F_1(x_{1,t}/\mu_{1,t}), \ldots, F_K(x_{K,t}/\mu_{K,t})) + \sum_{i=1}^{K} (\ln f_i(x_{i,t}/\mu_{i,t}) - \ln \mu_{i,t}) \\ &= \frac{1}{2} \ln |\mathbf{R}^{-1}| - \frac{1}{2} \mathbf{q}_t' \mathbf{R}^{-1} \mathbf{q}_t + \frac{1}{2} \mathbf{q}_t' \mathbf{q}_t \\ &+ \sum_{i=1}^{K} \left[\phi_i \ln \phi_i - \ln \Gamma(\phi_i) - \ln x_{i,t} + \phi_i \left(\ln x_{i,t} - \ln \mu_{i,t} - \frac{x_{i,t}}{\mu_{i,t}} \right) \right]. \\ &= (\text{copula contribution})_t + (\text{marginals contribution})_t. \end{aligned}$$

The contribution of the t-th observation to the inference can then be decomposed in the contribution of the copula plus the contribution of the marginals. The contribution of the marginals depends only on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, whereas the contribution of the copula depends on $\mathbf{R}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$. This implies

$$\frac{\partial l}{\partial \mathbf{R}^{-1}} = \frac{1}{2}(T\mathbf{R} - \mathbf{q}'\mathbf{q}) \Rightarrow \widehat{\mathbf{R}} = \frac{\mathbf{q}'\mathbf{q}}{T},$$

where $\mathbf{q} = (\mathbf{q}_1'; \ldots; \mathbf{q}_T')$ is a $T \times K$ matrix. Hence the unconstrained ML estimator of $\mathbf{R}$ has an explicit form.

Replacing the estimated $\mathbf{R}$ in the log-likelihood function we obtain a *concentrated log-likelihood*

$$\begin{aligned} lc &= -\frac{T}{2} \ln |\widehat{\mathbf{R}}| - \frac{1}{2} \sum_{t=1}^{T} q_t' (\widehat{\mathbf{R}}^{-1} - I) q_t + (\text{marginals contribution}) \\ &= -\frac{T}{2} \ln |\widehat{\mathbf{R}}| - \frac{T}{2} \left[K - \text{trace}(\widehat{\mathbf{R}}) \right] + (\text{marginals contribution}) \end{aligned}$$

In deriving the concentrated log–likelihood, $\mathbf{R}$ is estimated without imposing any constraint relative to its nature as correlation matrix ($\text{diag}(\mathbf{R}) = \mathbb{1}$ and positive definiteness). Computing directly the derivatives with respect to the off–diagonal elements of $\mathbf{R}$ we obtain, after some algebra, that the ML estimate of $\mathbf{R}$ satisfies the following equations:

$$(\mathbf{R}^{-1})_{ij} - (\mathbf{R}^{-1})_{i\cdot} \frac{\mathbf{q}'\mathbf{q}}{T} (\mathbf{R}^{-1})_{\cdot j} = 0$$

for $i \neq j = 1, \ldots, K$, where $\mathbf{R}_{i\cdot}$ and $\mathbf{R}_{\cdot j}$ indicate, respectively, the i–th row and the j–th

column of the matrix $\mathbf{R}$. Unfortunately, these equations do not have an explicit solution.[10]

An acceptable compromise which should increase efficiency, although formally it cannot be interpreted as an ML estimator, is to adopt the sample correlation matrix of the $\mathbf{q}_t$'s as a constrained estimator of $\mathbf{R}$, that is

$$\widetilde{\mathbf{R}} = \mathbf{D}_Q^{-\frac{1}{2}} \mathbf{Q} \mathbf{D}_Q^{-\frac{1}{2}}.$$

where

$$\mathbf{Q} = \frac{\mathbf{q}'\mathbf{q}}{T} \qquad \mathbf{D}_Q = \mathrm{diag}(Q_{11}, \ldots, Q_{KK}).$$

This solution can be justified observing that the copula contribution to the likelihood depends on $\mathbf{R}$ exactly as if it were the correlation matrix of i.i.d. r.v. $\mathbf{q}_t$ normally distributed with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}$. Using this new estimate of $\mathbf{R}$, the trace of $\widetilde{\mathbf{R}}$ is now constrained to K and the concentrated log-likelihood simplifies to

$$\begin{aligned} lc &= -\frac{T}{2} \ln |\widetilde{\mathbf{R}}| + (\text{marginals contribution}) \\ &= -\frac{T}{2} \left[\ln |\mathbf{q}'\mathbf{q}| - \sum_{i=1}^{K} \ln(\mathbf{q}_{\cdot i}' \mathbf{q}_{\cdot i}) \right] + (\text{marginals contribution}). \end{aligned}$$

3.4.2 Derivatives of the concentrated log-likelihood

In what follows, let us consider $\boldsymbol{\lambda} = (\boldsymbol{\theta}; \boldsymbol{\phi})$. As

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \ln |\mathbf{q}'\mathbf{q}| &= \frac{2}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\lambda}} \mathbf{q}_t' \mathbf{Q}^{-1} \mathbf{q}_t \\ \nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{K} \ln(\mathbf{q}_{\cdot i}' \mathbf{q}_{\cdot i}) &= \frac{2}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\lambda}} \mathbf{q}_t' \mathbf{D}_Q^{-1} \mathbf{q}_t \end{aligned}$$

the derivative of lc is given by

$$\nabla_{\boldsymbol{\lambda}} lc = \sum_{t=1}^{T} \left[\nabla_{\boldsymbol{\lambda}} \mathbf{q}_t' [\mathbf{D}_Q^{-1} - \mathbf{Q}^{-1}] \mathbf{q}_t + \nabla_{\boldsymbol{\lambda}} (\text{marginals contribution})_t \right].$$

[10] Even when $\mathbf{R}$ is a $(2,2)$ matrix, the value of $\mathbf{R}_{12}$ has to satisfy a cubic equation as the following:

$$\mathbf{R}_{12}^3 - \mathbf{R}_{12}^2 \frac{q_1' q_2}{T} + \mathbf{R}_{12} \left[\frac{q_1' q_1}{T} + \frac{q_2' q_2}{T} - 1 \right] - \frac{q_1' q_2}{T} = 0.$$

To develop further this formula we need to distinguish the derivatives $\nabla_{\boldsymbol{\lambda}}\mathbf{q}_t'$ and $\nabla_{\boldsymbol{\lambda}}(\text{marginals contribution})_t$ with respect to $\boldsymbol{\theta}$ (the parameters of $\boldsymbol{\mu}_t$) and $\boldsymbol{\phi}$. After some algebra we obtain

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathbf{q}_t' &= -\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'\mathbf{D}_{1t} \\
\nabla_{\boldsymbol{\theta}}(\text{marginals contribution})_t &= \nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'\mathbf{v}_{1t} \\
\nabla_{\boldsymbol{\phi}}\mathbf{q}_t' &= \mathbf{D}_{2t} \\
\nabla_{\boldsymbol{\phi}}(\text{marginals contribution})_t &= \mathbf{v}_{2t},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{D}_{1t} &= \operatorname{diag}\left(\frac{f_i(\widehat{\varepsilon}_{i,t})\widehat{\varepsilon}_{i,t}}{\phi(q_{i,t})\mu_{i,t}} : i = 1,\ldots,K\right) \\
\mathbf{v}_{1t} &= \left(\phi_i\frac{\widehat{\varepsilon}_{i,t}-1}{\mu_{i,t}} : i = 1,\ldots,K\right) \\
\mathbf{D}_{2t} &= \operatorname{diag}\left(\frac{1}{\phi(q_{i,t})}\frac{\partial F_i(\widehat{\varepsilon}_{i,t})}{\partial\phi_i} : i = 1,\ldots,K\right) \\
\mathbf{v}_{2t} &= \left(\ln\phi_i - \psi(\phi_i) + \ln(\widehat{\varepsilon}_{i,t}) - \widehat{\varepsilon}_{i,t} + 1 : i = 1,\ldots,K\right).
\end{aligned}$$

and $\widehat{\varepsilon}_{i,t} = x_{i,t}/\mu_{i,t}$. Then

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} lc &= \sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'\left[-\mathbf{D}_{1t}(\mathbf{D}_Q^{-1}-\mathbf{Q}^{-1})\mathbf{q}_t + \mathbf{v}_{1t}\right] \\
\nabla_{\boldsymbol{\phi}} lc &= \sum_{t=1}^{T}\left[\mathbf{D}_{2t}(\mathbf{D}_Q^{-1}-\mathbf{Q}^{-1})\mathbf{q}_t + \mathbf{v}_{2t}\right]
\end{aligned}$$

We remark that the derivatives $\dfrac{\partial F_i(\widehat{\varepsilon}_{i,t})}{\partial\phi_i}$ must be computed numerically.

3.4.3 Details about the estimation of $\boldsymbol{\theta}$

The estimation of the parameter $\boldsymbol{\theta}$ involved in the equation of $\boldsymbol{\mu}_t$ requires further details. We explain them considering the more general version of $\boldsymbol{\mu}_t$ defined in section 3.3

$$\begin{aligned}
\boldsymbol{\mu}_t &= \boldsymbol{\omega} + \boldsymbol{\alpha}\mathbf{x}_{t-1} + \boldsymbol{\gamma}\mathbf{x}_{t-1}^{(-)} + \boldsymbol{\delta}\mathbf{x}_{t-1}^{(s)} + \boldsymbol{\beta}\boldsymbol{\mu}_{t-1} \\
&= \boldsymbol{\omega} + \boldsymbol{\alpha}^*\mathbf{x}_{t-1}^* + \boldsymbol{\beta}\boldsymbol{\mu}_{t-1}
\end{aligned}$$

(see the cited section for details and notation). Such a structure for $\boldsymbol{\mu}_t$ depends in the general case from $K + 4K^2$ parameters. For instance, when $K = 3$ there are 39 parameters. We think that, in general, data do not provide enough information to capture asymmetries of both types ($\mathbf{x}_{t-1}^{(s)}$ and $\mathbf{x}_{t-1}^{(-)}$) but even removing one of these two components the parameters become $K + 3K^2$ (30 when $K = 3$).

A reduction in the number of parameters can obtained estimating $\boldsymbol{\omega}$ from stationary con-

ditions. Imposing that $\boldsymbol{\mu}_t$ is stationary we have

$$\boldsymbol{\omega} = [\mathbf{I} - (\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}/2)]\boldsymbol{\mu},$$

where $\boldsymbol{\mu} = E(\mathbf{x}_t)$, and then

$$(\boldsymbol{\mu}_t - \boldsymbol{\mu}) = \boldsymbol{\alpha}(\mathbf{x}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\gamma}(\mathbf{x}_{t-1}^{(-)} - \boldsymbol{\mu}/2) + \boldsymbol{\delta}\mathbf{x}_{t-1}^{(s)} + \boldsymbol{\beta}(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}).$$

Replacing $\boldsymbol{\mu}$ with its natural estimate, that is the unconditional average $\overline{\mathbf{x}}$, we obtain

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_t &= \boldsymbol{\alpha}\widetilde{\mathbf{x}}_{t-1} + \boldsymbol{\gamma}\widetilde{\mathbf{x}}_{t-1}^{(-)} + \boldsymbol{\delta}\widetilde{\mathbf{x}}_{t-1}^{(s)} + \boldsymbol{\beta}\widetilde{\boldsymbol{\mu}}_{t-1} \\ &= \boldsymbol{\alpha}^*\widetilde{\mathbf{x}}_{t-1}^* + \boldsymbol{\beta}\widetilde{\boldsymbol{\mu}}_{t-1} \end{aligned} \tag{33}$$

where the symbol $\widetilde{\mathbf{x}}$ means the demeaned version of $\mathbf{x}$. This solution save K parameters in the iterative estimation and provides better performances than direct ML estimates of $\boldsymbol{\omega}$ in simulations.

To save time it is also useful take into account analytic derivatives of $\widetilde{\boldsymbol{\mu}}_t$ with respect to parameters. To this purpose rewrite (33) as

$$\widetilde{\boldsymbol{\mu}}_t = \widetilde{\mathbf{B}}_{t-1}' \operatorname{vec}(\boldsymbol{\beta}') + \widetilde{\mathbf{A}}_{t-1}^{*\prime} \operatorname{vec}(\boldsymbol{\alpha}^{*\prime}),$$

where

$$\widetilde{\mathbf{B}}_t = \begin{pmatrix} \widetilde{\boldsymbol{\mu}}_t & 0 & \dots & 0 \\ 0 & \widetilde{\boldsymbol{\mu}}_t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widetilde{\boldsymbol{\mu}}_t \end{pmatrix} \qquad \widetilde{\mathbf{A}}_t^* = \begin{pmatrix} \widetilde{\mathbf{x}}_t^* & 0 & \dots & 0 \\ 0 & \widetilde{\mathbf{x}}_t^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widetilde{\mathbf{x}}_t^* \end{pmatrix}$$

and the operator $\operatorname{vec}(.)$ stacks the columns of the matrix inside brackets. Taking derivatives with respect to $\operatorname{vec}(\boldsymbol{\beta}')$ and $\operatorname{vec}(\boldsymbol{\alpha}^{*\prime})$ and arranging results we obtain

$$\frac{\partial \widetilde{\boldsymbol{\mu}}_t}{\partial \operatorname{vec}(\boldsymbol{\theta})} = \begin{pmatrix} \widetilde{\mathbf{B}}_{t-1} \\ \widetilde{\mathbf{A}}_{t-1}^* \end{pmatrix} + \frac{\partial \widetilde{\boldsymbol{\mu}}_{t-1}}{\partial \operatorname{vec}(\boldsymbol{\theta})}$$

where $\operatorname{vec}(\boldsymbol{\theta}) = (\operatorname{vec}(\boldsymbol{\beta}'); \operatorname{vec}(\boldsymbol{\alpha}^{*\prime}))$.

