ICAE

Instituto Complutense de Análisis Económico

UNIVERSIDAD COMPLUTENSE

FACULTAD DE ECONOMICAS

Campus de Somosaguas

28223 MADRID

Teléfono 91 394 26 11 - FAX 91 294 26 13

Internet: http://www.ucm.es/info/icae/

E-mail: icaesec@ccee.ucm.es

Documento de trabajo

The Likelihood of Multivariate Garch Models is Ill-Conditioned

Miguel Jerez

José Casal

Sonia Sotoca

No.9904 Septiembre 1999

ICAE

Instituto Complutense de Análisis Económico

UNIVERSIDAD COMPLUTENSE

THE LIKELIHOOD OF MULTIVARIATE GARCH MODELS IS ILL-CONDITIONED

Miguel Jerez
José Casals
Sonia Sotoca

Universidad Complutense de Madrid
Campus de Somosaguas
28223 Madrid

ABSTRACT

The likelihood of multivariate GARCH models is ill-conditioned because of two facts. First, financial time series often display high correlations, implying that an eigenvalue of the conditional covariances fluctuates near the zero boundary. Second, GARCH models explain conditional covariances in terms of a linear combination of delayed squared errors and their conditional expectation; this functional form implies that the likelihood function is almost flat in the neighborhood of the optimal estimates. Building on this analysis we propose a linear transformation of data which, not only stabilizes the likelihood computation, but also provides insight about the statistical properties of data. The use of this transformation is illustrated by modeling the short-run conditional correlations of four nominal exchange rates.

RESUMEN

La verosimilitud de procesos GARCH multivariantes está mal condicionada por dos causas. En primer lugar, las series financieras a menudo están fuertemente correladas, lo cual implica que un autovalor de las matrices de covarianzas condicionales está próximo a cero. En segundo lugar, los modelos GARCH explican la varianza condicional en términos de errores cuadráticos retardados y de la esperanza condicional de éstos; esta forma funcional implica que la función de verosimilitud es prácticamente plana en el entorno de las estimaciones óptimas. A partir de este análisis, proponemos una transformación lineal de los datos que, no sólo estabiliza el cálculo de la verosimilitud, sino que ayuda a analizar las propiedades estadísticas de los datos. El uso de esta transformación se ilustra modelizando las correlaciones condicionales a corto plazo de cuatro tipos de cambio nominales.

Key words: ARCH, GARCH, maximum-likelihood

JEL classification: C130, C320, C510.

Mailing address: Departamento de Fundamentos del Análisis Económico II, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain, E-mail: mjerez@ccee.ucm.es.

1. Introduction.

Since the seminal paper of Engle (1982) many works describe the volatility of financial yields using models with conditional heteroskedastic errors. Univariate models in the ARCH family are useful to measure and forecast the volatility of single assets. While this is important, problems of risk-assessment, asset-allocation, hedging and options pricing require knowledge of the properties of multivariate series. Often, these properties can be represented adequately by means of a vector GARCH model.

According with our experience, maximum-likelihood (ML) estimation of multivariate GARCH models often implies:

1) a high computational cost,

2) sensitivity of the estimates to changes in both, the sample and the initial conditions of the iterative algorithm,

3) frequent iteration on solutions where conditional covariances have negative eigenvalues and, because of this,

4) non-convergence or convergence to solutions with nonzero gradient. This 'false convergence' situation happens because many nonlinear algorithms stop when changes in the function or parameter values are considered small enough. In an ill-conditioned case, these heuristic criteria can be satisfied in solutions with a nonzero gradient.

This paper analyzes the causes of such bad behavior. We conclude that it is due to a) the fact that financial time series often exhibit high unconditional correlations and b) identificability problems derived from the functional form of GARCH processes. We will refer to these problems as "high correlations" and "identificability".

Poor identificability is implied by the functional form the GARCH model. It explains the conditional covariance as a function of delayed cross-products of errors and the conditional expectation of these cross-products. Obviously these variables share much common information and, in the neighborhood of the optimal estimates, are deemed to be very similar. Therefore, point-estimates of the parameters will be highly correlated and imprecise. On the other hand, poor identificability does not affect the capacity of a GARCH model to describe and forecast volatility and, except in extreme situations, should not compromise the stability of ML algorithms.

The issue of high correlations is more critical. It implies that there is at least one eigenvalue of the unconditional covariance is close to zero. Then, the smallest eigenvalues of conditional covariances fluctuate near the zero boundary and, in a context of iterative nonlinear methods, it is easy to iterate on a

trial solution where conditional covariances are not positive-definite. In this situation computing the likelihood results in unbounded or mathematically undefined operations.

When both, identificability and high correlation problems occur, a) the likelihood function is almost flat in the neighborhood of the optimal estimates and b) this point is close to the zone of the parametric space where covariances are not positive-semidefinite. This situation spells disaster for iterative ML methods.

Building on this analysis we propose a linear transformation of data designed to project the eigenvalues of conditional covariances far from the zero boundary and to optimize their relative value. This transformation is closely related to principal components and results useful, not only to stabilize the computation of likelihood, but also to analyze the statistical properties of the sample.

The structure of the paper is as follows. Section 2 states the problem of likelihood computation on standard grounds. Section 3 describes in detail the problems summarized above and discusses its implications. Section 4 defines the stabilizing data transformation and characterizes its properties. Section 5 applies this data transformation to model the short-run conditional correlations of four nominal exchange rates. Finally, Section 6 discusses previous results and summarizes the main conclusions.

2. Problem statement and notation.

Consider a $(k\times 1)$ random vector y_t which, by means of an econometric model, is decomposed as $y_t = \mathrm{E}_{t-1}(y_t) + \varepsilon_t$, being $\mathrm{E}_{t-1}()$ the expectation of the argument conditional to the information set up to t-1, Ω_{t-1}. In a conditional heteroskedastic framework, the errors ε_t are such that $\varepsilon_t \sim \mathrm{iid}(\mathbf{0},\Sigma)$, $\varepsilon_t \mid \Omega_{t-1} \sim \mathrm{iid}(\mathbf{0},\Sigma_t)$.

Assume without loss of generality that $y_t = \varepsilon_t$. If the conditional covariance Σ_t depends on a vector θ of unknown parameters, the minus log gaussian likelihood of a sample of size N is:

$$\ell(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N \mid \theta) = \frac{1}{2}Nk\ln(2\pi) + \frac{1}{2}\sum_{t=1}^{N}\left(\ln\left|\Sigma_t(\theta)\right| + \varepsilon_t^T\Sigma_t(\theta)^{-1}\varepsilon_t\right) \qquad (1)$$

Literature proposes different ways to parametrize Σ_t. Many formulations are encompassed by the multivariate GARCH(p,q). To avoid unnecessary complications, in the rest of the paper we will assume that p=q=1. The vector GARCH(1,1) model is characterized by:

$$\mathrm{vech}(\Sigma_t) = w + \mathrm{A}\,\mathrm{vech}(\varepsilon_{t-1}\varepsilon_{t-1}^T) + \mathrm{B}\,\mathrm{vech}(\Sigma_{t-1}) \qquad (2)$$

where vech(.) denotes the vector-half operator, which stacks the lower triangle of an $N\times N$ symmetric matrix into a $[N(N+1)/2]\times 1$ vector.

