ELSEVIER

Journal of Econometrics 93 (1999) 93–111

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

Efficiency comparisons of maximum-likelihood-based estimators in GARCH models

Gloria González-Rivera[a,*], Feike C. Drost[b,1]

[a]Department of Economics, University of California, Riverside, Riverside, CA 92521, USA
[b]Department of Econometrics and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Netherlands

Received 1 December 1995; accepted 1 December 1998

Abstract

In this paper, we investigate the loss of asymptotic efficiency of semiparametric and quasi-maximum-likelihood estimators relative to maximum-likelihood estimators in models with generalized autoregressive conditional heteroscedasticity (GARCH). For a general time-varying location–scale model, the factors that contribute to differences in efficiency among the estimators can be divided in two categories. One pertains to the parametric specifications of the conditional mean and the conditional variance. The other corresponds to the shape characteristics of the conditional density of the standardized errors, summarized in the coefficients of skewness and kurtosis together with the Fisher information for location and scale. The quantification of these factors has practical implications since it can help to decide if the more complex semiparametric estimator provides sufficient efficiency gains with respect to the simplest quasi-maximum-likelihood estimator. We also prove that there is no probability density function, with the exception of the normal, for which the asymptotic efficiency of the three estimators is the same. Particular models are also considered, for which the efficiency comparisons are greatly simplified. © 1999 Elsevier Science S.A. All rights reserved.

JEL classification: C13; C14; C32

Keywords: Fisher information matrix; MLE; QMLE; Semiparametric estimation

* Corresponding author. Tel.: 909-787-5037; fax: 909-787-5685.
E-mail address: ggonzale@galaxy.ucr.edu (G. González-Rivera)
[1] E-mail: f.c.drost@kub.nl.

PII: S0304-4076(99)00005-6

1. Introduction

In this paper, we compare the efficiency properties of maximum-likelihood-based estimators in the context of generalized autoregressive conditionally heteroscedastic (GARCH) models. Asymptotically, consistency is both a desirable and required property of an estimator, but the property of maximal efficiency, though desirable, is not always attainable. Asymptotic efficiency is a function of the level of information available to the researcher.

We develop a strategy for evaluating efficiency gains when estimating several types of GARCH and GARCH-in-mean models, but the methodology will be extended to other time-varying location–scale models. The estimators we consider are all based on the likelihood principle. A likelihood function is constructed based on an assumed conditional probability density function. Depending on the amount of information available, we can estimate a model with a maximum-likelihood (ML) estimator where the conditional probability density function is fully known, a semiparametric (SP) estimator where the density is estimated with a data-based procedure, or a quasi-maximum-likelihood (QML) estimator where conditional normality is assumed, though this assumption is likely to be false. The efficiency gains are directly proportional to the amount of information available; hence, ML is more efficient than SP, which is more efficient than QML estimation.

The comparison of the asymptotic variance–covariance matrices of the ML, SP, and QML estimators reveal that, in the general time-varying location–scale model, differences in efficiency are the result of the interaction between the specified model and the shape characteristics of the conditional probability density function. These interactions can be disentangled when we focus on interesting specific location–scale models. Relevant factors that contribute to differences in efficiency among the three estimators are the Fisher information for location, the Fisher information for scale, the coefficient of kurtosis and the coefficient of skewness of the conditional density. Our results have practical implications. The empirical researcher can assess the ‘closeness’ of the SP estimator to the ML and QML estimators. When comparing a QML to an SP estimator, there is a practical trade-off between simplicity in implementation and potential gains in efficiency. If the SP estimator is ‘closer’ to the QML than to the ML, the potential efficiency gains would be weighed against the costs of implementing the more complex SP estimator.

We distinguish among five types of models, depending on the relation between the conditional mean and the conditional variance. The most general model consists of a time-varying conditional mean and a time-varying conditional variance, where mean and variance are mutually dependent upon each other. The GARCH-in-mean models are examples of this class. This general model can be particularized to provide four interesting specific models. First, we consider a pure time-varying location model with homoscedastic errors, where the parameters of interest are only the mean parameters. In this group, we may include dynamic specifications as ARMA models and general regression models. The second model is a pure time-varying scale model, where the parameters of interest are the conditional variance parameters. In this group, we may include any type of conditional heteroscedasticity specification, such as the classic ARCH and GARCH, Exponential GARCH, Nonlinear GARCH, Asymmetric GARCH. The third group of models consists of a conditional mean and a conditional variance specification such that the conditional mean does not depend on the parameters of the conditional variance, and the conditional variance does not depend on the parameters of the conditional mean. In this group, we may include regression models with heteroscedasticity other than GARCH, such as multiplicative heteroscedasticity, and ARMA type models, where the conditional variance is a function of past observations. The fourth group consists of a conditional variance that depends on the full set of mean and variance parameters, but the conditional mean does not depend on the conditional variance parameters. In this group, we restrict the conditional variance to be a symmetric function of the errors as in the classical ARCH and GARCH models. In the third and fourth groups, we also impose symmetry of the probability density function of the errors.