3.5 Estimating Functions inference

A different approach to estimation and inference of a vector MEM is provided by *Estimating Functions* (Heyde, 1997) which has less demanding assumptions relative to ML: in the simplest version, Estimating Functions require only the specification of the first two conditional moments of $\mathbf{x}_t$, as in (24) and (25). The advantage of not formulating assumptions about the shape of the conditional distribution of $\mathbf{x}_t$ is balanced by a loss in efficiency with respect to Maximum likelihood under correct specification of the complete model. Given that even in the univariate MEM we had interpreted the ML estimator as QML, the flexibility provided by the Estimating Functions (henceforth EF) seems promising.

3.5.1 Framework

We consider here the EF inference of the vector MEM defined by (22) and (23) or, equivalently, by (24) and (25). In this framework the parameters are collected in the p^*-dimensional vector $\boldsymbol{\lambda} = (\boldsymbol{\theta}; \mathrm{upmat}(\boldsymbol{\Sigma}))$, with $p^* = p + K(K+1)/2$: $\boldsymbol{\theta}$ includes the parameters of primary interest involved in the expression of $\boldsymbol{\mu}_t$; $\mathrm{upmat}(\boldsymbol{\Sigma})$ includes the elements inside and above the main diagonal of the conditional variance–covariance matrix $\boldsymbol{\Sigma}$ of the multiplicative innovations $\boldsymbol{\varepsilon}_t$ and represents a nuisance parameter with respect to $\boldsymbol{\theta}$. However, to simplify the exposition, from now on we denote the parameters of the model as $\boldsymbol{\lambda} = (\boldsymbol{\theta}; \boldsymbol{\Sigma})$, instead of $\boldsymbol{\lambda} = (\boldsymbol{\theta}; \mathrm{upmat}(\boldsymbol{\Sigma}))$.

Following Bibby *et al.* (2004), an *estimating function* for the p^*-dimensional vector $\boldsymbol{\lambda}$ based on a sample $\mathbf{x}_{(T)}$ is a p^*-dimensional function denoted as

$$\mathbf{g}(\boldsymbol{\lambda}; \mathbf{x}_{(T)}) \quad \text{in short} \quad \mathbf{g}(\boldsymbol{\lambda}).$$

The EF estimator $\widehat{\boldsymbol{\lambda}}$ is defined as the solution to the corresponding *estimating equation*:

$$\widehat{\boldsymbol{\lambda}} \text{ such that } \mathbf{g}(\widehat{\boldsymbol{\lambda}}) = 0. \tag{34}$$

To be an useful estimating function, regularity conditions on $\mathbf{g}$ are usually imposed. Following Heyde (1997, sect. 2.6) or Bibby *et al.* (2004, Theorem 2.3), we consider zero mean and square integrable martingale estimating functions that are sufficiently regular to guarantee that a WLLN and a CLT applies. The martingale restriction is often motivated observing that the score function, if it exists, is usually a martingale: so it is quite natural to approximate it using families of martingales estimating functions. Furthermore, in modelling time series processes, estimating functions that are martingales arise quite naturally from the structure of the process. We denote as $\mathcal{M}$ the class of EF satisfying the asserted regularity conditions.

Following the decomposition of $\boldsymbol{\lambda}$ into $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, let us decompose the estimating function $\mathbf{g}(\boldsymbol{\lambda})$ as

$$\mathbf{g}(\boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{g}_1(\boldsymbol{\lambda}) \\ \mathbf{g}_2(\boldsymbol{\lambda}) \end{pmatrix} : \tag{35}$$

$\mathbf{g}_1$ has dimension p and refers to $\boldsymbol{\theta}$; $\mathbf{g}_2$ has dimension $K(K+1)/2$ and refers to $\boldsymbol{\Sigma}$.

3.5.2 Inference on θ

From (24) and (25) we obtain immediately that "residuals" $\mathbf{u}_t = \mathbf{x}_t \oslash \boldsymbol{\mu}_t - \mathbb{1}$, where $\oslash$ denotes the element by element division, have the property to be martingale differences with a constant conditional variance matrix: in fact

$$\mathbf{u}_t | \mathcal{F}_{t-1} \sim [\mathbf{0}, \boldsymbol{\Sigma}]. \tag{36}$$

Using these as basic ingredient, we can construct the *Hutton–Nelson quasi-score function* (Heyde (1997, p. 32)) as an estimating function for $\boldsymbol{\theta}$:

$$\begin{aligned}\mathbf{g}_1(\boldsymbol{\lambda}) &= -\sum_{t=1}^{T} E\left[\nabla_{\boldsymbol{\theta}}\mathbf{u}_t'|\mathcal{F}_{t-1}\right] V\left[\mathbf{u}_t|\mathcal{F}_{t-1}\right]^{-1}\mathbf{u}_t \\ &= \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'[\operatorname{diag}(\boldsymbol{\mu}_t)\boldsymbol{\Sigma}\operatorname{diag}(\boldsymbol{\mu}_t)]^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_t). \qquad (37)\end{aligned}$$

Transposing to the vector MEM considered here the arguments in Heyde (1997), (37) is (for fixed $\boldsymbol{\Sigma}$) the "best" estimating function for $\boldsymbol{\theta}$ in the subclass of $\mathcal{M}$ composed by all estimating functions that are linear in the residuals $\mathbf{u}_t = \mathbf{x}_t \oslash \boldsymbol{\mu}_t - \mathbb{1}$. In fact, more precisely, the Hutton–Nelson quasi score function is *optimal* (both in the *asymptotic* and in the *fixed sample* sense – see Heyde (1997, ch. 2) for details) in the class of estimating functions

$$\mathcal{M}^{(1)} = \left\{ \mathbf{g} \in \mathcal{M} : \mathbf{g}(\boldsymbol{\lambda}) = \sum_{t=1}^{T} \mathbf{a}_t(\boldsymbol{\lambda})\mathbf{u}_t(\boldsymbol{\lambda}) \right\},$$

where $\mathbf{u}_t(\boldsymbol{\lambda})$ are $(K,1)$ martingale differences and $\mathbf{a}_t(\boldsymbol{\lambda})$ are (p,K) $\mathcal{F}_{t-1}$–measurable functions.

Very interestingly, in the 1-dimensional case (37) specializes to

$$\mathbf{g}_1(\boldsymbol{\lambda}) = \sigma^{-2}\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}}\mu_t \frac{x_t-\mu_t}{\mu_t^2},$$

which provides exactly the first order condition of the univariate MEM. Hence, formula (17) can be also justified from an EF perspective. The main substantial difference from the multivariate case, is that in the vector MEM explicit dependence from the nuisance parameter $\boldsymbol{\Sigma}$ cannot be suppressed. About this point, however, we claim that EF inferences of vector MEM are not too seriously affected by the presence of $\boldsymbol{\Sigma}$, as stated in the following section.

3.5.3 Inference on $\boldsymbol{\Sigma}$

A clear discussion of inferential issues with nuisance parameters can be found, among others, in Liang and Zeger (1995). The main inferential problem is that, in presence of nuisance parameters, some properties of estimating functions are no longer valid if we replace parameters with the corresponding estimators. For instance, unbiasedness of $\mathbf{g}_1(\boldsymbol{\lambda})$ is not guaranteed if $\boldsymbol{\Sigma}$ is replaced by an estimator $\widehat{\boldsymbol{\Sigma}}$. By consequence, optimality properties also are not guaranteed.

An interesting statistical handling of nuisance parameter in the estimating functions framework is provided in Knudsen (1999) and Jørgensen and Knudsen (2004). Their handling parallels in some aspects the notion of *Fisher-orthogonality* or, shortly, *F-orthogonality*, in ML estimation. F-orthogonality is defined by block diagonality of the Fisher informa-

tion matrix for $\boldsymbol{\lambda} = (\boldsymbol{\theta}; \boldsymbol{\Sigma})$. This particular structure guarantees the following properties of the ML estimator (see the cited reference for details):

1. asymptotic independence of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$;
2. efficiency-stable estimation of $\boldsymbol{\theta}$, in the sense that the asymptotic variance for $\boldsymbol{\theta}$ is the same whether $\boldsymbol{\Sigma}$ is treated as known or unknown;
3. simplification of the estimation algorithm;
4. $\widehat{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$, the estimate of $\boldsymbol{\theta}$ when $\boldsymbol{\Sigma}$ is given, varies only slowly with $\boldsymbol{\Sigma}$.

Starting from this point, Jørgensen and Knudsen (2004) extend F-orthogonality to EFs introducing the concepts of *nuisance parameter insensitivity* (henceforth NPI). NPI is an extended parameter orthogonality notion for estimating function, that guarantees properties 2-4 above. Among these, Jørgensen and Knudsen (2004) state that efficiency stable estimation is the crucial property and prove that, in a certain sense, NPI is a necessary condition for this.

Here we emphasize two major points concerning the vector MEM:

1. Jørgensen and Knudsen (2004) define and prove properties connected to NPI within a *fixed sample* framework. However definitions and theorems can be easily adapted and extended to the *asymptotic* framework, that is the typical reference when stochastic processes are of interest.
2. The estimating function (37) for estimating $\boldsymbol{\theta}$ can be proved nuisance parameter insensitive.

About the first point, key quantities for inferences on estimating functions estimators within the fixed sample framework are the *variance–covariance matrix* $V(\mathbf{g})$ and the *sensitivity matrix* $E(\nabla_{\boldsymbol{\lambda}}\mathbf{g}')$ of the EF g. These quantities are replaced in the asymptotic framework by their 'estimable' counterparts: the *quadratic characteristic*

$$\langle \mathbf{g} \rangle = \sum_{t=1}^{T} V(\mathbf{g}_t | \mathcal{F}_{t-1}),$$

and the *compensator*,

$$\overline{\mathbf{g}} = \sum_{t=1}^{T} E\left(\nabla_{\boldsymbol{\lambda}'} \mathbf{g}_t | \mathcal{F}_{t-1}\right),$$

of g, where the martingale EF g is represented as sum of martingale differences $\mathbf{g}_t$:

$$\mathbf{g} = \sum_{t=1}^{T} \mathbf{g}_t. \tag{38}$$

As pointed by Heyde (1997, p. 28), there is a close relation between $V(\mathbf{g})$ and $\langle \mathbf{g} \rangle$ and between $E(\nabla_{\lambda'}\mathbf{g})$ and $\overline{\mathbf{g}}$. But, above all, the optimality theory within the asymptotic

framework parallels substantially that within the fixed sample framework because these quantities share the same essential properties. These properties can be summarized taking into account two points. The first one: under regularity conditions (see Heyde (1997), Bibby *et al.* (2004) and Jørgensen and Knudsen (2004) for details) we have,

$$E(\nabla_{\boldsymbol{\lambda}'}\mathbf{g}) = C(\mathbf{g}, \mathbf{s})$$

and

$$\overline{\mathbf{g}} = \sum_{t=1}^{T} C(\mathbf{g}_t, \mathbf{s}_t | \mathcal{F}_{t-1}),$$

where $\mathbf{s} = \sum_{t=1}^{T} \mathbf{s}_t$ is the score function expressed as a sum of martingale differences. The second one: both the covariance and the sum of conditional covariances of the martingale difference components share the bilinearity properties typical of a scalar product. Just by virtue of these common properties, it is possibile to show that NPI can be defined and guarantees properties 2–4 above also in the asymptotic framework. Hence, we sketch here only the main points and we remind to Jørgensen and Knudsen (2004) for the original full treatment.

We partition the compensator and the quadratic characteristic conformably to $\mathbf{g}$ as in (35), that is

$$\overline{\mathbf{g}} = \begin{pmatrix} \overline{\mathbf{g}}_{11} & \overline{\mathbf{g}}_{12} \\ \overline{\mathbf{g}}_{21} & \overline{\mathbf{g}}_{22} \end{pmatrix}$$

and

$$\langle \mathbf{g} \rangle = \begin{pmatrix} \langle \mathbf{g} \rangle_{11} & \langle \mathbf{g} \rangle_{12} \\ \langle \mathbf{g} \rangle_{21} & \langle \mathbf{g} \rangle_{22} \end{pmatrix},$$

and we consider *Heyde information*

$$\mathbf{I} = \overline{\mathbf{g}}' \langle \mathbf{g} \rangle^{-1} \overline{\mathbf{g}} \tag{39}$$

as the natural counterpart, within the asymptotic framework, of the *Godambe information*

$$\mathbf{J} = E(\nabla_{\boldsymbol{\lambda}} \mathbf{g}') V(\mathbf{g})^{-1} E(\nabla_{\boldsymbol{\lambda}'} \mathbf{g}),$$

that instead is typical of the fixed sample framework. We remember that each of these versions of *information*, in the respective framework, provides the inverse of the asymptotic variance covariance matrix of the EF estimator.