The following remarks summarize some features of model (2) that will be used in the rest of the paper:

1) It has a large number of parameters, even for moderate sizes of k. Many authors worry about this lack of parsimony and suggest simplifying assumptions like diagonal structure (Bollerslev *et al.* 1988) or constant-correlations (Bollerslev, 1990).

2) The functional form (2) does not assure the positive-definiteness of conditional covariances. In fact, this is a very difficult condition to impose except in drastically simplified versions of the model.

3) By definition, the variables in the right-hand-side of (2) are such that:

$$\text{vech}(\varepsilon_\tau \varepsilon_\tau^T) = \text{vech}(\Sigma_\tau) + \nu_\tau, \text{ for all } \tau \tag{3}$$

where ν_τ is (conditional and unconditionally) a zero-mean uncorrelated process with a complex heteroskedasticity (Bollerslev, 1988, pp. 123).

4) Generalizing the univariate result in Bollerslev (1988), the decomposition (3) allows one to express (2) as a VARMA(1,1) model:

$$(I - \Phi L)\text{vec}(\varepsilon_t \varepsilon_t^T) = w + (I - \Theta L)\nu_t \tag{4}$$

where L is the lag operator, ν_t are the innovations defined in (3) and the AR and MA factors are related to the polynomials in (2) by $\Phi = A + B$ and $\Theta = B$, respectively. If model (2) is such that the roots of $|I - \Phi\lambda| = 0$ lie outside the unit circle, then (4) can be written as:

$$\text{vech}(\varepsilon_t \varepsilon_t^T) = \text{vech}(\Sigma) + (I - \Phi L)^{-1}(I - \Theta L)\nu_t \tag{5}$$

where the constant term is the vector-half of the unconditional covariance:

$$\text{vech}(\Sigma) = (I - \Phi L)^{-1} w = [I - A - B]^{-1} w \tag{6}$$

Unless otherwise indicated we will use the representation (5)-(6), keeping in mind that it is observationally equivalent to the standard form (2).

3. Sources of ill-conditioning in likelihood computation.

3.1 High correlations.

Financial time series often display high unconditional correlations. Some explanations of this empirical regularity may be a) common statistical features of data, b) common factors due to the nature of the series (*e.g.* exchange rates are often related to a single currency) or c) simultaneous volatility clusters. In terms of principal components, high correlations imply that there is at least one quasi-deterministic linear combination of the series, characterized by a small eigenvalue of the unconditional covariance. In this situation the smallest eigenvalues of conditional covariances will fluctuate near the zero boundary.

Taking into account the form of the log-likelihood function (1), this situation is dangerous because:

1) Iterating on a solution $\hat{\theta}$, where $\Sigma_t(\hat{\theta})$ has small eigenvalues, may yield floating-point errors or unbounded results when computing:
   1.1) the sequences $\ln|\Sigma_t(\hat{\theta})|$ and $\Sigma_t(\hat{\theta})^{-1}$ $(t = 1,\ldots,N)$ in (1).
   1.2) the first and second-order derivatives of (1), which are functions of $\Sigma_t(\hat{\theta})^{-1}$.

2) If $\Sigma_t(\hat{\theta})$ has some negative eigenvalues, computation of $\ln|\Sigma_t(\theta)|$ $(t = 1,\ldots,N)$ result in mathematically undefined operations. Besides, many ML algorithms rely on the use of Cholesky decomposition to avoid the explicit inversion of covariance matrices. As Cholesky factors require these matrices to be positive-definite, negative eigenvalues also induce errors by this way when computing the function (1) or its derivatives. According to our experience, simple perturbation techniques help to avoid runtime errors, but are not useful to achieve convergence.

The following example illustrates the effect of high correlations on the eigenvalues of conditional covariances.

Example 1. Consider the bivariate GARCH(1,1) model expressed in the form (5):

$$\begin{bmatrix} \varepsilon_{1t}^2 \\ \varepsilon_{1t}\varepsilon_{2t} \\ \varepsilon_{2t}^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 \\ \sigma_{12} \\ \sigma_2^2 \end{bmatrix} + \begin{bmatrix} 1-.97B & 0 & 0 \\ 0 & 1-.90B & 0 \\ 0 & 0 & 1-.85B \end{bmatrix}^{-1} \begin{bmatrix} 1-.86B & 0 & 0 \\ 0 & 1-.80B & 0 \\ 0 & 0 & 1-.73B \end{bmatrix} \begin{bmatrix} \nu_{1t} \\ \nu_{12t} \\ \nu_{2t} \end{bmatrix} \tag{7}$$

and the unconditional covariances:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1.0 & .8 \\ .8 & 1.0 \end{bmatrix} \text{ ; with eigenvalues: } \lambda_1 = 1.8,\ \lambda_2 = .2, \text{ and} \tag{8}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1.0 & .1 \\ .1 & 1.0 \end{bmatrix} \text{ ; with eigenvalues: } \lambda_1 = 1.1 \text{ and } \lambda_2 = .9 \quad , \tag{9}$$

Note that the ratio between the smallest and largest eigenvalues in the first case ($\lambda_2/\lambda_1 = .111$) is much lower than in the second case ($\lambda_2/\lambda_1 = .818$). This fact characterizes a (not extreme) ill-conditioned situation.

The example consists of:

1) Obtaining two realizations with N=300 of a bivariate white noise process ε_t, which conditional covariances are given by model (7)-(8) for the first series, and model (7)-(9) for the second series.

2) Computing the sequences of conditional covariances and the corresponding eigenvalues, using the true value of the parameters.

Figure 1 represents the smallest and highest eigenvalues of $\Sigma_t(\theta)$ in the ill-conditioned case ($\sigma_{12} = .8$). Note that the first sequence fluctuates very close to the zero boundary, being its extreme values min=.019 and max=.288. Figure 2 displays the same eigenvalues in the well-conditioned case ($\sigma_{12} = .1$). Note that the sequence of smallest eigenvalues (min=.354, max=.960) is farther from zero than in the previous case.

[Insert Figure 1]

[Insert Figure 2]

The sequences in Figures 1 and 2 have been computed with the true values of the parameters. A sensitivity analysis reveals that small perturbations of the parameters in the ill-conditioned case yield negative eigenvalues. For example, if the MA parameter of the covariance equation in (7) is set to .82 instead of its true value .80, then the sequence of conditional covariances has several negative eigenvalues, being the smallest -0.012. In the well-conditioned case, however, the eigenvalues are much more robust. Therefore, a nonlinear ML algorithm has a higher risk of iterating on a solution with negative eigenvalues when correlations between the series are high - like those in (8) - than when they are small.

