Feasible Generalized Least Squares estimation of multivariate GARCH(1,1) models

Federico Poloni[*] and Giacomo Sbrana[†]

We provide a feasible generalized least squares estimator for (unrestricted) multivariate GARCH(1,1) models. We show that the estimator is consistent and asymptotically normally distributed under mild assumptions. Unlike the (quasi) maximum likelihood method, the feasible GLS is considerably fast to implement and does not require any complex optimization routine.

We present numerical experiments on simulated data showing the performance of the GLS estimator, and discuss the limitations of our approach.

Keywords: Multivariate GARCH(1,1) models; Feasible Generalized Least Squares estimation; Maximum Likelihood estimation

1. Introduction

Consider the unrestricted multivariate GARCH(1,1) model

$$y_t = H_t^{1/2}\epsilon_t, \quad t = 1, 2, \dots, n,$$

where $\epsilon_t \in \mathbb{R}^d$ is an i.i.d. noise vector with mean 0 and variance I_d, and the conditional covariance matrix H_t is given by

$$h_t = c + Ax_{t-1} + Bh_{t-1}, \quad t = 2, 3, \dots, n, \tag{1}$$

with $h_t := \operatorname{vech}(H_t)$, $x_t := \operatorname{vech}(y_t y_t^T)$, $\bar{d} = \frac{d(d+1)}{2}$, $A, B \in \mathbb{R}^{\bar{d}\times\bar{d}}, c \in \mathbb{R}^{\bar{d}}$.

Here, $\operatorname{vech}(M)$ represents the operator that stacks the elements of the lower triangular part of a symmetric matrix M to form a $\bar{d} \times 1$ vector.

[*]Dipartimento di Informatica; Largo B. Pontecorvo, 3; 56127, Pisa, Italy. fpoloni@di.unipi.it. Partly supported by INDAM (Istituto Nazionale di Alta Matematica).

[†]NEOMA Business School; 1, rue du Maréchal Juin, 76130 Mont-Saint-Aignan, France. gsb@neoma-bs.fr

Multivariate GARCH models are traditionally difficult to estimate, due to the large number of free parameters Bauwens et al. [2006], Francq and Zakoïan [2010]. Quasi-maximum likelihood (QML) estimators are slow to converge and complicated to implement (see Bauwens et al. [2006]). In Sbrana and Poloni [2013], we recently introduced a closed-form estimator based on the use of the moments and on linear algebra techniques. This estimator is consistent under standard assumptions; namely,

Assumption 1 H_t is positive definite almost surely;

Assumption 2 $\rho(A + B) < 1$ and $\rho(B) < 1$;

Assumption 3 the fourth moments of y_t exist and are finite;

Assumption 4 the GARCH model is identifiable, stationary, ergodic and strongly mixing.

When the distribution of ϵ_t is spherical, [Hafner, 2003, Theorem 3] gives an algebraic condition equivalent to Assumption 3 that is easy to test in practice. However, we do not need to assume sphericity here.

Assumption 4 may appear strong at first sight, but [Francq and Zakoïan, 2010, Theorem 11.5] and Boussama [2006] prove that it holds under mild sufficient conditions on the noise. Normality of the estimator can be proved only under the condition that the *eighth* moments of y_t exist finite, however. Another drawback is that its convergence rate is determined by that of the moments, which is notoriously quite slow. In Sbrana and Poloni [2013], we suggest using this estimator as a starting point for the QML maximization procedure; however, this means that we are led back to QML, with all its numerical issues.

In this paper, we suggest an alternative that avoids completely the use of QML, that is, a FGLS-type estimator for the unrestricted multivariate GARCH(1,1). We prove that this estimator is consistent and asymptotically normal distributed (with only Assumption 3 on the moments) when started with a consistent initial value, such as the aforementioned closed-form estimator. Moreover, its asymptotic variance is equal to that of the QML estimator. The accuracy of this new two-step estimator is good enough to be comparable with that of QML, while it is considerably faster and the whole procedure is simple enough to describe and implement, without relying on advanced optimization techniques.

Our proof generalizes the one in [Francq and Zakoïan, 2010, Theorem 6.3] for the univariate ARCH(p), and follows in some steps the one of Comte and Lieberman [2003] and Hafner and Preminger [2009] for the asymptotic properties of the QML estimator. Nevertheless, we aim to give a self-contained exposition in this paper.

The following additional assumption is needed for proving the asymptotic properties.

Assumption 5 $C = \operatorname{vech}^{-1}(c) > 0$, and A and B are such that $A \operatorname{vech}(X) > 0$ and $B \operatorname{vech}(X) > 0$ for each $X \geq 0$, $X \neq 0$.

This is a technical hypothesis; note that, if it is adopted, Assumption 1 and the second part of Assumption 2 are automatically satisfied. This assumption reduces to $A > 0$, $B > 0$, $c > 0$ for the univariate GARCH; it is a direct generalization of the similar assumptions made for the univariate ARCH(p) in [Francq and Zakoïan, 2010, Theorem 6.3].

Identifiability in Assumption 4 means that the task of estimating the model is well-posed and has a unique solution; it is therefore unavoidable.

Lastly, a word of warning on another important issue is needed. Identifying a simple condition on A, B and c that is both necessary and sufficient to guarantee Assumption 1 is still an open problem; hence, it is possible that the parameter values produced by estimators such as ours do not represent a well-posed model. Projection or truncation methods can be introduced at different stages to fix this issue; for instance, whenever the recurrence produces a matrix H_t which is not positive-definite, one can alter artificially its negative eigenvalues and set them to zero; or one can modify A, B and c to enforce Assumption 5.

This is a problem that may well arise in empirical analysis, especially when dealing with small samples or when the GARCH model only approximates the data (misspecification issues). The asymptotic properties of the estimator that we are considering are not affected by this issue, since they only consider the limit for $n \to \infty$ for a time series produced by a model satisfying the assumptions; similarly, any ad-hoc method to fix or exclude non-positive-definiteness will not affect the asymptotic properties of an estimator, since it will only be applied when the parameters are outside of a suitable neighbourhood of their limit values. Still, this issue has to be considered and treated in numerical code when implementing estimators for multivariate GARCH models working on empirical data.