We say that the marginal EF $\mathbf{g}_1$ is *nuisance parameter insensitive* if $\overline{\mathbf{g}}_{12} = 0$. Using block–matrix algebra, we check easily that NPI of $\mathbf{g}_1$ implies that the asymptotic variance–covariance matrix of the EF estimator of $\boldsymbol{\theta}$ is given by

$$\overline{\mathbf{g}}_{11}^{-1\prime} \langle \mathbf{g} \rangle_{11} \overline{\mathbf{g}}_{11}^{-1}. \tag{40}$$

Employing concepts and definitions above is possible to show, exactly using the same devices, that the whole theory in Jørgensen and Knudsen (2004), and in particular their *insensitivity theorem*, preserves substantially unaltered.

Returning now to the vector MEM, we can discuss the second point above: the EF (37) for estimating $\boldsymbol{\theta}$ is $\boldsymbol{\Sigma}$–insensitive. In fact, expressing (37) as sum of martingale differences $\mathbf{g}_{1,t} = \nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'(\operatorname{diag}(\boldsymbol{\mu}_t)\boldsymbol{\Sigma}\operatorname{diag}(\boldsymbol{\mu}_t))^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_t)$ as in (38), we have

$$\nabla_{\sigma_{ij}}\mathbf{g}_{1,t}(\boldsymbol{\lambda}) = -\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t' \operatorname{diag}(\boldsymbol{\mu}_t)^{-1}\boldsymbol{\Sigma}^{-1}\nabla_{\sigma_{ij}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\operatorname{diag}(\boldsymbol{\mu}_t)^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_t),$$

whose conditional expected value is $\mathbf{0}$. This implies $\overline{\mathbf{g}}_{12} = \mathbf{0}$ and hence $\boldsymbol{\Sigma}$-insensitivity of the EF $\mathbf{g}_1$ in (37). As a consequence, the asymptotic variance–covariance matrix of $\boldsymbol{\theta}$ is given by (40), that for the model considered becomes

$$\overline{\mathbf{g}}_{11}^{-1\prime}\langle\mathbf{g}\rangle_{11}\overline{\mathbf{g}}_{11}^{-1} = \langle\mathbf{g}\rangle_{11}^{-1} = \left[\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_t'[\operatorname{diag}(\boldsymbol{\mu}_t)\boldsymbol{\Sigma}\operatorname{diag}(\boldsymbol{\mu}_t)]^{-1}\nabla_{\boldsymbol{\theta}'}\boldsymbol{\mu}_t\right]^{-1}. \tag{41}$$

Expressions (37) and (41) require values for $\boldsymbol{\Sigma}$. Considering the conditional distribution of the "residuals" $\mathbf{u}_t = \mathbf{x}_t \oslash \boldsymbol{\mu}_t - \mathbb{1}$ in (36), $\boldsymbol{\Sigma}$ can be estimated using the estimating function:

$$\mathbf{g}_2(\boldsymbol{\lambda}) = \sum_{t=1}^{T}(\mathbf{u}_t\mathbf{u}_t' - \boldsymbol{\Sigma}). \tag{42}$$

This implies

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'.$$

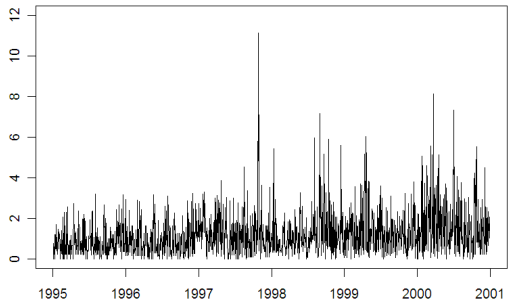
This estimator generalizes (19) to the multivariate model. We note that the EF (42) is unbiased, but in general does not satisfy optimality properties: improving (42) is possible but at the cost of adding further assumptions about higher order moments of $\mathbf{x}_t|\mathcal{F}_{t-1}$, and will not be pursued here.

4 Empirical Analysis

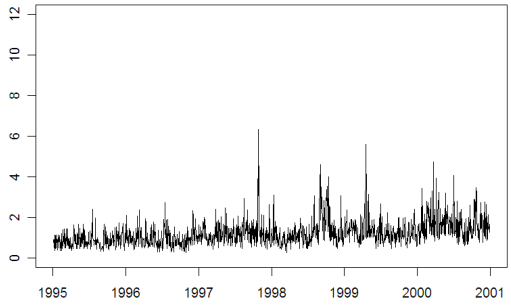
The multivariate extension of the MEM model is illustrated on the GE stock for which the three variables are available as in Engle and Gallo (2006), namely absolute returns $|r_t|$, high-low range hl_t and realized volatility rv_t derived from 5-minutes returns as in Andersen *et al.* (2003). The observations were rescaled so as to have the same mean as absolute returns.

In Table 1 we report a few descriptive statistics relative to the variables employed. As one may expect, daily range and realized volatility have high maximum values and there are a few zeroes in the absolute returns. We also report the unconditional correlations across the whole sample period showing that daily range has a relatively high correlation with both absolute returns and realized volatility (which are less correlated with one another). Finally, the time series behavior of the series is depicted in Figure 1.

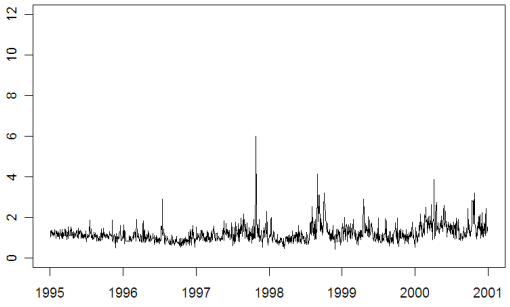
Figure 1: Time series of the indicators $|r_t|$, hl_t, rv_t – GE stock, 01/03/1995-12/29/2001.



(a) $|r_t|$



(b) hl_t



(c) rv_t

Table 1: Some descriptive statistics of the indicators $|r_t|$, hl_t, rv_t – GE stock, 01/03/1995-12/29/2001 (1515 obs.)

	Indicator		
Statistics	$\|r_t\|$	hl_t	rv_t
min	0	0.266	.429
max	11.123	6.318	6.006
mean	1.194	1.194	1.194
standard deviation	1.058	0.614	0.424
skewness	2.20	2.03	2.65
kurtosis	12.42	10.97	18.69
n. of zeros	60	0	0
correlations	$\|r_t\|$	hl_t	rv_t
$\|r_t\|$		0.767	0.440
hl_t			0.757

We applied the three estimators, namely, the Maximum Likelihood estimator with the Normal Copula function (labelled 'ML–R'), the Estimating Equations-based estimator (labelled 'EE') and the Maximum Likelihood estimator equation by equation (labelled 'ML–I own'). For each method a general–to–specific model selection strategy was followed to propose a parsimonious model. The first table (4 compares single coefficients (with their robust standard errors) by estimation method. ML–R and EE give the same specification for both stocks, while consistently ML–I gives a different result. For comparison purposes, we added an extra column by variable to include the results of the equation–by–equation estimation imposing the specification found with the system methods (labelled 'ML-I').

We can note that the selected model for GE is in line with the findings of Engle and Gallo (2006) for the S&P500, in that the daily range has significant lagged (be they overall or asymmetric) effects on the other variables, while it is not affected by them.[11] A synthetic picture of the total influence of variables can be measured by means of an *Impact Matrix* **A**, previously defined in (32). We propose a comparison of the estimated values of the non–zero elements of the matrix **A** in Table 3.

Overall, the estimated coefficient values do not differ too much between the system estimation with copulas and with estimating equations. The estimates equation by equation are quite different even when derived from the same specification and even more so when derived from the own selection.

Similar comments can be had for the estimated characteristic roots of the impact matrix ruling the dynamics of the system: in this particular case, the estimated persistence is underestimated by the equation–by–equation estimator relative to the other two methods.

A comment on the shape of the distribution as it emerges from the estimation methods

[11] This is not to say that one should expect the same result on other assets. An extensive analysis of what may happen when several stocks are considered is beyond the scope of the present paper.

based on the Gamma distribution is in order (cf. Table 5). The estimated ϕ parameters are quite different from one another (there is no major difference according to whether the system is estimated equation–by–equation or jointly) reflecting the different characteristics of each variable (with a decreasing skewness passing from absolute returns to realized volatility).

5 Conclusions

In this paper we have presented a general discussion of the vector specification of the Multiplicative Error Models introduced by Engle (2002): a positive valued process is seen as the product of a scale factor which follows a GARCH type specification and a unit mean innovation process. Engle and Gallo (2006) estimate a system version of the MEM by adopting a dynamically interdependent specification for the scale factors (each variable enters other variables' specifications with a lag) but keeping a diagonal variance–covariance matrix for the Gamma–distributed innovations. The extension to a multivariate process requires interdependence among the innovation terms. The specification in a multivariate framework cannot exploit multivariate Gamma distributions because they appear too restrictive. The maximum likelihood estimator can be derived by setting the multivariate innovation process in a copula with Gamma marginals framework. As an alternative, which provides estimators which are not based on any distributional assumptions, we also suggested the use of estimating equations. The empirical results with daily data on the GE stock show the feasibility of both procedures: both system estimators provide fairly similar values, and at the same time show deviations from the equation by equation approach.

A Useful univariate statistical distributions

In this section we summarize the properties of the statistical distributions employed in the work.

A.1 The Gamma distribution

A r. v. X follows a $Gamma(\phi, \beta)$ distribution if its pdf if given by

$$f(x) = \frac{\beta^{\phi}}{\Gamma(\phi)} x^{\phi-1} e^{-\beta x}$$

for $x > 0$ and 0 otherwise, where $\phi, \beta > 0$. The main moments are

$$\begin{aligned} E(X) &= \phi/\beta \\ V(X) &= \phi/\beta^2. \end{aligned}$$

An useful property of a Gamma r.v. is the following:

$$X \sim Gamma(\phi, \beta) \text{ and } c > 0 \Rightarrow Y = cX \sim Gamma(\phi, \beta/c).$$

Some interesting special cases of a $Gamma$ distributions are the Exponential distribution and the χ^2 distribution:

$$\begin{aligned} Exponential(\beta) &= Gamma(1, \beta) \\ \chi^2(n) &= Gamma(n/2, 1/2). \end{aligned}$$

In this work we often employ $Gamma$ distributions restricted to have mean 1, that implies $\beta = \phi$.

A.2 The GED distribution

The Generalized Error distribution (GED) is also known as Exponential Power distribution or as Subbotin distribution.

A random variable X follow a $GED(\mu, \sigma, \phi)$ distribution distribution if its pdf if given by

$$f(x) = \frac{\phi^{\phi-1}}{2\Gamma(\phi)\sigma} \exp\left(-\phi \left|\frac{x-\mu}{\sigma}\right|^{1/\phi}\right)$$

for $x \in \mathbb{R}$, where $\mu \in \mathbb{R}$, $\sigma, \phi > 0$. The main moments are

$$E(X) = \mu$$
$$V(X) = \frac{\Gamma(3\phi)}{\phi^{2\phi}\Gamma(\phi)}\sigma^2.$$

An useful property of a GED r.v. is the following:

$$X \sim GED(\mu, \sigma, \phi) \text{ and } c_1, c_2 \in \mathbb{R} \Rightarrow Y = c_1 + c_2 X \sim GED(c_1 + c_2\mu, |c_2|\sigma, \phi).$$

This property qualifies the GED as a location–scale distribution.

Some interesting special cases of a GED distributions are the Laplace distribution and the Normal distribution:

$$Laplace(\mu, \sigma) = GED(\mu, \sigma, 1)$$
$$N(\mu, \sigma) = GED(\mu, \sigma, 1/2).$$

In the work we consider in particular GED distributions with $\mu = 0$.

A.3 Relations GED–Gamma

The following relations links the GED and Gamma distributions:

$$X \sim GED(0, \sigma, \phi) \Rightarrow Y = \left|\frac{X - \mu}{\sigma}\right|^{1/\phi} \sim Gamma(\phi, \phi)$$

Useful particular cases are:

$$X \sim GED(0, 1, \phi) \Rightarrow Y = |X|^{1/\phi} \sim Gamma(\phi, \phi)$$
$$X \sim Laplace(0, 1) \Rightarrow Y = |X| \sim Exponential(1)$$
$$X \sim N(0, 1) \Rightarrow Y = X^2 \sim \chi^2(1).$$

B A summary on copulas

In this section we summarize main concepts and properties of copulas in a rather informal way and without proofs. For a more detailed and rigorous exposition see, among others, Embrechts *et al.* (2001) and Bouyé *et al.* (2000). For a simpler treatment, we assume that all r.v. considered in this section are absolutely continuous.