3.2. Poor identificability.

As we said in the Introduction, poor identificability is due to the functional form of the GARCH model. To simplify the analysis, we will discuss this problem in a univariate framework. Assume therefore that $y_t = \varepsilon_t$, $\varepsilon_t \sim \text{iid}(0,\sigma^2)$, $\varepsilon_t | \Omega_{t-1} \sim \text{iid}(0,\sigma_t^2)$. A GARCH(1,1) in the standard form (2) is:

$$\sigma_t^2 = w + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{10}$$

According to (3), the variables in the right-hand-side of (10) are related by:

$$\varepsilon_{t-1}^2 = \sigma_{t-1}^2 + v_{t-1} \tag{11}$$

being v_{t-1} an uncorrelated, zero-mean heteroskedastic noise. Eqs. (10)-(11) imply that:

1) The variables in the right-hand-side of (10), ε_{t-1}^2 and σ_{t-1}^2, are such that: $E_{t-2}(\varepsilon_{t-1}^2) = \sigma_{t-1}^2$.

2) The term v_{t-1} in (11) can be interpreted as the information in ε_{t-1}^2 which is not contained in σ_{t-1}^2. Then, if the information (or variance) of v_{t-1} is low, it will be difficult to obtain independent estimates of α and β, whereas some linear combination of these parameters will be identified.

Therefore, the likelihood of (10) is very flat in some directions of the parametric space. It is difficult to say when this problem will be important, because the support of v_{t-1} changes in time (Bollerslev, 1988, pp. 123) so its variance is almost impossible to describe analytically. One may guess that if model (10) shows high persistence - *i.e.* if $\alpha + \beta \approx 1$ - the parameters will be more identifiable because σ_{t-1}^2 would be less adaptive to ε_{t-1}^2 than in a model with less persistence.

The following example illustrates the poor identificability of a GARCH(1,1) model using simulated data.

Example 2. Consider 500 samples of the process $\varepsilon_t \sim \text{iid}(0,\sigma^2)$, $\varepsilon_t | \Omega_{t-1} \sim \text{iid}\, N(0,\sigma_t^2)$ with conditional variances following a GARCH(1,1) model in ARMA form:

$$\varepsilon_t^2 = \sigma^2 + \frac{1 - \theta B}{1 - \phi B} v_t \tag{12}$$

with $\sigma^2 = 1$, $\theta = .6$ and $\phi = .7$. The ML estimates of the parameters in (12), their correlations and the corresponding principal components are summarized in Table 1.

[Insert Table 1]

Note that:

1) Point estimates are close to the true values.

2) The estimate of the unconditional variance is almost orthogonal to the rest of the parameters. This situation is characterized both by a) small correlations of $\hat{\sigma}^2$ with $\hat{\phi}$ and $\hat{\theta}$, and b) an eigenvalue of 1.0 associated with the eigenvector [1 .01 -.1].

3) Correlation between $\hat{\phi}$ and $\hat{\theta}$ is .98. The highest eigenvalue (1.98) is associated with the eigenvector [.04 .71 .71], showing that the sum of both parameters is well identified. On the other hand, the smallest eigenvalue (.02) is associated with the eigenvector [.05 .71 -.71]. The difference between both estimates - which is the α parameter in (10) - is then ill-identified.

Figure 3 shows the optimal estimates (represented by a '+' sign) corresponding to a log-likelihood of 720.840, and the isoquantas of the log-likelihood conditional to $\hat{\sigma}^2 = 1.065$. The isoquantas are chosen to represent confidence regions for ϕ and θ, from a 5% confidence (given by the inner conic section) up to 95% in increments of 10 percent points. The first three isoquantas are labeled with the corresponding likelihood value. This Figure shows that a) big zones of the parametric space have a likelihood similar to the optimal and b) confidence regions are wide and, therefore, point-estimates result very uncertain.

[Insert Figure 3]

4. Stabilizing likelihood computation.

According to previous analysis, let be ε_t a $(k\times 1)$ random vector such that:

$$\varepsilon_t \sim \text{iid}(\mathbf{0}, \Sigma) \tag{13}$$

$$\varepsilon_t | \Omega_{t-1} \sim \text{iid}(\mathbf{0}, \Sigma_t) \tag{14}$$

and consider the linear transformation:

$$\varepsilon_t^* = V\varepsilon_t \tag{15}$$

where V is a $(k\times k)$ matrix of real numbers such that $|V| \neq 0$.

The problem of high correlations, discussed in Section 3.1, arises when an eigenvalue of Σ is relatively small. Then, the data can be optimally scaled by choosing:

$$V = \Lambda^{-1/2} M^T \tag{16}$$

where matrices in the right-hand-side of (16) are given by the eigenvalue-eigenvector decomposition:

$$\Sigma = M\Lambda M^T \tag{17}$$

4.1. Analytic properties of the stabilizing linear transformation.

The following propositions relate the stochastic properties of ε_t^* with those of ε_t.

Proposition 1. The unconditional and conditional distributions of ε_t^* are:

$$\varepsilon_t^* \sim \text{iid}(\mathbf{0}, I) \tag{18}$$

$$\varepsilon_t^* | \Omega_{t-1} \sim \text{iid}(\mathbf{0}, \Sigma_t^*), \text{ with } \Sigma_t^* = V\Sigma_t V^T \tag{19}$$

Proof. The result follows immediately from (13)-(17).

Note that the result in (18) implies that the transformation defined by (15)-(17) is optimal, as it scales all the eigenvalues of the unconditional covariance to unity, thus achieving the optimal condition number of one. An additional advantage is that the transformed values ε_t^* have a meaningful statistical interpretation, as standardized principal components of ε_t.

Proposition 2. If Σ_t is such that:

$$\text{vech}(\Sigma_t) = w + \text{A}\,\text{vech}(\varepsilon_{t-1}\varepsilon_{t-1}^T) + \text{B}\,\text{vech}(\Sigma_{t-1}) \tag{20}$$

then Σ_t^* follows the GARCH(1,1) motion law:

$$\text{vech}(\Sigma_t^*) = w^* + \text{A}^*\,\text{vech}(\varepsilon_{t-1}^*\varepsilon_{t-1}^{*T}) + \text{B}^*\,\text{vech}(\Sigma_{t-1}^*) \tag{21}$$

where: $$w^* = P^{-1} w \tag{22}$$

$$\text{A}^* = P^{-1}\,\text{A}\,P \tag{23}$$

$$\text{B}^* = P^{-1}\,\text{B}\,P \tag{24}$$

$$P = \Delta_1 (V^{-1} \otimes V^{-1}) \Delta_2,\ V^{-1} = M\Lambda^{1/2} \tag{25}$$

and Δ_1, Δ_2 are 0-1 matrices such that, for any symmetric matrix S, $\text{vech}(S) = \Delta_1 \text{vec}(S)$ and $\text{vec}(S) = \Delta_2 \text{vech}(S)$, being vec(.) the operator which stacks the columns of an $N\times N$ matrix into a $N^2\times 1$ vector.

Proof. See Appendix A.

Corollary 1. If the variance model is expressed in the form (5):

$$\text{vech}(\varepsilon_t \varepsilon_t^T) = \text{vech}(\Sigma) + (I - \Phi L)^{-1}(I - \Theta L)\nu_t \tag{26}$$

the cross-products of the transformed data follow the VARMA model:

$$\text{vech}(\varepsilon_t^* \varepsilon_t^{*T}) = I + (I - \Phi^* L)^{-1}(I - \Theta^* L)\nu_t^* \tag{27}$$

where:

$$\Phi^* = P^{-1} \Phi P \tag{28}$$

$$\Theta^* = P^{-1} \Theta P \tag{29}$$

Proposition 3. $\ell(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) = \ell(\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_N^*) + \frac{N}{2} \log |\Lambda|$, being $\ell()$ the minus log gaussian density of a sample.