2. The multivariate feasible generalized least-squares estimator

We shall denote with $M^{1/2}$ the unique positive definite square root of a positive definite matrix M, and with $\rho(M)$ the spectral radius, i.e., the maximum of $|\lambda|$ over the eigenvalues λ of M. Finally, with $\|M\|$ we denote the Euclidean norm on vectors, and the induced operator norm $\|M\| := \max_{\|v\|=1} \|Mv\|$ on matrices. By $\|M\|_F$ we denote the Frobenius norm of a matrix, i.e., the sum-of-squares norm, $\|M\|_F := \|\mathrm{vec}(F)\|$.

We define a sequence of estimators

$$\hat{\theta}_\ell := \begin{bmatrix} \hat{c}_\ell \\ \mathrm{vec}(\hat{A}_\ell) \\ \mathrm{vec}(\hat{B}_\ell) \end{bmatrix},$$

starting from the closed-form estimator $(\hat{c}_0, \hat{A}_0, \hat{B}_0)$ of the previous section, each one a refinement of the previous.

We describe in the following the generic step of computing $\hat{\theta}_{\ell+1}$ from $\hat{\theta}_\ell$; for simplicity, we drop the subscript ℓ.

The idea of the estimator is that the family of random variables $\mathit{\Xi}_t := H_t^{-1/2}(y_t y_t^T - H_t)H_t^{-1/2} = \epsilon_t \epsilon_t^T - I$ are i.i.d. with zero mean; hence we can run a least squares estimator on them. We use the definition of H_t to rewrite this as

$$\mathit{\Xi}_t = H_t^{-1/2}(y_t y_t^T - \mathrm{vech}^{-1}(c + Ax_{t-1} + Bh_{t-1}))H_t^{-1/2}, \quad t = 2, 3, \ldots, n. \tag{2}$$

Now we employ the previous estimate $\hat{\theta}_\ell = \hat{\theta}$ to derive an approximation $\hat{H}_t$ to use in (2) instead of the unknown exact value H_t. Hence we choose an arbitrary initial value $\hat{h}_1$, and define recursively

$$\hat{h}_t = \hat{c} + \hat{A}x_{t-1} + \hat{B}\hat{h}_{t-1}, \quad t = 2, 3, \dots, n.$$

Equation (2) becomes

$$\hat{\Xi}_t = \hat{H}_t^{-1/2}(y_t y_t^T - \mathrm{vech}^{-1}(c + Ax_{t-1} + B\hat{h}_{t-1}))\hat{H}_t^{-1/2}, \quad t = 2, 3, \dots, n. \tag{3}$$

This is a linear function in c, A, B; therefore, we can use linear least squares to compute

$$(\hat{c}_{\ell+1}, \hat{A}_{\ell+1}, \hat{B}_{\ell+1}) = \arg\min_{c,A,B} \sum_{t=2}^{n} \left\| \hat{H}_t^{-1/2}(y_t y_t^T - \mathrm{vech}^{-1}(c + Ax_{t-1} + B\hat{h}_{t-1}))\hat{H}_t^{-1/2} \right\|_F, \tag{4}$$

which is the next step of our iterative estimator.

Now, we vectorize everything to transform this least squares problem into a more common vector form. For a symmetric matrix H and $h = \mathrm{vech}\, H$, it holds that $\|H\|_F = h^T W h$, where W is the diagonal matrix with elements

$$W_{ii} = \begin{cases} 1 & \text{if } i = \frac{j(j+1)}{2} \text{ for some integer } j, \\ 2 & \text{otherwise.} \end{cases} \tag{5}$$

This weighting matrix is needed because off-diagonal elements appear twice in $\|H\|_F$, while diagonal elements only once. We note that $W = D_d^T D_d$, where D_d is the duplication matrix Abadir and Magnus [2005][1].

We define $\dot{h}_t = \begin{bmatrix} 1 & x_{t-1}^T & h_{t-1}^T \end{bmatrix} \otimes I_{\bar{d}}$, so that (1) can be rewritten as

$$h_t = \dot{h}_t \theta, \quad \theta = \begin{bmatrix} c \\ \mathrm{vec}(A) \\ \mathrm{vec}(B) \end{bmatrix}.$$

Note that $\frac{\partial h_t}{\partial \theta} = \dot{h}_t$, which justifies our choice of the notation.

Moreover, we denote by $\hat{\mathcal{H}}_t$ the $\bar{d} \times \bar{d}$ matrix such that for every symmetric $M \in \mathbb{R}^{d \times d}$ we have $\hat{\mathcal{H}}_t \,\mathrm{vech}\, M = \mathrm{vech}(\hat{H}_t^{1/2} M \hat{H}_t^{1/2})$; using the language of elimination and duplication matrices Abadir and Magnus [2005], we could write it explicitly as

$$\hat{\mathcal{H}}_t = L_d(\hat{H}_t^{1/2} \otimes \hat{H}_t^{1/2}) D_d.$$

However, no further manipulation of duplication matrices is required in the following.