B.1 Definitions

Copula. Defined "operationally", a K–dimensional *copula* C is the cdf of a continuous uniform r.v. defined on the unit hypercube $[0,1]^K$, in the sense that each of the univariate components of the r.v. has $Uniform(0,1)$ marginals, whenever they may be not independent.

Copula density. A copula C is then a continuous cdf with particular characteristics. For some purposes it is however useful the associated *copula density* c, defined as

$$c(\mathbf{u}) = \frac{\partial^K C(\mathbf{u})}{\partial u_1 \dots \partial u_K}. \tag{43}$$

Ordering. A copula C_1 is *smaller* then C_2, in symbols $C_1 \prec C_2$ if $C_1(\mathbf{u}) \leq C_2(\mathbf{u})$ $\forall \mathbf{u} \in [0,1]^K$. In practice this means that the picture of C_1 stay below that of C_2.

Some particular copulas. Among copulas, three play an important role: the *lower Frèchet bound* C^-, the *product copula* $C^\perp$ and the *upper Frèchet bound* C^+, defined as follows [12]

$$\begin{aligned} C^-(\mathbf{u}) &= \max\left(\sum_{i=1}^K u_i - K + 1, 0\right) \\ C^\perp(\mathbf{u}) &= \prod_{i=1}^K u_i \\ C^+(\mathbf{u}) &= \min(u_1, \dots, u_K). \end{aligned}$$

We could show that

$$C^-(\mathbf{u}) \prec C^\perp(\mathbf{u}) \prec C^+(\mathbf{u}).$$

More in general, for any copula C,

$$C^-(\mathbf{u}) \prec C(\mathbf{u}) \prec C^+(\mathbf{u}).$$

The concept of ordering is interesting because is connected with measures of concordance (like the correlation coefficient). Considering two r.v., the product copula $C^\perp(\mathbf{u})$ identifies the situation of independence; copulas smaller then $C^\perp(\mathbf{u})$ are connected with situations of discordance; copulas greater then $C^\perp(\mathbf{u})$ describe situations of concordance.

The ordering on the set of copulas, called *concordance ordering*, is however only a partial ordering, since not every pair of copulas is comparable in this way. However many important parametric families of copulas are totally ordered on the base of the values of the parameters (examples are the Frank copula and the Normal copula).

[12] C^- is not a copula for $K \geq 3$ but we use this notation for convenience.

B.2 Understanding copulas

We attempt to explain the usefulness of copulas starting from two important results.

The first one is relative to these well known relations concerning univariate r.v.:

$$X \sim F \Rightarrow U = F(X) \sim Uniform(0,1)$$

and, conversely,

$$U \sim Uniform(0,1) \Rightarrow F^{-1}(U) \sim F$$

where with the symbol $X \sim F$ we mean 'X distributed with cdf F'. F^{-1} is often called the *quantile function* of X.

The second one is Sklar's theorem. Sklar's theorem is perhaps the most important result regarding copulas.

Theorem 1 ***(Sklar's theorem)****. Let F a K–dimensional cdf with univariate continuous marginals $F_1, \ldots, F_K$. Then there exist an unique K–dimensional copula C such that, $\forall \mathbf{x} \in \mathbb{R}^K$,*

$$F(x_1, \ldots, x_K) = C(F_1(x_1), \ldots, F_K(x_K)).$$

Conversely, if C is a K–dimensional copula and $F_1, \ldots, F_K$ are univariate cdf , then the function F defined above is a K–dimensional cdf with marginals $F_1, \ldots, F_K$.

With these results in mind, one can verify that if F is a cdf with univariate marginals $F_1, \ldots, F_K$, and we take $u_i = F_i(x_i)$ then $F(F_1^{-1}(u_1), \ldots, F_K^{-1}(u_K))$ is a copula that guarantees the representation in the Sklar's theorem. We write then

$$C(u_1, \ldots, u_K) = F(F_1^{-1}(u_1), \ldots, F_K^{-1}(u_K)). \tag{44}$$

In this framework the main result of Sklar's theorem it is then, not the existence of a copula representation, but the fact that this representation is unique. We note also that given a particular cdf and its univariate marginals, (44) represent a direct way to find functions that are copulas.

Conversely, if in equation (44) we replace each u_i by its probability representation $F_i(x_i)$ we obtain

$$C(F_1(x_1), \ldots, F_K(x_K)) = F(x_1, \ldots, x_K) \tag{45}$$

that is exactly the formula within Sklar's theorem.

In practice, the copula of a continuous K–dimensional r.v. $\mathbf{X}$ (unique by the Sklar's theorem) is the representation of the distribution of the r.v. in the 'probability domain' (those of u_i's), rather then in the 'quantile domain' (those of x_i's, in practice the original scale). It is however clear these two ways are not equivalent: the former forget completely the information of the marginals but isolates the structure of dependence among the components of $\mathbf{X}$. This implies that in the study of a multivariate r.v. it is possible to proceed in 2 steps:

1. identification of the marginal distributions;
2. definition of the appropriate copula in order to represent the structure of dependence among the components of the r.v.

This two step analysis is emphasized also by the structure of the pdf of the K–variate r.v. $\mathbf{X}$ that follows from the copula representation (45):

$$f(x_1, \ldots, x_K) = c(F_1(x_1), \ldots, F_K(x_K)) \prod_{i=1}^{K} f_i(x_i). \tag{46}$$

where $f_i(x_i)$ is the pdf of the univariate marginals X_i and c is the copula density. It is a very interesting formulation: the product of the marginal densities, which corresponds to the multivariate density in case of independence among the X_i's, is corrected by the structure of dependence isolated by the copula's density c. We note also as (46) is a way to find functions representing copulas.

B.3 The Normal copula

Amongst the copulas presented in letterature an important role is played by the Normal copula, whose density is given by

$$c(\mathbf{u}) = |\mathbf{R}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}'(\mathbf{R}^{-1} - \mathbf{I})\mathbf{q}\right], \tag{47}$$

where $\mathbf{R}$ is a correlation matrix, $q_i = \Phi^{-1}(u_i)$, $\Phi(.)$ means the cdf of the standard Normal distribution. Normal copulas are interesting because possess interesting characteristics: the capability of capture a broad range of dependencies (the bivariate Normal copula, according to the value of the correlation parameter is capable of attaining the lower Frechèt bound, the product copula and the upper Frechèt bound), the analytical tractability, the simple simulation.

C Useful multivariate statistical distributions

C.1 Multivariate Gamma distributions

The Cheriyan and Ramabhadran multivariate Gamma (shortly $GammaCR$) is defined as follows[13]. Let $Y_0, Y_1, \ldots, Y_K$ independent random variables with the following distributions:

$$Y_0 \sim Gamma(\phi_0, 1) \qquad Y_i \sim Gamma(\phi_i - \phi_0, 1) \text{ for } i = 1, \ldots, K,$$

[13] We adapt the definition in (Johnson *et al.*, 2000, chapter 48) to our framework.

where $0 < \phi_0 < \phi_i$ for $i = 1, \ldots, K$. Then, the K–dimensional r.v. $\mathbf{X} = (X_1; \ldots; X_K)$, where

$$X_i = \frac{1}{\beta_i}(Y_i + Y_0), \tag{48}$$

and $\beta_i > 0$, has $GammaCR$ distribution with parameters ϕ_0, $\boldsymbol{\phi} = (\phi_1; \ldots; \phi_K)$ and $\boldsymbol{\beta} = (\beta_1; \ldots; \beta_K)$ that is

$$\mathbf{X} \sim GammaCR(\phi_0, \boldsymbol{\phi}, \boldsymbol{\beta}).$$

The pdf is given by

$$f(\mathbf{x}) = \frac{1}{\Gamma(\phi_0)} \prod_{i=1}^{K} \frac{\beta_i}{\Gamma(\phi_i - \phi_0)} e^{-\sum_{i=1}^{K} \beta_i x_i} \int_0^v y_0^{\phi_0 - 1} e^{(K-1)y_0} \prod_{i=1}^{K} (\beta_i x_i - y_0)^{\phi_i - \phi_0 - 1} dy_0,$$

where $v = \min(\beta_1 x_1, \ldots, \beta_K x_K)$ and the integral leads in general to very complicated expressions. Each univariate r.v. X_i has however a $Gamma(\phi_i, \beta_i)$ distribution. The main conjoint moments can be computed via formula (48); in particular the variance matrix has elements

$$C(X_i, X_j) = \frac{\phi_0}{\beta_i \beta_j}$$

and then

$$\rho(X_i, X_j) = \frac{\phi_0}{\sqrt{\phi_i \phi_j}}.$$

This implies that parameters ϕ_0 plays an important role in determining the correlation among the X_i's and that the $GammaCR$ distribution admits only positive correlations among the univariate components.

C.2 Normal copula – Gamma marginals distribution

The *Normal copula – Gamma marginals* distribution can be defined as a multivariate distribution whose univariate marginals are Gamma distributions and the structure of dependence is represented by a Normal copula. In symbols we write

$$\mathbf{X} \sim N(\mathbf{R}) - \prod_{l=1}^{K} Gamma(\phi_i, \beta_i)$$

where $\boldsymbol{\phi} = (\phi_1; \ldots; \phi_K)$, $\boldsymbol{\beta} = (\beta_1; \ldots; \beta_K)$, and $\mathbf{R}$ is a correlation matrix.

Song (2000) defines and discusses multivariate distributions whose univariate marginals are dispersion distributions and the structure of dependence is represented by a Normal copula. Since the Gamma distribution is a particular case of dispersion distribution, the Normal copula – Gamma marginals case is covered by the cited work. The author discusses various properties of multivariate dispersion models generated from Normal copula. In particular, exploiting first order approximations (as in Song (2000, pages 433–4))

and adapting notation, we obtain that the variance matrix has approximated elements

$$C(x_i, x_j) \simeq \mathbf{R}_{ij} \frac{\sqrt{\phi_i \phi_j}}{\beta_i \beta_j}$$

and then

$$\rho(x_i, x_j) \simeq \mathbf{R}_{ij}.$$

With respect to the $GammaCR$ distribution in section C.1, that considered in the present section admits negative correlations also.

D A summary on estimating functions

Estimating Functions (EFs, henceforth) are a method for making inference on the unknown parameters of a given statistical model. Even if some ideas appeared before, the whole development of the theory of EFs begins from the works of Godambe (1960) and Durbin (1960). Ever since, the method became increasingly popular, with important theoretical developments and illuminating applications in different domains: biostatistics and models for stochastic processes in particular. EFs are relatively less diffuse in the econometric literature, but the 'gap' is progressively filling up, perhaps even as thanks to some stimulating reviews (Vinod (1998), Bera *et al.* (2006)).

Broadly speaking, EFs have close links with other inferential approaches: in particular, analogies can be found with the Quasi Maximum Likelihood method (QML) and with the Generalized Method of Moment method (GMM) for just identified problems. However some other inferential procedures, as for instance M-estimation, have connections with EFs. About the relations between EFs and other inferential methods see, among others, Desmond (1997), Contreras and Ryan (2000) and the reviews cited above.

This summary aims to show how EFs can be usefully employed for making inference on time series models. In particular, we focus our attention on martingale EFs and this choice is motivated by at least three reasons. Firstly: martingale EFs are largely applicable, because in modelling time series, EFs that are martingales arise often naturally from the structure of the process. Secondly: martingale EFs have particularly nice properties and a relatively simple asymptotic theory. Thirdly: because an estimating function has some analogies with the score function, and because the score function, if it exists, is usually a martingale, it is quite natural to approximate it using families of martingales EFs.

An interesting historical review on EFs can be found in Bera *et al.* (2006), that provides also a long list of references. About the topics handled in this summary, important references can be found, among others, in Heyde (1997), Sørensen (1999), Bibby *et al.* (2004).

This summary is organized as follows: in section D.1 we list some necessary concepts about martingale processes; in section D.2 we discuss the likelihood framework as an important reference for interpreting EFs; in section D.3 we define EFs in general and martingale EFs in particular; in section D.4 we sketch the asymptotic theory of martingale

EFs; in section D.5 we consider optimal EFs; in section D.6 we show how nuisance parameters can be handled.

D.1 Notation and preliminary concepts

As stated before, we aim to illustrate how martingale EFs can be usefully employed for making inference on time series models. To this goal, we recall some concepts on martingale processes. We precede them summarizing the mathematical notation used.