Proof. See Appendix B.

Note that, replacing (18) by $\varepsilon_t^* \sim \text{iid}(\mathbf{0}, V \Sigma V^T)$, propositions 1 and 2 hold for any choice of V. A general result analogous to Proposition 3 is easy to derive following the proof in Appendix B, as only the final simplification relies in the particular choice of V given in (16).

4.2. Econometric implementation.

The results in Section 4.1 were derived for the true values of the data generating process. Building on them, the following empirical implementation is straightforward:

Step 1: Starting from a sample $\{\varepsilon_t\}_{t=1,\ldots,N}$, compute an estimate of the unconditional covariance matrix, $\hat{\Sigma}$, the eigenvalue-eigenvector decomposition (17), the matrix $\hat{V}$ using the sample analogue of (16) and the transformed series $\{\varepsilon_t^*\}_{t=1,\ldots,N}$ using (15). Specify a GARCH model for ε_t^*. We will assume that it is a GARCH(1,1) in the form (2).

Step 2:

Step 2.1: Compute consistent estimates for the parameters in (21), $\hat{\varpi}^*$, $\hat{\mathrm{A}}^*$ and $\hat{\mathrm{B}}^*$. If ML is used, assure that the corresponding gradient is small enough.

Step 2.2: Compute the covariances $\{\hat{\Sigma}_t^*\}_{t=1,\ldots,N}$ according to (21). Check the smallest eigenvalue to assure that it is positive.

Step 2.3: If required, obtain estimates of the parameters in (2) through the expressions:

$$\hat{\varpi} = \hat{P} \hat{\varpi}^* \tag{30}$$

$$\hat{\mathrm{A}} = \hat{P} \hat{\mathrm{A}}^* \hat{P}^{-1} \tag{31}$$

$$\hat{\mathrm{B}} = \hat{P} \hat{\mathrm{B}}^* \hat{P}^{-1} \tag{32}$$

where $\hat{P}$ denotes the sample analogue of P, see Eq. (25). Finally, compute estimates of the conditional covariances using:

$$\hat{\Sigma}_t = \hat{V}^{-1} \hat{\Sigma}_t^* (\hat{V}^{-1})^T \tag{33}$$

Expressions (30)-(32) follow immediately from Eq. (22)-(25) and (33) follows from (19). Note that consistency is assured by the Theorem of Slutsky. If ML were employed to compute the estimates in Step 2.1, Proposition 3 assures that the estimates $\hat{\varpi}$, $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$ are asymptotically equivalent to ML estimates. It also can be applied to compute information criteria or LR statistics.

Step 3: If required, compute estimates of the covariances of $\hat{\varpi}$, $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$ using the following Proposition:

Proposition 4. If $\hat{\text{cov}}(\hat{\varpi}^*)$, $\hat{\text{cov}}(\hat{\mathrm{A}}^*)$ and $\hat{\text{cov}}(\hat{\mathrm{B}}^*)$ are consistent estimates of the covariances of $\hat{\varpi}^*$, $\hat{\mathrm{A}}^*$ and $\hat{\mathrm{B}}^*$, respectively, then the expressions:

$$\hat{\text{cov}}(\hat{\varpi}) = \hat{P}\, \hat{\text{cov}}(\hat{\varpi}^*) \hat{P}^T \tag{34}$$

$$\text{vec}[\hat{\text{cov}}(\hat{\mathrm{A}})] = [(\hat{P}^{-1})^T \otimes \hat{P}]\, \text{vec}[\hat{\text{cov}}(\hat{\mathrm{A}}^*)][(\hat{P}^{-1})^T \otimes \hat{P}]^T \tag{35}$$

$$\text{vec}[\hat{\text{cov}}(\hat{\mathrm{B}})] = [(\hat{P}^{-1})^T \otimes \hat{P}]\, \text{vec}[\hat{\text{cov}}(\hat{\mathrm{B}}^*)][(\hat{P}^{-1})^T \otimes \hat{P}]^T \tag{36}$$

provide consistent estimates of the covariances of $\hat{\varpi}$, $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$.

Proof. Expression (34) follows immediately from (30). Applying the vec() operator to both sides of (31) we obtain:

$$\text{vec}(\hat{\mathrm{A}}) = [(\hat{P}^{-1})^T \otimes \hat{P}]\, \text{vec}(\hat{\mathrm{A}}^*) \tag{37}$$

which implies (35). The proof of (36) is analogous to this one. ■

This implementation allows one to obtain results for original data from those corresponding to transformed data. The following example illustrates its application.

5. Empirical example: short-run alignment of exchange rates.

It is well known that many exchange rates fluctuate in the same direction and in similar proportions. This co-movement can be explained by competitive appreciation or depreciation policies, by international agreements or just by the fact that all the rates are expressed in terms of a common numeraire (often the US Dollar) which performance affects them all.

Long-term comovements can be effectively measured through sample correlations. On the other hand, short-term fluctuations may deviate substantially from the alignment implied by the long-run correlation matrix. In this Section we model short-run comovements of four relevant currencies through the conditional correlations implied by a vector GARCH model.

Consider the spot bid exchange rates of Deutsche Mark (DM), French Franc (FF), British Pound (BP) and Japanese Yen (JY) against US Dollar, observed in the London Market during 695 weeks, from January 1985 to April 1998. The data has been logged, differenced and scaled by a factor of 100, to obtain the corresponding log percent yields. Excess returns are then computed by substracting the sample mean.

Table 2 summarizes the main descriptive statistics of the excess returns. Note that a) all the series exhibit excess kurtosis and some asymmetry, perhaps relevant for BP and JY, b) the correlations are high, ranging from .48 (BP-JY) to .98 (DM-FF), according to this fact and c) the ratio between the lowest and highest eigenvalues of the covariance matrix ($\lambda_{min}/\lambda_{max}$ = .0069) suggests that there will be a problem of high correlations. Note that the scaled eigenvectors in the last panel of Table 2 are the sample analogues of V in (16).

[Insert Table 2]

We tried to fit diagonal GARCH(1,1) models to all the possible pairs of the excess returns. Most of the attempts converged to solutions with a nonzero gradient and some negative eigenvalue in the conditional covariances. Convergence was obtained only when JY was included in the pair. Taking into account the analysis in Section 3.1 this was to be expected, as the correlation between JY returns and those of the other currencies is relatively small. All the attempts to build a model for three series failed to converge. Therefore, we will use the data transformation defined in Section 4.

Inspection of data scaled according to (15)-(17) reveals that the first series has a big outlier (-12.8 standard deviations) in the second week of April 1986. The corresponding scaled eigenvector implies that this series is roughly the difference between the returns of DM and FF (see Table 2). This anomaluos value does not occur in a cluster of high volatility and its source was traced to a) a high positive fluctuation of the FF exchange rate (+2.77 standard deviations), combined with b) a simultaneous small negative variation of the DM (-.69 standard deviations). As the correlation between both series is .98, this combination is unlikely.