[1] One could get rid of W by changing norm in (4) as follows:

$$\arg\min_{c,A,B} \sum_{t=2}^{n} \left\| \mathrm{vech} \left[\hat{H}_t^{-1/2}(y_t y_t^T - \mathrm{vech}^{-1}(c + Ax_{t-1} + B\hat{h}_{t-1}))\hat{H}_t^{-1/2} \right] \right\|.$$

We have now all the ingredients to restate (4) in vector form as

$$\begin{bmatrix} W^{1/2}\hat{\mathcal{H}}_2^{-1}\hat{\dot{h}}_2 \\ W^{1/2}\hat{\mathcal{H}}_3^{-1}\hat{\dot{h}}_3 \\ \vdots \\ W^{1/2}\hat{\mathcal{H}}_n^{-1}\hat{\dot{h}}_n \end{bmatrix} \hat{\theta}_{\ell+1} = \begin{bmatrix} W^{1/2}\hat{\mathcal{H}}_2^{-1}x_2 \\ W^{1/2}\hat{\mathcal{H}}_3^{-1}x_3 \\ \vdots \\ W^{1/2}\hat{\mathcal{H}}_n^{-1}x_n \end{bmatrix}$$

Forming normal equations, we obtain an explicit solution as $\hat{\theta}_{\ell+1} = \hat{Q}^{-1}\hat{R}$, with

$$\hat{Q} = \frac{1}{n}\sum_{t=2}^{n} \hat{\dot{h}}_t^T \hat{\mathcal{H}}_t^{-1} W \hat{\mathcal{H}}_t^{-1} \hat{\dot{h}}_t, \quad \hat{R} = \frac{1}{n}\sum_{t=2}^{n} \hat{\dot{h}}_t^T \hat{\mathcal{H}}_t^{-1} W \hat{\mathcal{H}}_t^{-1} x_t.$$

In the scalar case ($d = 1$), the normal equations reduce to [Kristensen and Linton, 2006, Equation 17].

3. Asymptotic properties

We use the symbols $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{\mathcal{L}}$ to denote almost sure convergence and convergence in law, respectively. The following consistency and normality results hold.

Theorem 1 (consistency)**.** *Suppose that one step of the FGLS estimator is run with initial values $\hat{\theta}$ such that $\hat{\theta} \xrightarrow{a.s.} \theta$, and that Assumptions 1–5 hold. Then, $\hat{Q}^{-1}\hat{R} \xrightarrow{a.s.} \theta$.*

Theorem 2 (normality)**.** *Suppose that one step of the FGLS estimator is run with initial values $\hat{\theta}$ such that $\hat{\theta} \xrightarrow{a.s.} \theta$, and that Assumptions 1–5 hold. Then $\sqrt{n}\left(\hat{Q}^{-1}\hat{R} - \theta\right) \xrightarrow{\mathcal{L}} N(0, Q^{-1}VQ^{-1})$, where $V := \mathrm{Var}\left[\dot{h}_t^T \mathcal{H}_t^{-1} W \,\mathrm{vec}(\epsilon_t\epsilon_t^T - I)\right]$ and $Q := \mathbb{E}\left[\dot{h}_t^T \mathcal{H}_t^{-1} W \mathcal{H}_t^{-1} \dot{h}_t\right]$.*

If the noise is Gaussian, a simpler expression for the asymptotic covariance holds.

Corollary 3. *If the GARCH model is driven by innovations ε_t that are Gaussian and independent, then $V = 2Q$ and thus (under the previous assumptions) $\sqrt{n}(\hat{Q}^{-1}\hat{R} - \theta) \xrightarrow{\mathcal{L}} N(0, 2Q^{-1})$.*

All proofs can be found in the Appendix. We note that the asymptotic variance of this estimator is the same as that of the QML estimator, as provided in Comte and Lieberman [2003]. This is an important observation, because it shows that the asymptotic efficiency of the two is the same. However, the two estimators do not coincide for finite samples.

4. Numerical experiments

Here we evaluate the quality of the Feasible GLS estimator through a Monte Carlo simulation. Restrictions on the matrices A, B, c are usually imposed since the number of parameters in (1) might be very large when the dimension d of the GARCH increases. In the numerical experiments we consider the so-called Diagonal VEC Bollerslev et al.

[1988], which is a popular choice among practitioners (Chrétien and Ortega [2012]). Other restrictions can be used in an analogous way. Indeed, the setup of embedding the parameters in a vector θ and defining $h_t = \dot{h}_t\theta$ makes it easy to adapt our results to different parametrizations. In the Diagonal VEC case one assumes that the conditional covariance matrix follows

$$H_t = C^\circ + A^\circ \odot (y_{t-1}y_{t-1}^T) + B^\circ \odot H_{t-1}, \tag{6}$$

where $\odot$ is the Hadamard (i.e., component-by-component) product. This is a special case of (1) with $A = \operatorname{diag}[\operatorname{vech}(A^\circ)]$, $B = \operatorname{diag}[\operatorname{vech}(B^\circ)]$, $c = \operatorname{vech}(C^\circ)$. Assumption 5 is then satisfied when the symmetric matrices $C^\circ, A^\circ, B^\circ$ as well as H_1 are positive definite.

The FGLS refinement procedure in this case takes a simpler form, since there are less parameters to estimate. Namely, we can set

$$\theta = \begin{bmatrix} \operatorname{vech} C^\circ \\ \operatorname{vech} A^\circ \\ \operatorname{vech} B^\circ \end{bmatrix},$$

(and $\hat{\theta}$ accordingly) and replace $\hat{\dot{h}}_t$ with

$$\hat{\dot{h}}_t^\circ = \begin{bmatrix} I_{\bar{d}} & \operatorname{diag}(x_{t-1}) & \operatorname{diag}(\hat{h}_{t-1}) \end{bmatrix}. \tag{7}$$

Notice that $\hat{\dot{h}}_t^\circ$ is a subset of the columns of $\hat{\dot{h}}_t$ in the unrestricted case. Therefore both the formulas for the FGLS estimator and the proofs continue to hold.

We simulate the Diagonal GARCH process as in (6) with three different sets of parameter values and three different dimensions $d = 2, 3, 4$. The three sets of parameter values have been generated as follows:

$$\text{Model 1} \quad (C^\circ)_{ij} = \begin{cases} 0.2 & i = j \\ 0.15 & i \neq j \end{cases} \quad (A^\circ)_{ij} = \begin{cases} 0.15 & i = j \\ 0.1 & i \neq j \end{cases} \quad (B^\circ)_{ij} = \begin{cases} 0.25 & i = j \\ 0.2 & i \neq j, \end{cases}$$

$$\text{Model 2} \quad (C^\circ)_{ij} = \begin{cases} 0.2 & i = j \\ 0.15 & i \neq j \end{cases} \quad (A^\circ)_{ij} = \begin{cases} 0.25 & i = j \\ 0.2 & i \neq j \end{cases} \quad (B^\circ)_{ij} = \begin{cases} 0.35 & i = j \\ 0.3 & i \neq j, \end{cases}$$

$$\text{Model 3} \quad (C^\circ)_{ij} = \begin{cases} 0.2 & i = j \\ 0.15 & i \neq j \end{cases} \quad (A^\circ)_{ij} = \begin{cases} 0.35 & i = j \\ 0.3 & i \neq j \end{cases} \quad (B^\circ)_{ij} = \begin{cases} 0.45 & i = j \\ 0.4 & i \neq j. \end{cases}$$

We recall here that Assumption 2 for a diagonal GARCH translates to the fact that all *elements* of $A^\circ + B^\circ$ and of B° are between -1 and 1.