D.1.1 Mathematical notation

We use the following mathematical notations:

- if $\mathbf{x} \in \mathbb{R}^K$, then $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2}$ (the Euclidean norm); if $\mathbf{x} \in \mathbb{R}^{H \times K}$ then $\|\mathbf{x}\| = \left(\sum_{i=1}^{H} \sum_{j=1}^{K} x_{ij}^2 \right)^{1/2}$ (the Frobenius norm);
- if $\mathbf{f}(\mathbf{x})$ is a vector function of $\mathbf{x} \in \mathbb{R}^K$, the matrix of its first derivatives is denoted by $\nabla_{\mathbf{x}'}\mathbf{f}(\mathbf{x})$ or $\nabla_{\mathbf{x}'}\mathbf{f}$;
- given two square, symmetric and non negative definite (nnd) matrices $\mathbf{A}$ and $\mathbf{B}$, we say $\mathbf{A}$ not greater than $\mathbf{B}$ (or $\mathbf{A}$ not smaller than $\mathbf{B}$) in the Löwner ordering if

  $$\mathbf{B} - \mathbf{A}$$

  is nnd and we write

  $$\mathbf{A} \preceq \mathbf{B};$$

- $\xrightarrow{p}$ denotes convergence in probability;
- $\overset{a}{\sim}$ denotes almost sure (with probability 1) convergence.

D.1.2 Martingales

Let $(\Omega, \mathcal{F}, P)$ a probability space and let $\{\mathbf{x}_t\}$ a discrete time ($t \in \mathbb{Z}^+$) K-dimensional time series process, that is $\mathbf{x}_t : \Omega \longrightarrow \mathbb{R}^K$. Denoting as $\mathcal{F}_t$ ($\subseteq \mathcal{F}$) the σ-algebra representing the information at time t, we recall that $\{\mathbf{x}_t\}$:

- is *predictable* if $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable; this implies

  $$E(\mathbf{x}_t|\mathcal{F}_{t-1}) = \mathbf{x}_t;$$

- is a *martingale* if $\mathbf{x}_t$ is $\mathcal{F}_t$-measurable and

$$E(\mathbf{x}_t|\mathcal{F}_{t-1}) = \mathbf{x}_{t-1};$$

- is a *martingale difference* if $\mathbf{x}_t$ is $\mathcal{F}_t$-measurable and

$$E(\mathbf{x}_t|\mathcal{F}_{t-1}) = \mathbf{0}.$$

Using a Dobb's representation, a martingale $\{\mathbf{x}_t\}$ can be always represented as

$$\mathbf{x}_t = \mathbf{x}_0 + \sum_{s=1}^{t} \Delta\mathbf{x}_s, \tag{49}$$

where

$$\Delta\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1} \tag{50}$$

is a martingale difference and $\mathbf{x}_0$ is the starting r.v. A martingale $\{\mathbf{x}_t\}$ is *null at* 0 if $\mathbf{x}_0 = \mathbf{0}$. In this case (49) becomes

$$\mathbf{x}_t = \sum_{s=1}^{t} \Delta\mathbf{x}_s \tag{51}$$

and $E\left(\mathbf{x}_t|\mathcal{F}_{t-1}\right) = \mathbf{0}$ for all t. A martingale $\{\mathbf{x}_t\}$ null at 0 is *square integrable* if $\|V\left(\mathbf{x}_t\right)\| < \infty$ for all t.

Taking into account, from now on, square integrable martingales null at 0, and considering their representation as in (51)-(50), the following quantities can be defined:

- $\langle \mathbf{x}, \mathbf{y} \rangle_t$, the *mutual quadratic characteristic* between $\mathbf{x}_t$ and $\mathbf{y}_t$:

$$\langle \mathbf{x}, \mathbf{y} \rangle_t = \sum_{s=1}^{t} C(\Delta\mathbf{x}_s, \Delta\mathbf{y}_s|\mathcal{F}_{s-1});$$

- $\langle \mathbf{x} \rangle_T$, the *quadratic characteristic* of $\mathbf{x}_t$:

$$\langle \mathbf{x} \rangle_T = \sum_{s=1}^{t} V(\Delta\mathbf{x}_s|\mathcal{F}_{s-1}).$$

Furthermore, if a martingale $\{\mathbf{x}_t\}$ null at 0 depends on a parameter $\boldsymbol{\lambda} \in \mathbb{R}^p$, we define the *compensator* of $\nabla_{\boldsymbol{\lambda}'}\mathbf{x}_t$ as

$$\overline{\mathbf{x}}_t = \sum_{s=1}^{t} E\left(\nabla_{\boldsymbol{\lambda}'}\Delta\mathbf{x}_s|\mathcal{F}_{s-1}\right).$$

D.2 Set-up and the likelihood framework

The likelihood framework provides an important reference for the theory of EFs. In fact, even if they can be usefully employed just in situations where the likelihood is unknown or is difficult to compute, EFs generalize the score function in many senses and share many of its properties. Moreover, the score function plays an important role in many aspects of the EFs theory.

Hence, we consider a discrete time K-dimensional time series process $\{\mathbf{x}_t\}$ and we denote as $\mathbf{x}_{(T)} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ a sample from $\{\mathbf{x}_t\}$. We assume that:

- the possible probability measures for $\mathbf{x}_{(T)}$ are $\mathcal{P} = \{P_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, a union of families of parametric models indexed by $\boldsymbol{\lambda}$, a p-dimensional parameter belonging to an open set $\boldsymbol{\Lambda} \subseteq \mathbb{R}^p$;
- each $(\Omega, \mathcal{F}, P_{\boldsymbol{\lambda}})$, where $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, is a complete probability space;
- each $P_{\boldsymbol{\lambda}}$, where $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, is absolutely continuous with respect to some σ-finite measure. In this situation a likelihood is defined (even if it can be unknown).

We use the following notation:

- $L_{(T)}(\boldsymbol{\lambda}) = f(\mathbf{x}_{(T)}; \boldsymbol{\lambda})$ is the *likelihood function* (the density of $\mathbf{x}_{(T)}$);
- $l_{(T)}(\boldsymbol{\lambda}) = \ln L_T(\boldsymbol{\lambda})$ is the *log-likelihood function*;
- $\mathbf{s}_{(T)}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}'} l_{(T)}(\boldsymbol{\lambda})$ (provided that $l_{(T)}(\boldsymbol{\lambda}) \in \mathcal{C}^1(\boldsymbol{\Lambda})$ a.s.) is the *score function*.

Usually, the likelihood function for time series samples can be expressed as product of the conditional densities $L_t(\boldsymbol{\lambda}) = f(\mathbf{x}_t | \mathcal{F}_{t-1}; \boldsymbol{\lambda})$ of the single observations in the sample. In this case

$$
\begin{aligned}
L_{(T)}(\boldsymbol{\lambda}) &= \prod_{t=1}^{T} L_t(\boldsymbol{\lambda}) \\
l_{(T)}(\boldsymbol{\lambda}) &= \sum_{t=1}^{T} l_t(\boldsymbol{\lambda}) \\
\mathbf{s}_{(T)}(\boldsymbol{\lambda}) &= \sum_{t=1}^{T} \mathbf{s}_t(\boldsymbol{\lambda})
\end{aligned} \tag{52}
$$

where $l_t(\boldsymbol{\lambda}) = \ln L_t(\boldsymbol{\lambda})$ and $\mathbf{s}_t(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}'} l_t(\boldsymbol{\lambda})$.

We assume that the score function $\mathbf{s}_{(T)}$[14] satisfies the usual regularity conditions: for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, and for all $T \in \mathbb{Z}^+$

1. $\mathbf{s}_{(T)}$ is a square integrable martingale null at 0;

[14] In many circumstances we suppress explicit dependence from the parameter $\boldsymbol{\lambda}$.

2. $\mathbf{s}_{(T)}$ is regular, in the sense that $\mathbf{s}_{(T)} \in \mathcal{C}^1(\mathbf{\Lambda})$ a.s.;
3. $\mathbf{s}_{(T)}$ is smooth, in the sense that differentiation and integration can be interchanged in differencing expectations of $\mathbf{s}_t$ with respect to $\boldsymbol{\lambda}$.

Under above regularity conditions, the ML estimator is the solution of the score equation

$$\mathbf{s}_{(T)}(\boldsymbol{\lambda}) = \mathbf{0}.$$

Using the notation of section D.1, the *Fisher information* is defined by

$$\mathbf{I}_T = \langle \mathbf{s} \rangle_T = -\bar{\mathbf{s}}_T,$$

where the second equality is named *second Bartlett identity*. We can check easily that $\bar{\mathbf{s}}_T$ and $\langle \mathbf{s} \rangle_T$ are linked, respectively, to the sum of the Hessians and to the outer product of the gradients of l_t's.

D.3 Martingale estimating functions

Within above framework, an *estimating function* (EF) for $\boldsymbol{\lambda}$ is a p-dimensional function of the parameter $\boldsymbol{\lambda}$ and of the sample $\mathbf{x}_{(T)}$:

$$\mathbf{g}_{(T)}(\boldsymbol{\lambda}; \mathbf{x}_{(T)}).$$

Usually we suppress explicit dependence from the observations and sometimes from the parameter also. In these cases we indicate compactly the EF as

$$\mathbf{g}_{(T)}(\boldsymbol{\lambda}) \quad \text{or} \quad \mathbf{g}_{(T)}.$$

An estimate for $\boldsymbol{\lambda}$, is the value $\widehat{\boldsymbol{\lambda}}_T$ that solves the corresponding *estimating equation* (EE)

$$\mathbf{g}_{(T)}(\boldsymbol{\lambda}) = \mathbf{0}. \tag{53}$$

The score function is then a particular case of estimating function. However, there is a remarkable difference between the score function and a generic EF. In fact $\mathbf{s}_{(T)}(\boldsymbol{\lambda})$ has $l_{(T)}(\boldsymbol{\lambda})$ as its potential function, since $\mathbf{s}_{(T)}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}'} l_{(T)}(\boldsymbol{\lambda})$. On the contrary, in general an EF may not possess a potential function $G_{(T)}(\boldsymbol{\lambda})$ such that $\mathbf{g}_{(T)}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}'} G_{(T)}(\boldsymbol{\lambda})$ (Knudsen (1999, p. 2)).

As stated in Heyde (1997, p. 26), since the score function, if it exists, is usually a martingale, it is quite natural to approximate it using families of martingales EFs. Furthermore, in modelling time series processes, EFs that are martingales arise often naturally from the structure of the process. In this spirit, we stress that many of the concepts discussed in this section are particular to martingale EFs. For a more general handling see, among others, Sørensen (1999).

Hence we consider martingale EFs that are square integrable and null at 0. As in (51), we

represent them as sums *martingale differences* $\mathbf{g}_t$,

$$\mathbf{g}_{(T)} = \sum_{t=1}^{T} \mathbf{g}_t$$

where $\mathbf{g}_t = \mathbf{g}_{(T)} - \mathbf{g}_{(T-1)}$. We denote the *quadratic characteristic* of $\mathbf{g}_{(T)}$ as

$$\langle \mathbf{g} \rangle_T = \sum_{t=1}^{T} V(\mathbf{g}_t | \mathcal{F}_{t-1})$$

and the *compensator* of $\nabla_{\boldsymbol{\lambda}'} \mathbf{g}_{(T)}$ as

$$\overline{\mathbf{g}}_T = \sum_{t=1}^{T} E\left(\nabla_{\boldsymbol{\lambda}'} \mathbf{g}_t | \mathcal{F}_{t-1}\right).$$

As in the likelihood theory, some regularity conditions on EFs $\mathbf{g}_{(T)}(\boldsymbol{\lambda})$ are usually imposed. In particular it is required that, $\forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda}$,

1. $\mathbf{g}_{(T)}$ is a *square integrable martingale null at* 0;
2. $\mathbf{g}_{(T)}$ is *regular*, in the sense that $\mathbf{g}_{(T)} \in \mathcal{C}^1(\boldsymbol{\Lambda})$ a.s. and $\overline{\mathbf{g}}_T$ is non-singular;
3. $\mathbf{g}_{(T)}$ is *smooth*, in the sense that differentiation and integration can be interchanged in differencing expectations of $\mathbf{g}_t$ with respect to $\boldsymbol{\lambda}$.

(We remark the analogies with the corresponding regularity conditions for the score function in section D.2). We denote as $\mathcal{M}_T^{(1)}$ the family of estimating functions satisfying conditions 1, 2, 3 above and we name it as the family of *regular and smooth* EFs.

Regular and smooth EFs satisfy some interesting properties. Among these (details in Knudsen (1999, ch. 1), Heyde (1997, ch. 2)):

- The compensator of $\nabla_{\boldsymbol{\lambda}'} \mathbf{g}_{(T)}$ is strictly linked to the mutual quadratic characteristic between the EF and the score function, since

  $$\overline{\mathbf{g}}_T = -\langle \mathbf{g}, \mathbf{s} \rangle_T. \tag{54}$$

  This follows from the relation

  $$E\left(\nabla_{\boldsymbol{\lambda}'} \mathbf{g}_t | \mathcal{F}_{t-1}\right) = -C\left(\mathbf{g}_t, \mathbf{s}_t | \mathcal{F}_{t-1}\right). \tag{55}$$

- $\overline{\mathbf{g}}_T$ non-singular (see the regularity condition 2 above) implies $\langle \mathbf{g} \rangle_T$ positive-definite. This implication can be obtained representing $\overline{\mathbf{g}}_T$ as in (54) and using considerations similar to Knudsen (1999, p. 4).