This paper proceeds as follows: In Section 2, we describe the general time-varying location–scale model and the various estimation methodologies. In Section 3, we consider efficiency comparisons among ML, QML, and SP estimators for the general location–model. In Section 4, we examine the efficiency comparisons for four specific location–scale models. Section 5 concludes the paper.

2. Location–scale models and estimation methodology

2.1. The general location-scale model

Consider a discrete time stochastic process $\{y_t\}$ parameterized by a finite parameter vector θ. Conditioning on available information up to time $t-1$, the random variables y_t have conditional mean $m_t(\theta_0)$ and conditional variance $h_t(\theta_0)$, where θ_0 denotes the true but unknown parameter. The functions $m_t(\theta)$ and $h_t(\theta)$ may be function of past information, including lagged exogenous variables $x_t, x_{t-1}, x_{t-2},\dots,$ and of the parameter vector θ, i.e. $m_t(\theta)=m(x_t, x_{t-1}, x_{t-2},\dots,y_{t-1}, y_{t-2},\dots;\theta)$, $h_t(\theta)=h(x_t, x_{t-1}, x_{t-2},\dots,y_{t-1}, y_{t-2},\dots;\theta)$. Such processes are known as regression models with Generalized Autoregressive Conditionally Heteroscedasticity (GARCH), possibly including GARCH-in-mean behavior. We assume that the starting values of the process are taken from the stationary distribution and, hence, the process, itself, is assumed stationary. This implies that the scores will be stationary too.

Alternatively, if the starting values are observed, Drost and Klaassen (1997) show that replacing the true non-stationary scores by the corresponding stationary ones has no influence asymptotically, see also Koul and Shick (1997) for general conditions. Our model will be

$$y_t = m_t(\theta_0) + \sqrt{h_t(\theta_0)}\,u_t, \tag{1}$$

where $\{u_t\}$ is an i.i.d. sequence with zero mean, unit variance, finite fourth moments, and an absolutely continuous probability density function (pdf) g with derivative g', such that the Fisher information for location and the Fisher inform ation for scale are finite. In general, the pdf g can be defined by additional parameters, say η, that are considered nuisance parameters, but may contain relevant information for the estimation of the vector of parameters of interest, θ.

This paper is concerned with the specification of the conditional density function g and with the efficiency properties of maximum-likelihood-based estimators of the parameter vector θ. We assume that the models for $m_t(\theta)$ and $h_t(\theta)$ are correctly specified. We use three estimation methods, maximum-likelihood (ML), quasi-maximum-likelihood (QML), and semiparametric (SP) estimation, depending on how much information we have available regarding the pdf. The issue of efficiency is directly related to knowledge surrounding g.

The assumption of i.i.d. $\{u_t\}$ innovations may be too restrictive for parametric maximum-likelihood methods; in fact, a weaker assumption as $\{u_t\}$ being a martingale difference sequence suffices to render the QML estimator asymptotically normal (Bollerslev and Wooldridge, 1992). In this paper, we retain the i.i.d. assumption, which is mostly adopted in the current parametric and semiparametric literature. Optimality in a more general framework is a subject that deserves further research.

2.2. Notation

Let $u_t(\theta) = (y_t - m_t(\theta))/\sqrt{h_t(\theta)}$ denote calculated residuals. Inserting the true but unknown value of the parameter θ_0 yields $u_t(\theta_0) = u_t$. We define the vector function $\psi = (\psi_1, \psi_s)'$, based upon the location–scale scores, by

$$\psi_1(u_t) = -\frac{g'(u_t)}{g(u_t)}$$

and

$$\psi_s(u_t) = -\left(1 + u_t\frac{g'(u_t)}{g(u_t)}\right).$$

In the special case of a standard normal density the function ψ reduces to

$$F(u) = \begin{pmatrix} u \\ u^2 - 1 \end{pmatrix}.$$

Define the matrix W as

$$W_t = [W_{lt}, W_{st}] = \left[\frac{1}{\sqrt{h_t(\theta)}}\frac{\partial m_t(\theta)}{\partial\theta}, \frac{1}{2h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}\right].$$

where l and s stands for location and scale, respectively.

Finally, we will use the notation $\langle X, Y\rangle = \mathrm{E}(XY')$ and $\|X\|^2 = \langle X, X\rangle = \mathrm{E}(XX')$.

2.3. Estimation methodologies

First, we consider the case for which the pdf g is fully known to the researcher and the object of interest is the estimation of θ. Maximum-likelihood estimation produces optimal estimators under a set of regularity conditions. MLE estimators are consistent and asymptotically efficient since they achieve the Cramér–Rao lower bound.