The anomalous FF excess return was corrected using a simple intervention model, see Box and Tiao (1975). Table 3 summarizes both, the new scaled eigenvector matrix and the Box-Ljung Q statistics of cross-products of the transformed series. This test rejects the null of no conditional heteroskedasticity. Figure 4 shows the resulting scaled series.

[Insert Table 3]

[Insert Figure 4]

A standard analysis of the scaled series and their cross-products suggests that a diagonal GARCH(1,1) will be adequate to capture most of the conditional heteroskedasticity. Table 4 summarizes the ML estimates of this model, expressed in the VARMA form (5). Note that:

1) All the parameters are much higher than its standard errors. As the scaled data is not gaussian, this is only informal evidence of statistical significance.

2) Many AR parameters are close to one, which implies a high persistence of variance effects.

3) The parameters in the constant term, which are the unconditional covariances, have been constrained to identity matrix values, in coherence with the properties of data transformation, see Eq. (18). Free estimates of these parameters (not shown here) are very similar to these and a likelihood-ratio test would not reject the null of that the unconditional covariance is equal to identity.

4) True convergence has been achieved, as the square root norm of gradient in both cases is small.

5) After convergence, we have computed the sequences of conditional covariances implied by the model both, for the scaled and original data. The minimum eigenvalues of both sequences, shown in the last two rows of Table 4, are positive.

[Insert Table 4]

Table 5 summarizes the descriptive statistics of standardized residuals. Apart from a typical excess kurtosis, there are no symptoms of misspecification. In particular, the Box-Ljung statistics do not reject the null of conditional homoskedasticity.

[Insert Table 5]

Figure 5 shows the conditional volatilities (square roots of conditional variances) implied by the model. Note that: a) volatilities of DM and FF returns are almost equal, b) BP returns share common periods of volatility with DM and FF yields and c) JY is more stable than the European currencies.

[Insert Figure 5]

Figure 6 show the conditional correlations implied by the model, which have clear and intuitive patterns. First, conditional correlations between DM and FF returns are close to unity, with transitory deviations in the last half of the sample. This result is hardly surprising, as both currencies are in the hard core of the EMS. Second, conditional correlations of BP returns with other European currencies are weaker (around .80, with highs and lows of .93 and .45 respectively) and there is a decreasing trend in the last part of the sample. Finally correlations of JY returns with those of European currencies are relatively small, around .5 to .6 with highs and lows of .95 and 0, respectively.

[Insert Figure 6]

6. Concluding remarks.

The first part of this paper concludes that iterative ML estimation of multivariate GARCH models is prone to diverge due to negative eigenvalues in the conditional covariances. Literature is unanimously concerned about the positive definiteness of these matrices and is conscious that ML estimation of multivariate ARCH models results difficult. Many authors, *e.g.,* Engle and Kroner (1995), worry also about the large number of parameters of unconstrained ARCH processes.

Whereas lack of parsimony contributes to instability of ML, two reasons suggest that it is not such a serious problem by itself. First, in a context of high-frequency financial data, availability of huge datasets somewhat balances overparametrization. Second, simplified ARCH models (*e.g.,* diagonal GARCH) often show the same instability of unconstrained specifications. We think that the high correlations and identificability problems discussed in sections 3 and 4 provide a more direct explanation than lack of parsimony. Besides, they suggest how to detect the potential problem before model building and how to improve the behavior of ML algorithms.

The issue of high correlations is obviously the most important of both, as it compromises the validity of estimates. This problem is easy to detect before model building, using the eigenvalues of a sample unconditional correlation matrix and the corresponding condition number.

Except in extreme cases, the problem of identificability is important only when combined with high correlations. By itself, it implies that point-estimates will be highly correlated and imprecise. On the other hand, it does not affect the capacity of GARCH specifications to describe and forecast volatility and can be dealt with by restrictions on the model parameters, *e.g.*, imposing IGARCH constraints. Existence of cofeatures in variance, see Engle and Kozicki (1993), also allows one to improve identificability by simplifying the model dynamic structure.

We have shown that the econometric implementation outlined in Section 4, which is closely related to factor-ARCH modeling, see Engle *et al.* (1990), contributes to the stability of likelihood computation. It also confirms that instability in likelihood computation is mainly due to the relative scale of the unconditional covariance eigenvalues. On the other hand it has clear limitations, as it does not assure conditional covariances to be positive-definite. This requires using a different parametrization like, *e.g.,* the previously mentioned constant correlations form or the BEKK model, see Engle and Kroner (1995).

The proposed transformation has three additional advantages. First, working with original or transformed data is indifferent for practical purposes, as the propositions in Section 4 define one-to-one relationships between their main stochastic properties. Second, the transformed variables, besides an obvious financial interpretation as yields of orthogonal portfolios, have a clear statistical meaning and may help in model building, *e.g.*, by revealing unlikely comovements, as was illustrated in the empirical example in Section 5. Third, as the unconditional covariance of the transformed variables is identity, imposing the corresponding constraints reduces the number of free parameters in the model and improves identificability.

Empirical evidence, not shown here, suggests that the data transformation improves the performance of ML algorithms even when using stable parametrizations as, for example, the BEKK model, see Engle and Kroner (1995). We think that this happens because the transformation improves the scaling of both, the data and the conditional covariance eigenvalues. Obviously if a model assures that conditional covariances are positive-definite, negative eigenvalues are not an issue. However, ill-conditioning problems also arise when some eigenvalues are positive but close to zero.

Acknowledgements.

Alfonso Novales made useful comments and suggestions to previous versions of this work. Sonia Sotoca acknowledges financial support from CICYT, project PB95-0912/95 and Fundación Caja de Madrid.

References.

Bollerslev, T., 1988. On the Correlation Structure for the Generalized Autoregressive Conditional Heteroskedastic Process. *Journal of Time Series Analysis*, 9, 2, 121-131.

Bollerslev, T., 1990. Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Approach. *Review of Economics and Statistics*, 72, 498-505.

Bollerslev, T., R.F. Engle and J.M. Wooldridge, 1988. A Capital-Asset Pricing Model with Time-Varying Covariances. *Journal of Political Economy*, 96/1, 116-131.

Box, G.E.P. and G.C. Tiao, 1975. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70-79.

Engle, R.F., 1982. Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation. *Econometrica*, 50, 987-1008.

Engle, R.F., V.K. Ng and M. Rotschild, 1990. Asset Pricing with a FACTOR-ARCH Covariance Structure: Empirical Estimates for Treasure Bills. *Journal of Econometrics*, 45, 213-237

Engle, R.F. and S. Kozicki, 1993. Testing for Common Features. *Journal of Business and Economic Statistics*, 11, 369-380.

Engle, R.F. and K.F. Kroner, 1995. Multivariate Simultaneous Generalized ARCH. *Econometric Theory*, 11, 122-150.

Appendix A. Proof of Proposition 2.