- If $\mathbf{A}(\boldsymbol{\lambda})$ is a (p, p) non-random, non-singular matrix belonging to $C^1(\boldsymbol{\Lambda})$ a.s., then also the linear transformation
$$\mathbf{A}(\boldsymbol{\lambda})\mathbf{g}_{(T)}(\boldsymbol{\lambda}) \tag{56}$$
is an EF belonging to $\mathcal{M}_T^{(1)}$.
- If $\boldsymbol{\xi}(\boldsymbol{\lambda})$ is a smooth 1-to-1 reparameterization, then also $\mathbf{g}_{(T)}(\boldsymbol{\lambda}(\boldsymbol{\xi}))$ is a regular and smooth EF for $\boldsymbol{\xi}$ (Knudsen (1999, p. 7)). In practice, invariance and smoothness conditions are *invariant* under smooth 1-to-1 reparameterizations.

Using (56), from any EF $\mathbf{g}_{(T)} \in \mathcal{M}_T^{(1)}$ we can derive another EF

$$\mathbf{g}_{(T)}^{(s)} = -\overline{\mathbf{g}}_T' \langle \mathbf{g} \rangle_T^{-1} \mathbf{g}_{(T)},$$

called *standardized version* of $\mathbf{g}_{(T)}$. $\mathbf{g}_{(T)}^{(s)}$ assumes a special role: as argued in section D.4, $\mathbf{g}_{(T)}^{(s)}$ and $\mathbf{g}_{(T)}$ are *equivalent*, in the sense that they produce the same estimator with the same asymptotic variance, but $\mathbf{g}_{(T)}^{(s)}$ is more directly comparable with the score function.

D.4 Asymptotic theory

A central part of the theory is devoted to find EFs that produce good, and possibly optimal, estimators. In this work, however, we focused attention to martingale EFs: the optimality criteria employed in this context have essentially an asymptotic justification. Following these motivations, in this section we sketch the asymptotic theory concerning martingale EFs. More details can be found, among others, in Heyde (1997), Sørensen (1999), Bibby *et al.* (2004).

In generic terms, the asymptotic theory of martingale EFs moves from a Taylor expansion of the EF in a neighborhood of the true parameter value and the consequent application of some law of large numbers (LLN) and central limit theorem (CLT) for square integrable martingales.

Hence, we consider the following Taylor expansion of $\mathbf{g}_{(T)}(\widehat{\boldsymbol{\lambda}}_T) = \mathbf{0}$ around the true parameter value $\boldsymbol{\lambda}_0$:

$$\mathbf{0} = \mathbf{g}_{(T)}(\widehat{\boldsymbol{\lambda}}_T) = \mathbf{g}_{(T)}(\boldsymbol{\lambda}_0) + \nabla_{\boldsymbol{\lambda}'}\mathbf{g}_{(T)}(\boldsymbol{\lambda}^*)(\widehat{\boldsymbol{\lambda}}_T - \boldsymbol{\lambda}_0), \tag{57}$$

where $\boldsymbol{\lambda}^* = \alpha\boldsymbol{\lambda}_0 + (1-\alpha)\widehat{\boldsymbol{\lambda}}_T$ with $\alpha \in (0, 1)$. Adding some other regularity conditions (they are useful for inverting $\nabla_{\boldsymbol{\lambda}'}\mathbf{g}_{(T)}(\boldsymbol{\lambda}^*)$, for approximating it by $\nabla_{\boldsymbol{\lambda}'}\mathbf{g}_{(T)}(\boldsymbol{\lambda}_0)$ and for convergence of this last quantity to the compensator $\overline{\mathbf{g}}_T(\boldsymbol{\lambda}_0)$ – details in Heyde (1997, ch. 2)) we have

$$\widehat{\boldsymbol{\lambda}}_T - \boldsymbol{\lambda}_0 \simeq -\overline{\mathbf{g}}_T(\boldsymbol{\lambda}_0)^{-1}\mathbf{g}_{(T)}(\boldsymbol{\lambda}_0). \tag{58}$$

Under mild regularity conditions a LLN,

$$\overline{\mathbf{g}}_T^{-1}\mathbf{g}_{(T)} \xrightarrow{p} \mathbf{0}, \tag{59}$$

and a CLT,

$$\langle \mathbf{g} \rangle_T^{-1/2}\mathbf{g}_{(T)} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p), \tag{60}$$

apply. Combined with (58), these imply

$$\widehat{\boldsymbol{\lambda}}_T \xrightarrow{p} \boldsymbol{\lambda}_0 \tag{61}$$

and

$$\langle \mathbf{g} \rangle_T(\boldsymbol{\lambda}_0)^{-1/2}\overline{\mathbf{g}}_T(\boldsymbol{\lambda}_0)(\widehat{\boldsymbol{\lambda}}_T - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p). \tag{62}$$

We call $\mathcal{M}_T^{(2)} \subseteq \mathcal{M}_T^{(1)}$ the class of martingale EFs $\mathbf{g}_{(T)}$ that belong to $\mathcal{M}_T^{(1)}$ and satisfy the additional regularity conditions needed for (61) and (62). Summarizing above results, EFs belonging to $\mathcal{M}_T^{(2)}$ give consistent and asymptotically normal estimators with asymptotic variance matrix

$$V_\infty\left(\widehat{\boldsymbol{\lambda}}_T\right) = \overline{\mathbf{g}}_T(\boldsymbol{\lambda}_0)^{-1\prime}\langle \mathbf{g} \rangle_T(\boldsymbol{\lambda}_0)\overline{\mathbf{g}}_T(\boldsymbol{\lambda}_0)^{-1} = \mathbf{J}_T(\boldsymbol{\lambda}_0)^{-1}. \tag{63}$$

The heuristic meaning of this result is that a small asymptotic variance requires EFs with 'small' variability $\langle \mathbf{g} \rangle_T$ and 'large' sensitivity $\overline{\mathbf{g}}_T$. The random matrix

$$\mathbf{J}_T = \overline{\mathbf{g}}_T'\langle \mathbf{g} \rangle_T^{-1}\overline{\mathbf{g}}_T \tag{64}$$

can be interpreted as an information matrix: it is called *martingale information* (Heyde, 1997, p. 28) or *Heyde information* (Bibby *et al.*, 2004, p. 30).

Now, let us return to the standardized version of $\mathbf{g}_{(T)}$, that is

$$\mathbf{g}_{(T)}^{(s)} = -\overline{\mathbf{g}}_T'\langle \mathbf{g} \rangle_T^{-1}\mathbf{g}_{(T)}. \tag{65}$$

We can check easily that, for EF belonging to $\mathcal{M}_T^{(2)}$, $\mathbf{g}_{(T)}$ and $\mathbf{g}_{(T)}^{(s)}$ are *equivalent*: they produce the same estimator and have $\mathbf{J}(\mathbf{g}^{(s)})_T = \mathbf{J}(\mathbf{g})_T$, that is they have the same Heyde information. However (65) is more directly comparable with the score function than $\mathbf{g}_{(T)}$. In fact, replacing (54) into (65), we can check that $\mathbf{g}_{(T)}^{(s)}$ can be interpreted as the orthogonal projection of the score function along the direction of $\mathbf{g}_{(T)}$. This means that, for any $\mathbf{g}_{(T)} \in \mathcal{M}_T^{(2)}$, $\mathbf{g}_{(T)}^{(s)}$ is the version of $\mathbf{g}_{(T)}$ "closest" to the score function, in the sense that

$$\left(\mathbf{s}_{(T)} - \mathbf{g}_{(T)}^{(s)}\right)\left(\mathbf{s}_{(T)} - \mathbf{g}_{(T)}^{(s)}\right)' \preceq \left(\mathbf{s}_{(T)} - \mathbf{A}\mathbf{g}_{(T)}\right)\left(\mathbf{s}_{(T)} - \mathbf{A}\mathbf{g}_{(T)}\right)'$$

for any linear transformation of $\mathbf{g}_{(T)}$ as in (56). Moreover, we can check that $\mathbf{g}_{(T)}^{(s)}$ satisfies a second Bartlett identity analogous to that of the score function:

$$\mathbf{J}(\overline{\mathbf{g}})_T = \mathbf{J}(\overline{\mathbf{g}}^{(s)})_T = \langle \mathbf{g}^{(s)} \rangle_T = -\overline{\mathbf{g}^{(s)}}_T. \tag{66}$$

D.5 Optimal estimating functions

The interpretation provided in section D.4 of the Heyde information and its close link with the asymptotic variance of $\widehat{\boldsymbol{\lambda}}_T$ suggest the criterion usually employed to choose optimal EF in the asymptotic sense: "maximize" the Heyde information. More precisely, once selected a subfamily $\mathcal{M}_T \subseteq \mathcal{M}_T^{(2)}$ of EFs we have:

Definition 1 (**A-optimality**) $\mathbf{g}^*_{(T)}$ *is* A-optimal *(optimal in the asymptotic sense) into* $\mathcal{M}_T \subseteq \mathcal{M}_T^{(2)}$ *if*

$$\mathbf{J}\left(\mathbf{g}^*\right)_T \succeq \mathbf{J}\left(\mathbf{g}\right)_T$$

a.s. for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$*, for all* $\mathbf{g}_{(T)} \in \mathcal{M}_T$ *and for all* $T \in \mathbb{Z}^+$*. An A-optimal estimating function* $\mathbf{g}^*_{(T)}$ *is called a* quasi score function*, whereas the corresponding estimator* $\widehat{\boldsymbol{\lambda}}_T$ *is called a* quasi likelihood estimator[15].

Returning for a moment to the Heyde information, we already underlined that $\mathbf{J}(\mathbf{g})_T$ equates the quadratic characteristic of the *standardized* version $\mathbf{g}^{(s)}_{(T)}$ (see (66)). Very interestingly, it can be shown that, when the score function exists, A-optimality as defined above is equivalent to find into $\mathcal{M}_T$ the standardized EF $\mathbf{g}^{*(s)}_{(T)}$ "closest" to the score $\mathbf{s}_{(T)}$, in the sense that

$$E\left[\left(\mathbf{s}_{(T)} - \mathbf{g}^{*(s)}_{(T)}\right)\left(\mathbf{s}_{(T)} - \mathbf{g}^{*(s)}_{(T)}\right)'\right] \preceq E\left[\left(\mathbf{s}_{(T)} - \mathbf{g}^{(s)}_{(T)}\right)\left(\mathbf{s}_{(T)} - \mathbf{g}^{(s)}_{(T)}\right)'\right] \tag{67}$$

for any other $\mathbf{g}^{(s)}_{(T)}$ into $\mathcal{M}_T$. (67) reveals another important result of the EFs theory: if $\mathcal{M}_T$ encloses $\mathbf{s}_{(T)}$, just the score function is asymptotically optimal with respect to any other EF $\mathbf{g}_{(T)}$.

As stated in Heyde (1997), the criterion in the of A-optimality definition or the equivalent formulation (67) are hard to apply directly. In practice the following theorem can be employed:

Theorem 2 (Heyde (1997, p. 29)) *If* $\mathbf{g}^*_{(T)} \in \mathcal{M}_T \subseteq \mathcal{M}_T^{(2)}$ *satisfies*

$$\overline{\mathbf{g}}_T^{*-1}\langle \mathbf{g}^* \rangle_T = \overline{\mathbf{g}}_T^{-1}\langle \mathbf{g}, \mathbf{g}^* \rangle_T, \tag{68}$$

for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$*, for all* $\mathbf{g}_{(T)} \in \mathcal{M}_T$ *and for all* $T \in \mathbb{Z}^+$*, then* $\mathbf{g}^*_{(T)}$ *is A-optimal in* $\mathcal{M}_T^{(2)}$*.*

Conversely, if $\mathcal{M}_T$ *is closed under addition and* $\mathbf{g}^*_{(T)}$ *is A-optimal in* $\mathcal{M}_T$ *then (68) holds.*

As we always work on subsets of $\mathcal{M}_T^{(2)}$, a crucial point in applications is to make a good choice of the set $\mathcal{M}_T$ of the EFs considered. About this, an important family of EFs that

[15] Despite a quasi likelihood function in the sense of potential function of $\mathbf{g}_{(T)}$ may not exist (see section D.3).