The number of parameters to estimate is 12, 18 and 30 respectively for $d = 2$, $d = 3$ and $d = 4$. For each choice of parameter values, we chose $n = 300$, 600 and 1000 observations, and simulated 1000 data sets for each of the 9 cases. For each generation, we estimate and compare the performance of the three competitors by reporting the mean squared errors. In each of them, we ran 10 iterations of the feasible GLS estimator, and chose among them the iterate that minimizes the error measure

$$\frac{1}{n}\sum_{t=1}^{n}\|x_t - h_t\|.$$

The results are compared with those of the Closed Form estimator (CF) from Sbrana and Poloni [2013] cited above, that we use as the initial value, and the Quasi-Maximum likelihood (ML), using the implementation included in the MFE Toolbox Sheppard [2009]. The results are reported in Table 1.

The CF estimator clearly reports the worst performance compared to the other alternatives. In particular, the CF estimator seems to be badly affected by the increase of dimensionality and the increase in the spectral radius of $A^\circ + B^\circ$ (as expected by the convergence theory). On the contrary, the QMLE reports the best performance. This is especially evident for larger values of $\rho(A^\circ + B^\circ)$ and also (but not so evidently) when the number of variables increases. The GLS is the second best, closer to the MLE in terms of performance. Indeed, for Model 1, QML and GLS report similar performance. As for Model 2, the GLS still reports good performance and the results for C° and A° are close to those of the QML. On the other hand, the results of the GLS for B° seem to degrade. Finally, for Model 3, the performance of the GLS is slightly better than those reported in Model 1 and Model 2. However, the results for B° are still clearly worse than those of the QMLE.

It should be noted that the GLS uses the CF estimator as starting value of the iterative estimation. Therefore, when the initial values are far from the true ones, the GLS is penalized. However, despite this issue, we can observe that, compared with the CF, the GLS improves this initial value remarkably in terms of performance and seems to be affected by neither the dimensionality nor the roots of $A^\circ + B^\circ$.

In terms of computation time, the GLS is remarkably faster than the QML. Indeed, it took us about one week to obtain the simulation results for the MLE in the case $d = 4, n = 1000$ and only six hours for the GLS. Moreover, it took about about 4 days to obtain the results for the QML in the case $d = 4, n = 600$ and only three hours for the GLS. Hence, the QML is heavily affected by the curse of dimensionality and it would get computationally unfeasible for matrices of dimension $d > 5$.

To sum up, the GLS might be either a good alternative to the QML when the roots of $A^\circ + B^\circ$ are not too high, or it might be used as starting values in the QML optimization routines.

5. Conclusions

Even on modern computers, estimating a multivariate GARCH model remains a challenging task. In this paper we propose a feasible generalized least squares estimator for unrestricted multivariate GARCH(1,1) models. We prove that the estimator is consistent and asymptotically normally distributed under mild assumptions. Unlike the

Table 1: MSE of the parameters estimates. Each cell multiplied $\times 10^{-3}$ is the MSE obtained in the simulations.

Model 1		d=2 CF	d=2 GLS	d=2 ML	d=3 CF	d=3 GLS	d=3 ML	d=4 CF	d=4 GLS	d=4 ML
$n = 300$	C°	5.6	7.4	7.1	4.3	6.3	5.1	4.0	5.4	3.0
	A°	70.4	5.5	5.3	51.2	3.8	3.5	45.2	2.9	2.7
	B°	61.3	40.9	91.3	61.5	37.4	65.8	58.1	35.1	46.6
$n = 600$	C°	5.0	7.3	2.6	4.2	2.8	2.7	2.5	2.3	1.2
	A°	60.5	3.7	2.1	55.8	1.9	2.2	43.9	1.9	1.8
	B°	61.3	39.9	37.9	57.3	28.1	24.8	42.7	26.4	21.5
$n = 1000$	C°	4.3	2.8	1.8	2.6	1.9	1.3	3.1	2.0	1
	A°	44.5	1.7	1.1	48.4	1.3	0.8	42.2	0.9	0.6
	B°	51.5	24.1	23.6	53.8	19.6	17.9	53.1	17.5	14.6

Model 2		CF	GLS	ML	CF	GLS	ML	CF	GLS	ML
$n = 300$	C°	16.4	7.5	4.1	17.2	5.5	2.8	21.6	4.4	2.5
	A°	64.7	6.2	5.7	67.9	4.6	4.2	64.0	3.8	3.2
	B°	68.2	37.2	28.7	77.3	34.3	20.8	91.5	27.8	18.0
$n = 600$	C°	11.0	5.2	2.5	14.7	3.8	1.2	15.9	2.7	1.3
	A°	41.5	3.5	3.3	55.2	2.0	2.9	38.5	1.9	1.9
	B°	49.4	24.6	16.0	68.4	20.2	11.4	82.7	16.8	7.6
$n = 1000$	C°	6.5	3.5	1.2	10.7	2.5	0.9	13.5	1.8	0.7
	A°	25.2	2.1	1.5	46.9	1.5	1.1	47.4	1.04	0.8
	B°	37.9	19.1	8.6	60.2	16.4	6.5	71.6	11.7	5.5