Eqs. (15) and (19) imply that:

$$\varepsilon_t = V^{-1}\varepsilon_t^* \tag{A.1}$$

$$\Sigma_t = V^{-1}\Sigma_t^*(V^{-1})^T \tag{A.2}$$

Substituting (A.1) and (A.2) in (20) yields:

$$\text{vech}[V^{-1}\Sigma_t^*(V^{-1})^T] = w + \text{A}\,\text{vech}[V^{-1}\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T(V^{-1})^T] + \text{B}\,\text{vech}[V^{-1}\Sigma_{t-1}^*(V^{-1})^T] \tag{A.3}$$

The next steps require to use the following algebraic result:

$$\text{vec}(ABA^T) = (A\otimes A)\text{vec}(B) \tag{A.4}$$

and the fact that the vech() and vec() operators are such that, for any symmetric matrix S, $\text{vech}(S) = \Delta_1\text{vec}(S)$ and $\text{vec}(S) = \Delta_2\text{vech}(S)$vector, being Δ_1, Δ_2 are 0-1 matrices.

Then, Exp. (A.3) in vec() form becomes:

$$\Delta_1\text{vec}[V^{-1}\Sigma_t^*(V^{-1})^T] = w + \text{A}\,\Delta_1\text{vec}[V^{-1}\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T(V^{-1})^T] + \text{B}\,\Delta_1\text{vec}[V^{-1}\Sigma_{t-1}^*(V^{-1})^T] \tag{A.5}$$

and by result (A.4):

$$\begin{aligned}\Delta_1[V^{-1}\otimes V^{-1}]\text{vec}(\Sigma_t^*) = w &+ \text{A}\,\Delta_1[V^{-1}\otimes(V^{-1})^T]\text{vec}[\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T] \\ &+ \text{B}\,\Delta_1[V^{-1}\otimes V^{-1}]\text{vec}(\Sigma_{t-1}^*)\end{aligned} \tag{A.6}$$

which can be expressed in vech() notation as:

$$\begin{aligned}\Delta_1[V^{-1}\otimes V^{-1}]\Delta_2\text{vech}(\Sigma_t^*) = w &+ \text{A}\,\Delta_1[V^{-1}\otimes V^{-1}]\Delta_2\text{vech}[\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T] \\ &+ \text{B}\,\Delta_1[V^{-1}\otimes V^{-1}]\Delta_2\text{vech}(\Sigma_{t-1}^*)\end{aligned} \tag{A.7}$$

Denoting: $P = \Delta_1[V^{-1}\otimes V^{-1}]\Delta_2$ simplifies (A.7) to:

$$P\,\text{vech}(\Sigma_t^*) = w + \text{A}P\,\text{vech}[\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T] + \text{B}P\,\text{vech}(\Sigma_{t-1}^*) \tag{A.8}$$

which implies:

$$\text{vech}(\Sigma_t^*) = P^{-1}w + P^{-1}\text{A}P\,\text{vech}[\varepsilon_{t-1}^*(\varepsilon_{t-1}^*)^T] + P^{-1}\text{B}P\,\text{vech}(\Sigma_{t-1}^*) \tag{A.9}$$

Finally, identifying the parameter matrices in (A.9) and (21) yields Exp. (22)-(25). ■

Appendix B. Proof of Proposition 3.

According with (14), the minus log gaussian likelihood of a sample of size N is:

$$\ell(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N) = \frac{1}{2}Nk\ln(2\pi) + \frac{1}{2}\sum_{t=1}^{N}(\ln|\Sigma_t| + \varepsilon_t^T\Sigma_t^{-1}\varepsilon_t) \tag{B.1}$$

Substituting (A.1) and (A.2) in (B.1) yields:

$$\ell(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N) = \frac{1}{2}Nk\ln(2\pi) + \frac{1}{2}\sum_{t=1}^{N}\{\ln|V^{-1}\Sigma_t^*(V^{-1})^T| + (\varepsilon_t^*)^T(V^{-1})^T[V^{-1}\Sigma_t^*(V^{-1})^T]^{-1}V^{-1}\varepsilon_t^*\} \tag{B.2}$$

and the terms inside the summatory are such that:

$$\ln|V^{-1}\Sigma_t^*(V^{-1})^T| = -2\ln|V| + \ln|\Sigma_t^*| = \ln|\Lambda| + \ln|\Sigma_t^*| \tag{B.3}$$

$$(\varepsilon_t^*)^T(V^{-1})^T[V^{-1}\Sigma_t^*(V^{-1})^T]^{-1}V^{-1}\varepsilon_t^* = (\varepsilon_t^*)^T(V^{-1})^TV^T(\Sigma_t^*)^{-1}VV^{-1}\varepsilon_t^* = (\varepsilon_t^*)^T(\Sigma_t^*)^{-1}\varepsilon_t^* \tag{B.4}$$

To understand the simplification in (B.3), note that (16) implies that $|V| = |\Lambda^{-1/2}|$, because the determinant of the eigenvector matrix M is one and, therefore, $\ln|V| = -\frac{1}{2}\ln|\Lambda|$.

Finally, substituting (B.3) and (B.4) in (B.2) implies that:

$$\ell(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N) = \ell(\varepsilon_1^*,\varepsilon_2^*,\ldots,\varepsilon_N^*) + \frac{N}{2}\log|\Lambda| \tag{B.5}$$

■

Fig. 1. Eigenvalues of the conditional covariances in the ill-conditioned case ($\sigma_{12} = .8$).

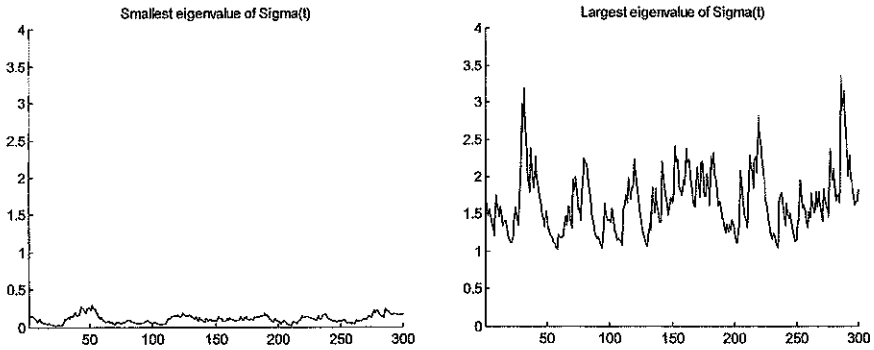

Fig. 2. Eigenvalues of the conditional covariances in the well-conditioned case ($\sigma_{12} = .1$).

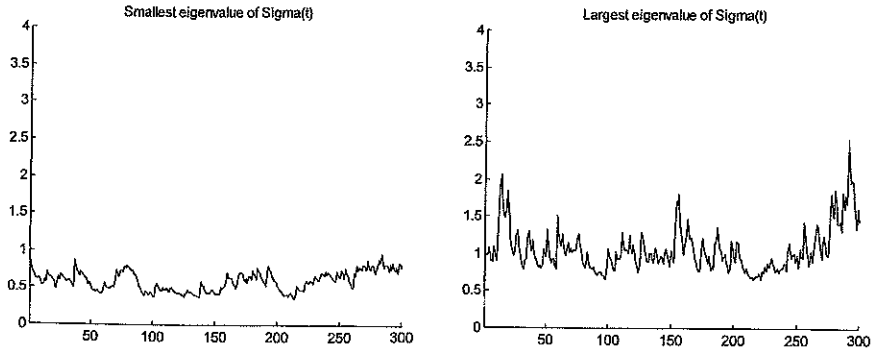

Fig. 3. Isoquantas of the log-likelihood function of model (12).