often can be usefully employed in time series processes is

$$\mathcal{M}_T = \left\{ \mathbf{g}_{(T)} \in \mathcal{M}_T^{(2)} : \mathbf{g}_{(T)}(\boldsymbol{\lambda}) = \sum_{t=1}^{T} \boldsymbol{\alpha}_t(\boldsymbol{\lambda}) \mathbf{v}_t(\boldsymbol{\lambda}) \right\},$$

where $\mathbf{v}_t(\boldsymbol{\lambda})$ is a K-dimensional martingale difference and $\boldsymbol{\alpha}_t(\boldsymbol{\lambda})$ is a (p, K) $\mathcal{F}_{t-1}$-measurable function. Tacking

$$\boldsymbol{\alpha}_t^* = -E\left(\nabla_{\boldsymbol{\lambda}} \mathbf{v}_t' | \mathcal{F}_{t-1}\right) V\left(\mathbf{v}_t | \mathcal{F}_{t-1}\right)^{-1}$$

we can check immediately that condition (68) of theorem 2 is satisfied and then

$$\mathbf{g}_{(T)}^* = -\sum_{t=1}^{T} E\left(\nabla_{\boldsymbol{\lambda}} \mathbf{v}_t' | \mathcal{F}_{t-1}\right) V\left(\mathbf{v}_t | \mathcal{F}_{t-1}\right)^{-1} \mathbf{v}_t \tag{69}$$

is A-optimal into $\mathcal{M}_T$. Furthermore, since $\overline{\mathbf{g}}_T^{*-1} \langle \mathbf{g}^* \rangle_T = -\mathbf{I}_p$ its Heyde information is given by

$$J\left(\mathbf{g}^*\right)_T = \langle \mathbf{g}^* \rangle_T = -\overline{\mathbf{g}}_T^* = \sum_{t=1}^{T} E\left(\nabla_{\boldsymbol{\lambda}} \mathbf{v}_t' | \mathcal{F}_{t-1}\right) V\left(\mathbf{v}_t | \mathcal{F}_{t-1}\right)^{-1} E\left(\nabla_{\boldsymbol{\lambda}'} \mathbf{v}_t | \mathcal{F}_{t-1}\right). \tag{70}$$

(69) is known as the *Hutton-Nelson quasi score function*. We note that when a martingale difference $\mathbf{v}_t(\boldsymbol{\lambda})$ can be defined from the model under analysis, the Hutton-Nelson solution provides a powerful and, at the same time, surprisingly simple tool: (69) and (70), the fundamental quantities for the inference, depend only on the conditional variance and on the conditional expectation of the first derivative of $\mathbf{v}_t(\boldsymbol{\lambda})$. See Heyde (1997, sect. 2.6) for a deeper discussion.

D.6 Estimating functions in presence of nuisance parameters

Sometimes, the p-dimensional parameter $\boldsymbol{\lambda}$ that appears in the likelihood can be partitioned as $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\psi})$, where:

- $\boldsymbol{\theta}$ is a p_1-dimensional parameter of scientific interest called *interest parameter*;
- $\boldsymbol{\psi}$ is a p_2-dimensional ($p_2 = p - p_1$) parameter which is not of scientific interest and is called *nuisance parameter*.

In this case the score function for $\boldsymbol{\lambda}$ can be partitioned as

$$\mathbf{s}_{(T)}(\boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{s}_{(T)1} \\ \mathbf{s}_{(T)2} \end{pmatrix},$$

where $\mathbf{s}_{(T)1} = \nabla_{\boldsymbol{\theta}'} l_{(T)}$ and $\mathbf{s}_{(T)2} = \nabla_{\boldsymbol{\psi}'} l_{(T)}$ are *marginal score functions* for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively.

This structure can be replicated if instead of a score function we consider an EF

$$\mathbf{g}_{(T)}(\boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{g}_{(T)1} \\ \mathbf{g}_{(T)2} \end{pmatrix},$$

where $\mathbf{g}_{(T)1}$ and $\mathbf{g}_{(T)2}$ are the *marginal estimating functions* for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. In particular, $\mathbf{g}_{(T)1}$ is meant for estimating $\boldsymbol{\theta}$ when $\boldsymbol{\psi}$ is known whereas $\mathbf{g}_{(T)2}$ is meant for estimating $\boldsymbol{\psi}$ when $\boldsymbol{\theta}$ is known.

A clear discussion of inferential issues with nuisance parameters can be found, among others, in Liang and Zeger (1995). The main inferential problem is that, in presence of nuisance parameters, some properties of estimating functions are no longer valid if we replace parameters with the corresponding estimators. For instance, unbiasedness of $\mathbf{g}_{(T)1}$ is not guaranteed if $\boldsymbol{\psi}$ is replaced by an estimator $\widehat{\boldsymbol{\psi}}$. By consequence, optimality properties also are not guaranteed.

An interesting statistical handling of nuisance parameter in the estimating functions framework is provided in Jørgensen and Knudsen (2004) (see also Knudsen (1999)). Their handling parallels, in some aspects, the notion of *Fisher-orthogonality* (*F-orthogonality*, henceforth) in ML estimation. F-orthogonality is defined by block diagonality of the Fisher information matrix for $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\psi})$. This particular structure guarantees the following properties of the ML estimator:

1. asymptotic independence of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$;
2. efficiency-stable estimation of $\boldsymbol{\theta}$, in the sense that the asymptotic variance for $\boldsymbol{\theta}$ is the same whether $\boldsymbol{\psi}$ is treated as known or unknown;
3. simplification of the estimation algorithm;
4. $\widehat{\boldsymbol{\theta}}(\boldsymbol{\psi})$, the estimate of $\boldsymbol{\theta}$ when $\boldsymbol{\psi}$ is given, varies only slowly with $\boldsymbol{\psi}$.

Among these, Jørgensen and Knudsen (2004) state that efficiency-stable estimation is the crucial property.

Starting from this point, Jørgensen and Knudsen (2004) extend F-orthogonality to EFs introducing the concept of *nuisance parameter insensitivity* (NPI). Even if these authors define and employ this concept within a finite sample optimality framework, NPI can be easily extended asymptotic optimality framework. We discuss this in the following. See the above reference for the original treatment.

We start our exposition partitioning the compensator $\overline{\mathbf{g}}_T$ and the quadratic characteristic $\langle \mathbf{g} \rangle_T$ as

$$\overline{\mathbf{g}}_T = \begin{pmatrix} \overline{\mathbf{g}}_{T,11} & \overline{\mathbf{g}}_{T,12} \\ \overline{\mathbf{g}}_{T,21} & \overline{\mathbf{g}}_{T,22} \end{pmatrix} \qquad \langle \mathbf{g} \rangle_T = \begin{pmatrix} \langle \mathbf{g} \rangle_{T,11} & \langle \mathbf{g} \rangle_{T,12} \\ \langle \mathbf{g} \rangle_{T,21} & \langle \mathbf{g} \rangle_{T,22} \end{pmatrix} \tag{71}$$

conformably to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Using this partition, the information $\mathbf{J}_T$ and the asymptotic variance matrix $\mathbf{J}_T^{-1}$ have a block structure

$$\mathbf{J}_T = \begin{pmatrix} \mathbf{J}_{T,11} & \mathbf{J}_{T,12} \\ \mathbf{J}_{T,21} & \mathbf{J}_{T,22} \end{pmatrix} \qquad \mathbf{J}_T^{-1} = \begin{pmatrix} [\mathbf{J}_T^{-1}]_{11} & [\mathbf{J}_T^{-1}]_{12} \\ [\mathbf{J}_T^{-1}]_{21} & [\mathbf{J}_T^{-1}]_{22} \end{pmatrix} \tag{72}$$

whose components can be derived from (71).

Definition 2 (**Nuisance parameter insensitivity**) *The marginal EF* $\mathbf{g}_{(T)1} \in \mathcal{M}_T^{(2)}$[16] *is NPI or* $\boldsymbol{\psi}$-insensitive *if*

$$\overline{\mathbf{g}}_{T,12} = 0.$$

Using block matrix algebra, we can check easily that $\boldsymbol{\psi}$-insensitivity of $\mathbf{g}_{(T)1}$ is a sufficient condition for

$$[\mathbf{J}_T^{-1}]_{11} = \overline{\mathbf{g}}_{T,11}'^{-1} \langle \mathbf{g} \rangle_{T,11} \overline{\mathbf{g}}_{T,11}^{-1}. \tag{73}$$

(In practice, this result represents the *only if* part of the *Insensitivity Theorem* – see theorem 3 below). It implies that, under $\boldsymbol{\psi}$-insensitivity, the asymptotic variance of $\boldsymbol{\theta}$ can be computed using the compensator and the quadratic characteristic of $\mathbf{g}_{(T)1}$ only (see property 2 above).

Another important consequence of NPI is that $\widehat{\boldsymbol{\theta}}(\boldsymbol{\psi})$, the estimate of $\boldsymbol{\theta}$ when $\boldsymbol{\psi}$ is given, varies only slowly with $\boldsymbol{\psi}$ (property 4 above). In fact, as argued in Jørgensen and Knudsen (2004, p. 97), if $\|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|$ is $O_p(v^{-1/2})$ then $\|\widehat{\boldsymbol{\theta}}(\boldsymbol{\psi}) - \boldsymbol{\theta}\|$ is $O_p(v^{-1})$ if $\mathbf{g}_{(T)1}$ is $\boldsymbol{\psi}$-insensitive, otherwise is $O_p(v^{-1/2})$.

Sensitivity can be removed by projection. In fact we can check easily that the marginal EF

$$\mathbf{g}^*_{(T)1} = \mathbf{g}_{(T)1} - \overline{\mathbf{g}}_{T,12} \overline{\mathbf{g}}_{T,22}^{-1} \mathbf{g}_{(T)2} \tag{74}$$

is $\boldsymbol{\psi}$-insensitive. In the Hilbert space sense, (74) is the projection of $\mathbf{g}_{(T)2}$ onto the orthocomplement of $\mathbf{s}_{(T)2}$. This means that $\mathbf{g}^*_{(T)1}$ carries the same information about $\boldsymbol{\theta}$ as $\mathbf{g}_{(T)1}$ and $\mathbf{g}_{(T)2}$ together. See Jørgensen and Knudsen (2004, p. 101) for a deeper discussion.

We think however that, from a theoretical point of view, the more interesting result of the cited paper is that NPI is not only sufficient but also necessary for (73). This result has been condensed by Jørgensen and Knudsen (2004) in the *Insensitivity Theorem*. We provide it below, adjusting the formulation to the asymptotic framework considered here. The theorem can be proved, with minor adjustments, as in the cited paper.

Theorem 3 ***(Insensitivity Theorem)*** *The marginal EF* $\mathbf{g}_{(T)1} \in \mathcal{M}_T^{(2)}$ *is* $\boldsymbol{\psi}$*-insensitive if and only if*

$$[\mathbf{J}_T^{-1}]_{11} = \overline{\mathbf{g}}_{T,11}'^{-1} \langle \mathbf{g} \rangle_{T,11} \overline{\mathbf{g}}_{T,11}^{-1}$$

for all regular marginal EFs $\mathbf{g}_{(T)2} \in \mathcal{M}_T^{(2)}$.

[16] Strictly speaking, the marginal EF $\mathbf{g}_{(T)1}$ does not belong $\mathcal{M}_T^{(2)}$. In fact, $\mathbf{g}_{(T)1}$ has dimension p_1 whereas members of $\mathcal{M}_T^{(2)}$ are p-dimensional functions. However, we use $\mathcal{M}_T^{(2)}$ also for EFs with different dimension, meaning that they have to satisfy the corresponding regularity conditions.

E Mathematical Appendix

We prove sufficient conditions for nonnegativity of the components of $\boldsymbol{\mu}_t$.

Proposition 1 *The relation*

$$\sum_{i=1}^{n} \left(a_i x_i^2 + b_i x_i\right) + c \geq 0 \tag{75}$$

is satisfied for all $x_i \geq 0$ $(i = 1, \ldots, n)$ if and only if the coefficients a_i, b_i and c satisfy all the following conditions:

1. $a_i \geq 0$ *for all* $i \in S_n$*;*
2. $b_i \geq 0$ *for all* $i \in S_n$ *such that* $a_i = 0$*;*
3. $c - \dfrac{1}{4} \displaystyle\sum_{i=1}^{n} \dfrac{b_i^2}{a_i} \mathrm{I}(b_i < 0)\, \mathrm{I}(a_i > 0) \geq 0.$

where $S_n = \{1, \ldots, n\}$.

Proof:

The minimum of

$$\sum_{i=1}^{n} (a_i x_i^2 + b_i x_i) + c \tag{76}$$

with respect to the x_i's is simply the sum of c and the minima of the additive quantities $(a_i x_i^2 + b_i x_i)$. Hence, considering assumption $x_i \geq 0$:

1. if $a_i < 0$, the minimum of $(a_i x_i^2 + b_i x_i)$ is always $-\infty$;
2. if $a_i = 0$, the minimum of $(a_i x_i^2 + b_i x_i) = b_i x_i$ is nonnegative (0) only if $b_i \geq 0$;
3. if $a_i > 0$, the minimum of $(a_i x_i^2 + b_i x_i)$ for $x_i \geq 0$ is reached for $x_i = -\dfrac{b_i}{2a_i} I(b_i < 0)$ and is given by $-\dfrac{b_i^2}{4a_i} I(b_i < 0)$.

Requiring that (76) has a nonnegative minimum, from the above conditions those in the proposition follow immediately.