Model 3		CF	GLS	ML	CF	GLS	ML	CF	GLS	ML
$n = 300$	C°	103.8	18.4	4.1	149.2	12.3	3.1	102.3	8.1	4.4
	A°	65.1	8.8	5.8	97.7	7.4	3.1	91.1	5.7	3.9
	B°	87.4	35.4	16.0	108.4	27.7	4.2	130.5	25.6	10.8
$n = 600$	C°	71.8	11.1	2.9	110.2	8.3	1.7	91.8	6.7	1.9
	A°	53.3	4.7	2.5	56.8	2.9	2.2	65.8	2.5	2.6
	B°	65.7	25.3	6.5	114.1	17.0	5.5	126.5	13.4	5.3
$n = 1000$	C°	64.1	9.1	0.9	101.9	6.1	0.7	116.4	5.7	0.6
	A°	27.7	2.2	1.8	45.9	1.1	1.2	58.7	1.2	1.2
	B°	62.3	18.1	3.1	109.6	12.9	2.4	129.9	9.3	2.1

quasi-maximum likelihood method, the feasible GLS is considerably fast to implement and does not require any complex optimization routine.

The numerical experiments show that the quasi-maximum likelihood estimation procedure still represents the best tool when both the dimension of the conditional covariance matrices and that of the sample size are small. The GLS represents a second best with respect to the QML, achieving similar performance in some of the tests, but in much shorter time. Finally, the suggested estimator can also be employed as starting value in the numerical maximization routines.

The main disadvantage of the suggested estimator is the lack of guarantees that the resulting parameters yield a model satisfying the mentioned assumptions, as discussed in the introduction; in particular, the estimated model may not be stationary or may have conditional volatility matrices H_t that are not positive semidefinite.

A. Proofs

A.1. Boundedness of $\hat{\mathcal{H}}_t^{-1}\hat{\hat{h}}_t$

Following the approach in Francq and Zakoïan [2010] for the univariate ARCH, as a preliminary step we state this lemma, whose proof turns out to be more involved in the multivariate case.

Lemma 4. *Under Assumption 5, $\left\|\hat{\mathcal{H}}_t^{-1}\hat{\hat{h}}_t\right\| = O(1)$ uniformly, for $\hat{\theta} \to \theta$ and for each $t = 2, 3, \ldots, n$.*

Proof. Since $C > 0$, $\lambda_{\min}(C) > 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue. Moreover, since the set $\Omega := \{\mathrm{vech}^{-1}(x) : x \in \mathbb{R}^{\bar{d}}, \|x\| = 1 \text{ and } \mathrm{vech}^{-1}(x) \geq 0\}$ is compact, the function $\lambda_{\min}(\mathrm{vech}^{-1}(Ax))$ has a minimum m_A on Ω. This minimum is strictly positive thanks to Assumption 5. Similarly, $m_B := \min_{x\in\Omega} \lambda_{\min}(\mathrm{vech}^{-1}(Bx)) > 0$. Hence, if $\hat{C}, \hat{A}, \hat{B}$ are close enough to their exact counterparts, then $\lambda_{\min}(\hat{C}) \geq \delta$, $m_{\hat{A}} \geq \delta$, $m_{\hat{B}} \geq \delta$ for a sufficiently small $\delta > 0$.

We shall prove the stronger fact that $\left\|\hat{\mathcal{H}}_t^{-1}\right\| \left\|\hat{\hat{h}}_t\right\| = O(1)$. First of all, we wish to replace $\left\|\hat{\mathcal{H}}_t^{-1}\right\|$ with $\left\|\hat{H}_t^{-1}\right\|$. One has for each symmetric M

$$\begin{aligned}\left\|\hat{\mathcal{H}}_t^{-1}\,\mathrm{vech}(M)\right\| &= \left\|\mathrm{vech}(\hat{H}_t^{-1/2} M \hat{H}_t^{-1/2})\right\| \leq \left\|W^{-1/2}\right\| \left\|\hat{H}_t^{-1/2} M \hat{H}_t^{-1/2}\right\|_F \\ &\leq \left\|W^{-1/2}\right\| \left\|\hat{H}_t^{-1/2}\right\|_F^2 \|M\|_F \leq \left\|W^{-1/2}\right\| \bar{d} \left\|\hat{H}_t^{-1/2}\right\|^2 \|M\|_F \\ &\leq \left\|W^{-1/2}\right\| \bar{d} \left\|\hat{H}_t^{-1}\right\| \|M\|_F \leq \left\|W^{-1/2}\right\| \bar{d} \left\|\hat{H}_t^{-1}\right\| \left\|W^{1/2}\right\| \|\mathrm{vech}(M)\|\end{aligned}$$

(we have used here the fact that $\|X\|_F \leq \sqrt{\bar{d}}\,\|X\|$ for each $\bar{d} \times \bar{d}$ matrix X), thus

$$\left\|\hat{\mathcal{H}}_t^{-1}\right\| \leq \left\|W^{-1/2}\right\| \left\|W^{1/2}\right\| \bar{d} \left\|\hat{H}_t^{-1}\right\|.$$

Since $\left\|\hat{\hat{h}}_t\right\| = \sqrt{1 + \|x_{t-1}\|^2 + \left\|\hat{h}_{t-1}\right\|^2}$, it is sufficient to prove separately that

$$\left\|\hat{H}_t^{-1}\right\| \leq \delta^{-1}, \quad \left\|\hat{H}_t^{-1}\right\| \|\hat{x}_{t-1}\| \leq \delta^{-1}, \quad \left\|\hat{H}_t^{-1}\right\| \left\|\hat{h}_{t-1}\right\| \leq \delta^{-1}.$$

The first inequality follows from

$$\hat{H}_t \geq \hat{C} \geq \lambda_{\min}(\hat{C}) I \geq \delta I,$$

the second from

$$\begin{aligned}\hat{H}_t \geq \operatorname{vech}^{-1}(\hat{A} x_{t-1}) &= \|x_{t-1}\| \operatorname{vech}^{-1}\left(\hat{A} \frac{x_{t-1}}{\|x_{t-1}\|}\right) \\ &\geq \|x_{t-1}\| \, \lambda_{\min}\left(\operatorname{vech}^{-1}\left(\hat{A} \frac{x_{t-1}}{\|x_{t-1}\|}\right)\right) I \geq \|x_{t-1}\| \, m_{\hat{A}} I \geq \|x_{t-1}\| \, \delta I,\end{aligned}$$

and the third is analogous. □

A.2. An average bound on $h_t - \hat{h}_t$ and related results

Lemma 5. *Under the previous assumptions, it holds almost surely that for n large enough* $\left(\frac{1}{n}\sum_t \left\|\hat{h}_t - h_t\right\|^2\right)^{1/2} = O(\hat{\theta} - \theta)$.