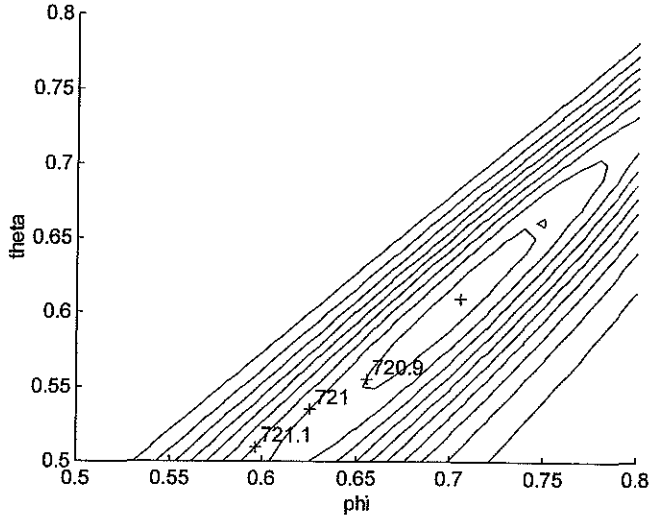

Figure 4. Scaled yields after intervention in FF.

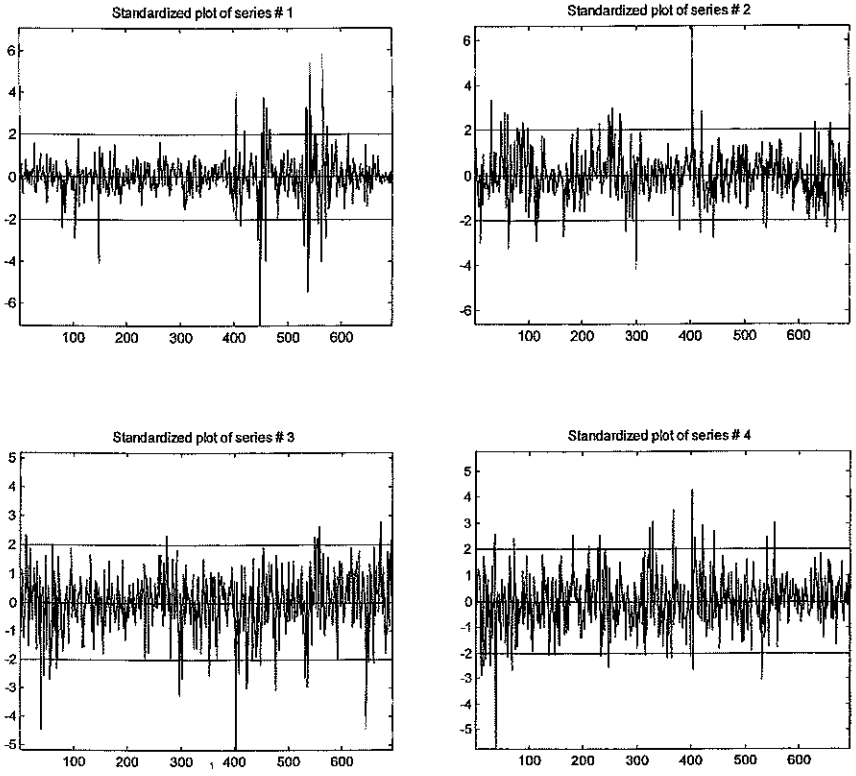

Figure 5. Estimated conditional volatilities.

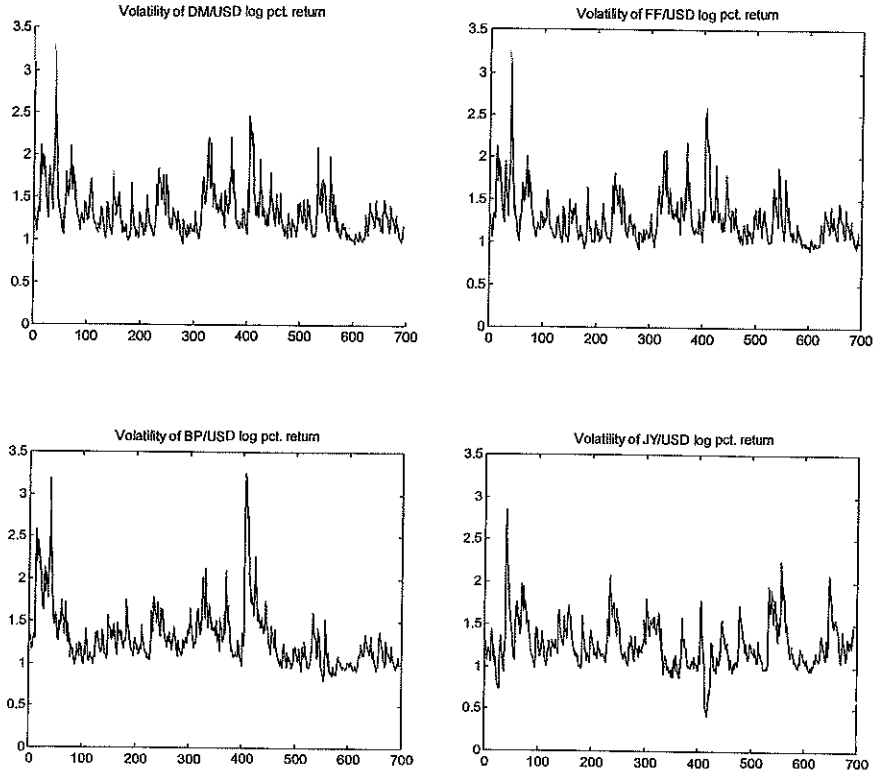

Figure 6. Estimated conditional correlations.