□

As a corollary of the previous proposition, we prove sufficient conditions for nonnegativity of the components of $\boldsymbol{\mu}_t$ in the vector MEM. For sake of generality, we give the result for the general formulation in which more lags are included in the $\boldsymbol{\mu}$ structure.

Corollary 1 *Let*

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \sum_{l=1}^{L} \left[\boldsymbol{\beta}_l \boldsymbol{\mu}_{t-l} + \boldsymbol{\alpha}_l \mathbf{x}_{t-l} + \boldsymbol{\gamma}_l \mathbf{x}_{t-l}^{(-)} + \boldsymbol{\delta}_l \mathbf{x}_{t-l}^{(s)} \right], \tag{77}$$

the equation that describes the evolution of $\boldsymbol{\mu}_t$ in the vector MEM of section 3.1, where:

- $\mathbf{x}_t$, $\boldsymbol{\mu}_t$ *($t = 1, \ldots, T$) and* $\boldsymbol{\omega}$ *are $(K, 1)$-vectors;*
- $\boldsymbol{\beta}_l$, $\boldsymbol{\alpha}_l$, $\boldsymbol{\gamma}_l$ *and* $\boldsymbol{\delta}_l$ *($l = 1, \ldots, L$) are (K, K) matrices;*
- *in the 'enriched' formulation:* $\mathbf{x}_t^{(-)} = \mathbf{x}_t \, \mathrm{I}(r_t < 0)$, $\mathbf{x}_t^{(s)} = \mathbf{x}_t^{1/2} \operatorname{sign}(r_t)$*;*
- *in the 'contagion' formulation:* $\mathbf{x}_t^{(-)} = \mathbf{x}_t \odot \mathrm{I}(\mathbf{r}_t < 0)$, $\mathbf{x}_t^{(s)} = \mathbf{x}_t^{1/2} \odot \operatorname{sign}(\mathbf{r}_t)$.

We assume that the starting values $\mathbf{x}_t$ and $\boldsymbol{\mu}_t$, for $t = 1, \ldots, L$, have non-negative components.

Then $\boldsymbol{\mu}_t$ has non-negative components if all the following conditions are satisfied ($i, j = 1, \ldots, K$, $l = 1, \ldots, L$):

1. $\beta_{ijl} \geq 0$, $\alpha_{ijl} \geq 0$, $\alpha_{ijl} + \gamma_{ijl} \geq 0$*;*
2. *if* $\alpha_{ijl} = 0$ *then* $\delta_{ijl} \geq 0$*; if* $\alpha_{ijl} + \gamma_{ijl} = 0$ *then* $\delta_{ijl} \leq 0$*;*
3. $\omega_i - \dfrac{1}{4} \displaystyle\sum_{j=1}^{K} \sum_{l=1}^{L} \delta_{ijl}^2 \left[\frac{\mathrm{I}(\delta_{ijl} < 0) \, \mathrm{I}(\alpha_{ijl} > 0)}{\alpha_{ijl}} + \frac{\mathrm{I}(\delta_{ijl} > 0) \, \mathrm{I}(\alpha_{ijl} + \gamma_{ijl} > 0)}{\alpha_{ijl} + \gamma_{ijl}} \right] \geq 0$

Proof:

We prove the result considering the 'enlarged' formulation. For the 'contagion' version the proof is almost identical.

The proof is by induction on t. Assume that $\mathbf{x}_{t-l}, \boldsymbol{\mu}_{t-l} \geq 0$ for $l = 1, \ldots, L$. Then the i-th subequation of (77) can be rewritten taking:

$$\begin{aligned}
(x_1; \ldots; x_n) &= (\boldsymbol{\mu}_{t-1}; \ldots; \boldsymbol{\mu}_{t-L}; \mathbf{x}_{t-1}; \ldots; \mathbf{x}_{t-L})^{1/2} \\
(a_1, \ldots, a_n) &= (\boldsymbol{\beta}_{i.1}, \ldots, \boldsymbol{\beta}_{i.l}, \boldsymbol{\alpha}_{i.1} + \boldsymbol{\gamma}_{i.1} \, \mathrm{I}(r_{t-1} < 0), \ldots, \boldsymbol{\alpha}_{i.L} + \boldsymbol{\gamma}_{i.L} \, \mathrm{I}(r_{t-L} < 0)) \\
(b_1, \ldots, b_n) &= (\mathbf{0}', \ldots, \mathbf{0}', \boldsymbol{\delta}_{i.1} \operatorname{sign}(r_{t-1}), \ldots, \boldsymbol{\delta}_{i.L} \operatorname{sign}(r_{t-L})) \\
c &= \omega_i,
\end{aligned}$$

where $\boldsymbol{\beta}_{i.l}$, $\boldsymbol{\alpha}_{i.l}$, $\boldsymbol{\gamma}_{i.l}$, and $\boldsymbol{\delta}_{i.l}$ denote the i-th row of the corresponding matrix of coefficients at lag l.

At this point we can apply proposition 1 to the formulation obtained. Rewriting conditions 1, 2 and 3 of the proposition in the model notation and considering that, in the whole time series, the returns r_{t-l} take both positive and negative signs, the result follows.

□

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Benjamin, M. A., Rigby, R. A., and Stasinopoulos, M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, **98**(461), 214–223.

Bera, A. K., Bilias, Y., and Simlai, P. (2006). Estimating functions and equations: An essay on historical developments with applications to econometrics. In T. C. Mills and K. Patterson, editors, *Palgrave Handbook of Econometrics*, volume 1. Palgrave MacMillan.

Bibby, B. M., Jacobsen, M., and Sørensen, M. (2004). Estimating functions for discretely sampled diffusion-type models. Working paper, University of Copenagen, Department of Applied Mathematics and Statistics.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000). Copulas for finance: A reading guide and some applications. Technical report, Groupe de Recherche Opérationelle, Credit Lyonnais, Paris.

Brownlees, C. T. and Gallo, G. M. (2005). Ultra high-frequency data management. In C. Provasi, editor, *Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione. Atti del Convegno S.Co. 2005*, pages 431–436. CLUEP, Padova.

Chou, R. Y. (2005). Forecasting financial volatilities with extreme values: The conditional autoregressive range (carr) model. *Journal of Money, Credit and Banking*, **37**(3), 561–582.

Contreras, M. and Ryan, L. M. (2000). Fitting nonlinear and constrained generalized estimating equations with optimization software. *Biometrics*, **56**, 1268–1272.

Davis, R. A., Dunsmuir, W. T. M., and Strett, S. B. (2002). Observation driven model for poisson counts. Working paper, Department of Statistics, Colorado State University.

Desmond, A. F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling (with discussion). *Journal of Statistical Planning and Inference*, **60**, 77–121.

Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**(1), 83–106.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society, Series B*, **22**, 139–153.

Embrechts, P., Lindskog, F., and McNeil, A. J. (2001). Modelling dependence with copulas. Technical report, Department of Mathematics, ETHZ, CH-8092 Zurich Switzerland, `www.math.ethz.ch/finance`.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, **17**, 425–446.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, **131**(1–2), 3–27.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, **66**, 1127–62.

Engle, R. F., Gallo, G. M., and Velucchi, M. (2005). An mem–based investigation of dominance and interdependence across markets. presented at the Computational Statistics and Data Analysis World Conference, Cyprus.

Gallo, G. M. and Velucchi, M. (2005). Ultra-high frequency measures of volatility: An mem-based investigation. In C. Provasi, editor, *Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione. Atti del Convegno S.Co. 2005*, pages 269–274. CLUEP, Padova.

Glosten, L., Jaganathan, R., and Runkle, D. (1993). On the relation between the expected value and volatility of the nominal excess return on stocks. *Journal of Finance*, **48**, 1779–1801.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1212.

Heyde, C. (1997). *Quasi-Likelihood and Its Applications*. Springer-Verlag, New York.

Johnson, N., Kotz, S., and Balakrishnan, N. (2000). *Continuous Multivariate Distributions*. John Wiley & Sons, New York.

Jørgensen, B. and Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, **31**, 93–114.

Knudsen, S. J. (1999). Estimating functions and separate inference. Monographs, Vol.1, Department of Statistics and Demography, University of Southern Denmark. Available from http://www.stat.sdu.dk/oldpublications/monographs.html.

Li, W. K. (1994). Time series models based on generalized linear models: Some further results. *Biometrics*, **50**, 506–511.

Liang, K.-Y. and Zeger, S. L. (1995). Inference based on estimating functions in presence of nuisance parameters (with discussion). *Statistical Science*, **10**, 158–199.

Manganelli, S. (2002). Duration, volume and volatility impact of trades. Working paper, European Central Bank, WP ECB 125 http://ssrn.com/abstract=300200.

Shephard, N. (1995). Generalized linear autoregressions. Working paper, Nuffield College, Oxford OX1 1NF, UK.

Song, P. X.-K. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, **27**, 305–320.

Sørensen, M. (1999). On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, **13**, 111–136.

Vinod, H. D. (1998). Using Godambe-Durbin estimating functions in econometrics. In I. W. Basawa, V. P. Godambe, and R. L. Taylor, editors, *Selected Proceedings of the Symposium on Estimating Functions*, volume 32 of *IMS Lecture Notes Series*. Institute of Mathematical Statistics.

Table 2: Multiple Indicator MEM: Comparison across Estimators. GE

	$\lvert r_t \rvert$				hl_t				rv_t			
	ML-R	EE	ML-I	ML-I (own)	ML-R	EE	ML-I	ML-I (own)	ML-R	EE	ML-I	ML-I (own)
constant	0.0206	0.0185	0	0.0174	0.0772	0.0764	0.0663	0.0548	0.1172	0.1201	0.1504	0.1455
μ_{t-1}	0.7925	0.7883	0.7934	0.8052	0.7795	0.774	0.749	0.7693	0.645	0.638	0.5573	0.5621
	0.0442	*0.0413*	*0.0602*	*0.0683*	*0.0287*	*0.0208*	*0.056*	*0.0452*	*0.0397*	*0.0222*	*0.0514*	*0.0506*
$\lvert r_{t-1} \rvert$												
$\lvert r_{t-1} \rvert^-$	0.0709	0.06	0	0	0.0674	0.0642	0.0417		0.0434	0.0418	0.0488	0.0539
	0.021	*0.0208*	*0*	*0*	*0.0106*	*0.0091*	*0.0324*		*0.0084*	*0.0077*	*0.0108*	*0.0069*
hl_{t-1}	0.1564	0.1678	0.2066	0.1538	0.1283	0.1361	0.1799	0.1892				
	0.0447	*0.0423*	*0.0602*	*0.0572*	*0.0184*	*0.017*	*0.0296*	*0.0329*				
hl_{t-1}^-				0.0556					0.0359	0.0419	0.0274	
				0.0232					*0.0177*	*0.0179*	*0.0233*	
rv_{t-1}									0.2418	0.2467	0.3028	0.3007
									0.0285	*0.0183*	*0.038*	*0.0375*
rv_{t-1}^-									-0.0302	-0.0351	-0.024	
									0.0154	*0.0155*	*0.0197*	

Note: Standard errors for the constant are not reported since the coefficient was estimated through variance targeting. 'ML-I' reports the estimated results equation by equation with the specification selected by the system estimators. 'ML-I (own)' reports the estimated results equation by equation with its own specification.

Table 3: Estimated Impact Matrices of the vMEM on the indicators $|r_t|$, hl_t, rv_t – GE stock, 01/03/1995-12/29/2000.

EE			
	$\|r_{t-1}\|$	hl_{t-1}	rv_{t-1}
$\|r_t\|$	0.81827	0.16775	0
hl_t	0.03208	0.91007	0
rv_t	0.02091	0.02097	0.86716
ML-R			
	$\|r_{t-1}\|$	hl_{t-1}	rv_{t-1}
$\|r_t\|$	0.82797	0.15642	0
hl_t	0.03371	0.90783	0
rv_t	0.02168	0.01797	0.87161
ML-I (own)			
	$\|r_{t-1}\|$	hl_{t-1}	rv_{t-1}
$\|r_t\|$	0.80522	0.18163	0
hl_t	0	0.95848	0
rv_t	0.02694	0	0.86286
ML-I			
	$\|r_{t-1}\|$	hl_{t-1}	rv_{t-1}
$\|r_t\|$	0.79342	0.20658	0
hl_t	0.02085	0.92892	0
rv_t	0.02439	0.0137	0.84803

Table 4: Characteristic roots of the estimated Impact Matrices of the vMEM on the indicators $|r_t|$, hl_t, rv_t – GE stock, 01/03/1995-12/29/2001.

Characteristic roots of **A**			
ML-R	0.80064	0.85259	0.98827
EE	0.80118	0.84151	0.98738
ML-I (own)	0.78945	0.80948	0.97853

Table 5: Estimated parameters of the Gamma marginals of the vMEM on the indicators $|r_t|$, hl_t, rv_t – GE stock, 01/03/1995-12/29/2001 (1515 obs.)

	ML-R	ML-I
$\|r_t\|$	1.2808	1.272
hl_t	6.6435	6.689
rv_t	20.6836	20.84