Proof. To establish this result in an easier way, we take a digression to express the GARCH recurrence in terms of matrices and vectors of dimension $n\bar{d} \times n\bar{d}$. We set

$$\underline{\mathbf{h}} = \frac{1}{\sqrt{n}} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}, \quad \underline{\mathbf{x}} = \frac{1}{\sqrt{n}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \underline{\mathbf{c}} = \frac{1}{\sqrt{n}} \begin{bmatrix} h_1 \\ c \\ c \\ \vdots \\ c \end{bmatrix},$$

$$\underline{\mathbf{A}} = \begin{bmatrix} 0 & & & & \\ A & 0 & & & \\ & A & 0 & & \\ & & \ddots & \ddots & \\ & & & A & 0 \end{bmatrix}, \quad \underline{\mathbf{B}} = \begin{bmatrix} 0 & & & & \\ B & 0 & & & \\ & B & 0 & & \\ & & \ddots & \ddots & \\ & & & B & 0 \end{bmatrix}.$$

Now we can rewrite the GARCH(1,1) model as

$$\underline{\mathbf{h}} = \underline{\mathbf{c}} + \underline{\mathbf{A}}\underline{\mathbf{x}} + \underline{\mathbf{B}}\underline{\mathbf{h}}$$

or

$$\underline{\mathbf{h}} = (I_{n\bar{d}} - \underline{\mathbf{B}})^{-1} \underline{\mathbf{c}} + (I_{n\bar{d}} - \underline{\mathbf{B}})^{-1} \underline{\mathbf{A}}\underline{\mathbf{x}}.$$

We have

$$\|\underline{\mathbf{x}}\|^2 = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{\bar{d}} (x_t)_i^2 \xrightarrow{\text{a.s.}} \operatorname{Trace} \mathbb{E}\left[x_t x_t^T\right],$$

thus $\|\underline{\mathbf{x}}\|^2 \leq 2 \operatorname{Trace} \mathbb{E}\left[x_t x_t^T\right] = O(1)$ holds almost surely. Moreover,

$$\|\underline{\mathbf{c}}\|^2 = \frac{1}{n}(\|h_1\|^2 + (n-1)\|c\|^2) = O(1)$$

By a continuity argument, using boundedness and the fact that $\|\underline{\mathbf{A}}\| = \|A\|$ and $\left\|\underline{\mathbf{B}}^k\right\| = \left\|B^k\right\|$ for each k, one gets that

$$\left\|\hat{\underline{\mathbf{h}}} - \underline{\mathbf{h}}\right\| = \left\|(I - \hat{\underline{\mathbf{B}}})^{-1}\left(\hat{\underline{\mathbf{c}}} + \hat{\underline{\mathbf{A}}}\underline{\mathbf{x}}\right) - (I - \underline{\mathbf{B}})^{-1}\left(\underline{\mathbf{c}} + \underline{\mathbf{A}}\underline{\mathbf{x}}\right)\right\| = O(\hat{\theta} - \theta).$$

□

From the previous result we can infer some additional results. Looking back at the definition of $\dot{h}_t$, it is clear that this relation also implies

$$\left(\sum_t \frac{1}{n}\left\|\hat{\dot{h}}_t - \dot{h}_t\right\|^2\right)^{1/2} = O(\hat{\theta} - \theta).$$

Since all norms are equivalent, $\left(\sum_t \frac{1}{n}\left\|\hat{H}_t - H_t\right\|^2\right)^{1/2} = O(\hat{\theta} - \theta)$ holds as well. We also have

$$\left\|\hat{H}_t^{-1} - H_t^{-1}\right\| = \left\|\hat{H}_t^{-1}(\hat{H}_t - H_t)H_t^{-1}\right\| \leq \left\|\hat{H}_t^{\,-1}\right\|\left\|\hat{H}_t - H_t\right\|\left\|H_t^{-1}\right\|,$$

thus, since the $\hat{H}_t$ are bounded away from zero,

$$\left(\sum_t \frac{1}{n}\left\|\hat{H}_t^{-1} - H_t^{-1}\right\|^2\right)^{1/2} = O(\hat{\theta} - \theta). \tag{8}$$

A.3. Consistency of $\hat{Q}$

Theorem 6. *Suppose that one step of the FGLS estimator is run with initial values $\hat{\theta}$ such that $\hat{\theta} \xrightarrow{a.s.} \theta$, and that Assumptions 1–5 hold. Then, $\hat{Q} \xrightarrow{a.s.} Q$.*

Proof. We have almost surely

$$\frac{1}{n}\sum\left\|\hat{H}_t^{-1}\hat{\dot{H}}_{t,j} - H_t^{-1}\dot{H}_{t,j}\right\| \leq \frac{1}{n}\sum\left\|\hat{H}_t^{-1}\right\|\left\|\hat{\dot{H}}_{t,j} - \dot{H}_{t,j}\right\| + \left\|\hat{H}_t^{-1} - H_t^{-1}\right\|\left\|\dot{H}_{t,j}\right\| = O(\hat{\theta} - \theta) \tag{9}$$

after applying the Cauchy-Schwarz inequality to both terms to reduce to the previously established bounds.