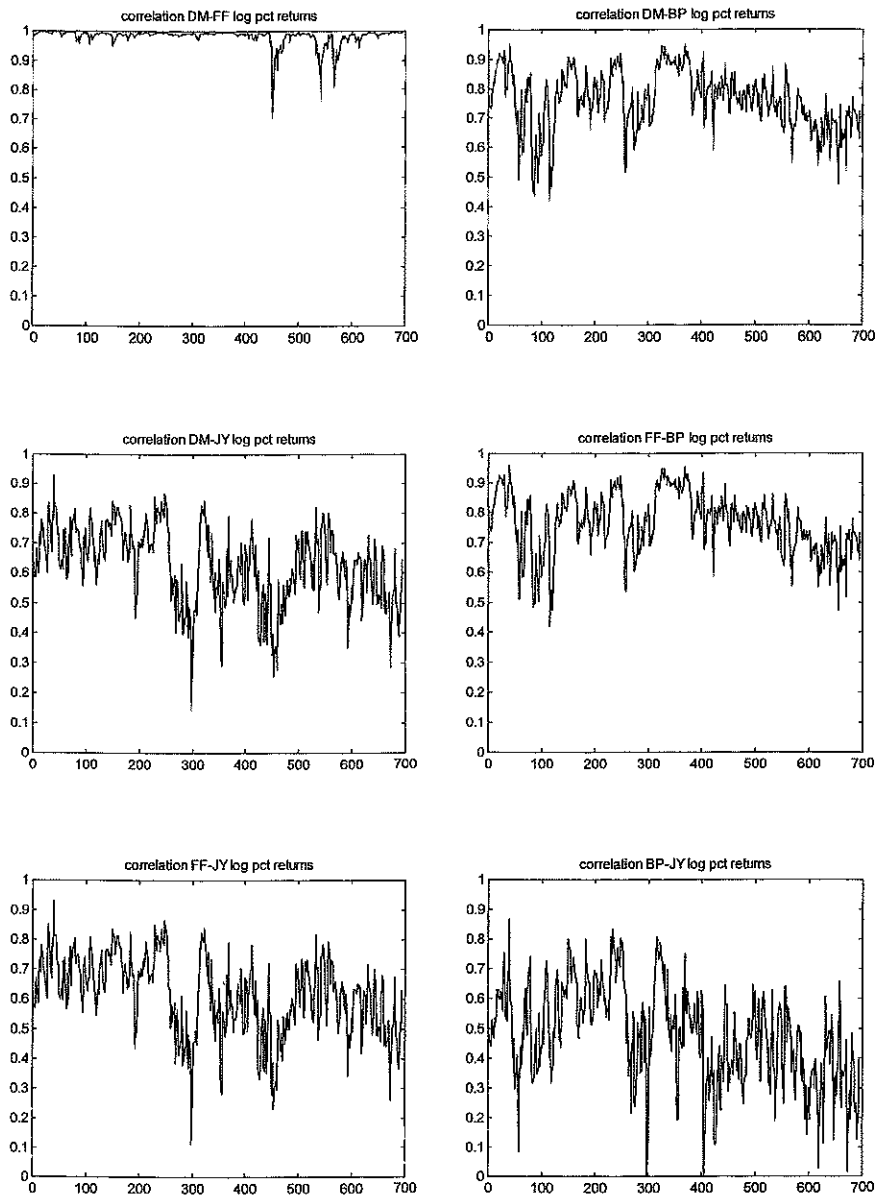

Table 1. ML estimates, correlations and principal components information.

True values	Estimates†	Correlations			Eigenvalues	Eigenvectors (by rows)		
$\sigma^2 = 1.0$	$\hat{\sigma}^2 = 1.065$ (.091)	1	--	--	1	1	0.01	-0.1
$\phi = .7$	$\hat{\phi} = .706$ (.203)	0.06	1	--	0.02	0.05	0.71	-0.71
$\theta = .6$	$\hat{\theta} = .609$ (.231)	0	0.98	1	1.98	0.04	0.71	0.71

† The figure in parentheses is the standard deviation of the estimate.

Table 2. Descriptive statistics of excess returns.

Statistic		**DM**	**FF**	**BP**	**JY**
Standard deviation		1.358	1.31	1.359	1.298
Skewness		-0.046	-0.021	0.432	-0.609
Excess Kurtosis		1.608	1.874	3.986	2.313
Sample correlations:					
DM		1	--	--	--
FF		0.978	1	--	--
BP		0.777	0.781	1	--
JY		0.635	0.623	0.477	1
Eigen-structure of the covariance matrix					
Eigenvalue	**% of var.**	**Scaled eigenvectors [matrix V in Eq. (16)]**			
0.039	0.55	3.535	-3.651	0.041	-0.078
0.472	6.66	-0.652	-0.628	1.062	0.413
0.954	13.45	-0.102	-0.123	-0.482	0.889
5.627	79.34	0.233	0.224	0.209	0.171

Table 3. Transformation coefficients and Q statistics of the scaled series.

Scaled eigenvectors [matrix V in Eq. (16)] after intervention				
	DM	FF	BP	JY
DM	4.032	-4.195	0.068	-0.088
FF	-0.652	-0.628	1.062	0.413
BP	-0.102	-0.123	-0.482	0.889
JY	0.233	0.224	0.209	0.171
Ljung-Box Q statistic (for 10 lags of the autocorrelation function of cross-products of the transformed series)†				
	Series #1	Series #2	Series #3	Series #4
Series #1	288.13	--	--	--
Series #2	42.45	19.67	--	--
Series #3	63.27	12.58	57.87	--
Series #4	28.6	23.09	84.04	25.19

† The 95% percentile of a χ^2_{10} is 18.3. As the data is not gaussian, this is only an orientative critical value of the statistic under the null of no autocorrelation.

Table 4. ML estimates of the GARCH(1,1) model (standard deviations in parentheses).

$\text{vech}(\varepsilon_t^* \varepsilon_t^{*T})$	$\hat{\sigma}_{ij}$	$\hat{\phi}_{ij}$	$\hat{\theta}_{ij}$
$(\varepsilon_{1t}^*)^2$	1 (--)	.955 (.010)	.683 (.017)
$\varepsilon_{1t}^* \varepsilon_{2t}^*$	0 (--)	.895 (.015)	.845 (.015)
$\varepsilon_{1t}^* \varepsilon_{3t}^*$	0 (--)	.273 (.009)	.238 (.008)
$\varepsilon_{1t}^* \varepsilon_{4t}^*$	0 (--)	.442 (.007)	.232 (.004)
$(\varepsilon_{2t}^*)^2$	1 (--)	.895 (.023)	.795 (.020)
$\varepsilon_{2t}^* \varepsilon_{3t}^*$	0 (--)	.936 (.012)	.846 (.014)
$\varepsilon_{2t}^* \varepsilon_{4t}^*$	0 (--)	.971 (.010)	.925 (.014)
$(\varepsilon_{3t}^*)^2$	1 (--)	.891 (.015)	.763 (.013)
$\varepsilon_{3t}^* \varepsilon_{4t}^*$	0 (--)	.957 (.025)	.880 (.020)
$(\varepsilon_{4t}^*)^2$	1 (--)	.895 (.018)	.745 (.018)
Diagnostics of estimation results:			
Gaussian likelihood (minus log) on convergence			3618.78
Square root norm of gradient			0.0773
Min. eigenvalue of scaled data covariances			0.0658
Min. eigenvalue of original data covariances			0.0046

† The parameters in this column are constrained to identity matrix values, according to the transformation (15)-(17). The minus log likelihood corresponding to this model with free covariances is 3614.52. Therefore, an LR test would not reject the constraints at the 95% confidence level.

Table 5. Statistics of standardized residuals.

	Series #1	**Series #2**	**Series #3**	**Series #4**
Skewness	-0.583	0.481	-0.735	-0.015
Excess Kurtosis	3.156	5.016	2.376	1.865
Ljung-Box Q statistic (for 10 lags of the autocorrelation function of cross-products of the standardized series)				
	Series #1	Series #2	Series #3	Series #4
Series #1	5.30	--	--	--
Series #2	4.08	4.48	--	--
Series #3	6.13	7.67	9.38	--
Series #4	8.80	16.11	5.19	9.90

† The 95% percentile of a χ^2_{10} is 18.3. As the data is not gaussian, this is only an orientative critical value of the statistic under the null of no autocorrelation.