The (i, j) entry of $\hat{Q}$ is given by

$$\frac{1}{n}\operatorname{tr}\sum_{t=1}^{n}\hat{\dot{H}}_{t,i}\hat{H}_t^{-1}\hat{\dot{H}}_{t,j}\hat{H}_t^{-1}. \tag{10}$$

(compare the last expression with [Comte and Lieberman, 2003, Equation (A3)] and Hafner and Herwartz [2008]). We use the triangle inequality and Cauchy-Schwarz to get

$$\begin{aligned}\frac{1}{n}\operatorname{tr}\sum\left\|\hat{\dot{H}}_{t,i}\hat{H}_t^{-1}\hat{\dot{H}}_{t,j}\hat{H}_t^{-1} - \dot{H}_{t,i}H_t^{-1}\dot{H}_{t,j}H_t^{-1}\right\| &\leq \frac{1}{n}\operatorname{tr}\sum\left\|\hat{\dot{H}}_{t,i}\hat{H}_t^{-1} - \dot{H}_{t,i}H_t^{-1}\right\|\left\|\hat{\dot{H}}_{t,j}\hat{H}_t^{-1}\right\| \\ &+ \frac{1}{n}\operatorname{tr}\sum\left\|\dot{H}_{t,i}H_t^{-1}\right\|\left\|\hat{\dot{H}}_{t,j}\hat{H}_t^{-1} - \dot{H}_{t,j}H_t^{-1}\right\| = O(\hat{\theta} - \theta),\end{aligned}$$

since $\left\|\hat{\dot{H}}_{t,j}H_t^{-1}\right\| = O(1)$ by Lemma 4.

Going back to the vectorized setting, this means that

$$\frac{1}{n}\sum\left(\hat{\dot{h}}_t^T\hat{\mathcal{H}}_t^{-1}W\hat{\mathcal{H}}_t^{-1}\hat{\dot{h}}_t - \dot{h}_t^T\mathcal{H}_t^{-1}W\mathcal{H}_t^{-1}\dot{h}_t\right) = O(\hat{\theta}-\theta).$$

By consistency of $\hat{\theta}$, $\left\|\hat{\theta}-\theta\right\| \xrightarrow{\text{a.s.}} 0$, so $\hat{Q}$ converges to the same limit as

$$\frac{1}{n}\sum\dot{h}_t^T\mathcal{H}_t^{-1}W\mathcal{H}_t^{-1}\dot{h}_t,$$

which equals Q by the law of large numbers. $\square$

The matrix Q is considered in [Comte and Lieberman, 2003, Appendix A], where it is proved that it is equal to the asymptotic value of the Hessian matrix of the Gaussian log-likelihood function, which coincides with the Fisher information matrix of the process. Under our assumptions, the argument in [Comte and Lieberman, 2003, Appendix A] holds verbatim and shows that it is nonsingular: essentially, if it had a zero eigenvector, then we could identify a nontrivial linear relation that holds almost surely between c, Ax_{t-1} and Bh_{t-1}, and use it to construct an alternative model of the GARCH, contradicting identificability.

A.4. Consistency of the FGLS estimator

Proof of Theorem 1. After recalling that $h_t = \dot{h}_t\theta$ and $\mathcal{H}_t^{-1}(x_t - h_t) = \mathrm{vech}(\epsilon_t\epsilon_t^T - I)$, we have

$$\begin{aligned}
\hat{R}-\hat{Q}\theta &= \frac{1}{n}\sum\hat{\dot{h}}_t^T\hat{\mathcal{H}}_t^{-1}W\hat{\mathcal{H}}_t^{-1}x_t - \frac{1}{n}\sum\hat{\dot{h}}_t^T\hat{\mathcal{H}}_t^{-1}W\hat{\mathcal{H}}_t^{-1}\hat{\dot{h}}_t\theta \\
&= \frac{1}{n}\sum\hat{\dot{h}}_t^T\hat{\mathcal{H}}^{-1}W\left(\hat{\mathcal{H}}_t^{-1}-\mathcal{H}_t^{-1}\right)x_t + \frac{1}{n}\sum\hat{\dot{h}}_t^T\hat{\mathcal{H}}^{-1}W\,\mathrm{vech}(\epsilon_t\epsilon_t^T - I) \qquad (11)\\
&\quad + \frac{1}{n}\sum\hat{\dot{h}}_t^T\hat{\mathcal{H}}^{-1}W\left(\hat{\mathcal{H}}_t^{-1}\hat{\dot{h}}_t - \mathcal{H}_t^{-1}\dot{h}_t\right)\theta.
\end{aligned}$$

We call S_1, S_2 and S_3 the three terms of this sum. All of them converge to zero almost surely: the first summand S_1 because of (8) and Lemma 4; the second term S_2 thanks to the law of large numbers for martingale difference sequences Meyn and Tweedie [2009], and S_3 by (9) and Lemma 4 once again.

This shows that $\hat{R}-\hat{Q}\theta \xrightarrow{\text{a.s.}} 0$, and thus also $\hat{Q}^{-1}\hat{R} \xrightarrow{\text{a.s.}} \theta$, since $\hat{Q}$ is asymptotically nonsingular. $\square$

A.5. Normality of the FGLS estimator

Proof of Theorem 2. The key of the proof is once again the decomposition (11). The terms $\sqrt{n}S_1$ and $\sqrt{n}S_3$ still converge almost surely to zero, thanks to the previous convergence and boundedness results. The term $\sqrt{n}S_2$ converges in law to $N(0,V)$ thanks to the central limit theorem for martingale difference sequences Meyn and Tweedie [2009].

Thus $\sqrt{n}\left(\hat{R}-\hat{Q}\theta\right) \xrightarrow{\mathcal{L}} N(0,V)$. The factor $\hat{Q}^{-1}$, which converges a.s. to Q^{-1}, gives the two outer factors in the covariance. $\square$

A.6. Gaussian noise

Lemma 7. *Suppose that the random variable ϵ follows a multivariate $N(0,I)$ distribution. Then,*

$$\operatorname{Var}\left[\operatorname{vech}(\epsilon\epsilon^T - I)\right] = 2W^{-1}$$

Proof. The entries of this covariance matrix have the form $\mathbb{E}\left[(\epsilon_i\epsilon_j - \delta_{ij})(\epsilon_k\epsilon_l - \delta_{kl})\right]$, with δ denoting the Kronecker delta symbol. We divide into the following cases:

1. $i \neq j$ and $k \neq l$, $\{i,j\} \neq \{k,l\}$: at least one of the components appears with exponent 1, thus $\mathbb{E}\left[\epsilon_i\epsilon_j\epsilon_k\epsilon_l\right] = 0$ by sphericity.
2. $i \neq j$ and $k \neq l$, $\{i,j\} = \{k,l\}$: thus we are in the case $\mathbb{E}\left[(\epsilon_i\epsilon_j)(\epsilon_i\epsilon_j)\right] = \mathbb{E}\left[\epsilon_i^2\right]\mathbb{E}\left[\epsilon_j^2\right] = 1$.
3. $i = j$, $k \neq l$ (or viceversa): $\mathbb{E}\left[(\epsilon_i^2 - 1)\epsilon_k\epsilon_l\right] = 0$ by sphericity, since one among k and l must differ from i.
4. $i = j$, $k = l \neq i$: $\mathbb{E}\left[(\epsilon_i^2-1)(\epsilon_k^2-1)\right] = \mathbb{E}\left[\epsilon_i^2\right]\mathbb{E}\left[\epsilon_j^2\right] - 1 = 0$.
5. $i = j = k = l$: $\mathbb{E}\left[(\epsilon_i^2-1)(\epsilon_i^2-1)\right] = \mathbb{E}\left[\epsilon_i^4\right] - 1 = 2$, since the kurtosis of a Gaussian variable is 3.

Hence, after checking which positions on the diagonal correspond to the case $i=j=k=l$ and which to the case $i = k \neq j = l$, we see that the required covariance matrix is diagonal and

$$\left(\operatorname{Var}\left[\operatorname{vech}(\epsilon\epsilon^T - I)\right]\right)_{ii} = \begin{cases} 2 & \text{if } i = \frac{j(j+1)}{2} \text{ for some integer } j, \\ 1 & \text{otherwise.} \end{cases}$$

Comparing this with the definition of W in (5), we obtain the desired result. $\square$

Proof of Corollary 3. The noise at time t, ε_t, is independent from $\mathcal{H}_t$ and $\dot{h}_t$, thus we can compute the expected value of

$$\begin{aligned} V &= \mathbb{E}\left[\dot{h}_t^T\mathcal{H}_t^{-1}W\operatorname{vec}(\epsilon_t\epsilon_t^T - I)\operatorname{vec}(\epsilon_t\epsilon_t^T - I)^T W\mathcal{H}_t^{-1}\dot{h}_t\right] \\ &= \mathbb{E}\left[\dot{h}_t^T\mathcal{H}_t^{-1}W\operatorname{Var}\left[\operatorname{vech}(\epsilon\epsilon^T - I)\right]W\mathcal{H}_t^{-1}\dot{h}_t\right] = 2Q. \quad \square \end{aligned}$$

References

K.M. Abadir and J.R. Magnus. *Matrix Algebra*. Econometric Exercises. Cambridge University Press, 2005. ISBN 9780521822893.

Luc Bauwens, Sébastien Laurent, and Jeroen V. K. Rombouts. Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006.

Tim Bollerslev, Robert F Engle, and Jeffrey M Wooldridge. A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–31, February 1988. URL `http://ideas.repec.org/a/ucp/jpolec/v96y1988i1p116-31.html`.

Farid Boussama. Ergodicité des chaînes de Markov à valeurs dans une variété algébrique: application aux modèles GARCH multivariés. *C. R. Math. Acad. Sci. Paris*, 343 (4):275–278, 2006. ISSN 1631-073X. doi: 10.1016/j.crma.2006.06.027. URL `http://dx.doi.org/10.1016/j.crma.2006.06.027`.

Stéphane Chrétien and Juan-Pablo Ortega. Multivariate GARCH estimation via a Bregman-proximal trust-region method. *Computational Statistics and Data Analysis*, 2012. To appear. Available as arXiv:1101.5475.

F. Comte and O. Lieberman. Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis*, 84(1):61 – 84, 2003. ISSN 0047-259X. doi: 10.1016/S0047-259X(02)00009-X. URL `http://www.sciencedirect.com/science/article/pii/S0047259X0200009X`.

C. Francq and J.M. Zakoïan. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley, 2010. ISBN 9780470683910.

Christian M. Hafner. Fourth moment structure of multivariate GARCH models. *Journal of Financial Econometrics*, 1(1):26–54, 2003. doi: 10.1093/jjfinec/nbg001. URL `http://jfec.oxfordjournals.org/content/1/1/26.abstract`.

Christian M. Hafner and Helmut Herwartz. Analytical quasi maximum likelihood inference in multivariate volatility models. *Metrika*, 67(2):219–239, 2008. ISSN 0026-1335. doi: 10.1007/s00184-007-0130-y. URL `http://dx.doi.org/10.1007/s00184-007-0130-y`.

Christian M. Hafner and Arie Preminger. On asymptotic theory for multivariate GARCH models. *J. Multivariate Anal.*, 100(9):2044–2054, 2009. ISSN 0047-259X. doi: 10.1016/j.jmva.2009.03.011. URL `http://dx.doi.org/10.1016/j.jmva.2009.03.011`.

Dennis Kristensen and Oliver Linton. A closed-form estimator for the GARCH $(1,1)$ model. *Econometric Theory*, 22(2):323–337, 2006. ISSN 0266-4666. doi: 10.1017/S0266466606060142. URL `http://dx.doi.org/10.1017/S0266466606060142`.

Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-73182-9. With a prologue by Peter W. Glynn.

Giacomo Sbrana and Federico Poloni. A closed-form estimator for the multivariate GARCH $(1, 1)$ model. *J. Multivariate Anal.*, 120:152–162, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2013.05.005. URL http://dx.doi.org/10.1016/j.jmva.2013.05.005.

Kevin Sheppard. MFE toolbox v. 4.0, 2009. Available from http://www.kevinsheppard.com/wiki/MFE_Toolbox.