For a sample of length T, the averaged log-likelihood function is given by

$$\mathscr{L}_T(\theta) = -\frac{1}{2T}\sum_t \log h_t(\theta) + \frac{1}{T}\sum_t \log g(u_t(\theta)). \tag{2}$$

The ML estimator is found by maximizing Eq. (2) with respect to the vector of parameters θ. The score function is given by

$$\begin{aligned} S_T^{\mathrm{ml}}(\theta) &= \frac{\partial\mathscr{L}_T(\theta)}{\partial\theta} \\ &= -\frac{1}{T}\sum_t \left\{\frac{1}{\sqrt{h_t(\theta)}}\frac{\partial m_t(\theta)}{\partial\theta}\frac{g'(u_t(\theta))}{g(u_t(\theta))}\right. \\ &\qquad \left. + \frac{1}{2h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}\left(1 + u_t(\theta)\frac{g'(u_t(\theta))}{g(u_t(\theta))}\right)\right\} \\ &= \frac{1}{T}\sum_t W_t\psi_t, \end{aligned} \tag{3}$$

where ψ_t is used as short-hand notation for $\psi(u_t)$. The ML estimator $\hat{\theta}_{\mathrm{ml}}$ solves the system of equations $S_T^{\mathrm{ml}}(\theta) = 0$. Since this system is nonlinear in θ, the solution is obtained via numerical techniques. Note that the two factors in the score function, $h_t(\theta)^{-1/2}(\partial m_t(\theta)/\partial\theta)$ and $\frac{1}{2}h_t(\theta)^{-1}(\partial h_t(\theta)/\partial\theta)$, depend solely on past information, and they rely on the specification of the conditional mean and the conditional variance equations. The other two factors, g'/g and $(1 + u_t g'/g)$, are functions of u_t and depend on the shape of the pdf g. It is easy to show that the terms in (3) form a martingale difference sequence. The expectation of the score is zero for any pdf since integration by parts results in $\mathrm{E}(g'/g) = 0$ and $\mathrm{E}(u_t g'/g) = -1$.

Proving consistency and asymptotic normality of the ML estimator for (G)ARCH processes is a non-trivial exercise. Basawa et al. (1976) provide a set of sufficient conditions for consistency and asymptotic normality of estimators for dependent processes. Results are only available under the assumption of conditional normality and only for a limited class of processes, mainly GARCH(1,1) and ARCH(p) (see Weiss, 1986; Lumsdaine, 1996; Lee and Hansen, 1994). Under a correct specification of the variance equation and of the pdf g, the ergodic theorem and a central limit theorem can be invoked to show that

$$\sqrt{T}(\hat{\theta}_{\mathrm{ml}} - \theta_0) \rightarrow \mathrm{N}(0,\ V_{\mathrm{ml}}), \tag{4}$$

where $V_{\mathrm{ml}}^{-1} = B_0^{\mathrm{ml}}$ is the expectation of the outer product of the score evaluated at the true parameter vector θ_0 and is given by

$$B_0^{\mathrm{ml}} = \mathrm{E}\left[T\left(\frac{\partial \mathscr{L}_T(\theta_0)}{\partial\theta}\right)\left(\frac{\partial \mathscr{L}_T(\theta_0)}{\partial\theta}\right)' \right]. \tag{5}$$

Under the assumption of a correctly specified model, the information matrix equality holds; that is, $B_0^{\mathrm{ml}} = A_0^{\mathrm{ml}}$, where the matrix A_0^{ml} is (minus) the expectation of the Hessian matrix given by

$$A_0^{\mathrm{ml}} = -\mathrm{E}\left(\frac{\partial^2 \mathscr{L}_T(\theta_0)}{\partial\theta\partial\theta'}\right). \tag{6}$$

The second methodology that we consider for the estimation of the parameter vector θ is quasi-maximum-likelihood estimation. In this case, the researcher does not have any knowledge of the pdf that characterizes the standardized innovations u_t and chooses the normal pdf. The quasi-maximum-likelihood estimator is the argument that maximizes the likelihood function under the assumption of conditional normality, even though this may be a false assumption. The score function corresponding to a quasi-maximum-likelihood function is

$$S_T^{\mathrm{qml}}(\theta) = -\frac{1}{T}\sum_t \left\{ -\frac{1}{\sqrt{h_t(\theta)}} \frac{\partial m_t(\theta)}{\partial \theta} u_t(\theta) + \frac{1}{2h_t(\theta)} \frac{\partial h_t(\theta)}{\partial \theta} (1 - u_t(\theta)^2) \right\}$$

$$= \frac{1}{T}\sum_t W_t F_t, \tag{7}$$

with $F_t = F(u_t)$. The score function $S_T^{\mathrm{qml}}(\theta)$ preserves the martingale difference property. It is easy to see that the expectation of the score is equal to zero because u_t is a standardized innovation, for which $\mathrm{E}(u_t) = 0$ and $\mathrm{E}(u_t^2) = 1$. This property holds for any g and is the basis for proving consistency and asymptotic normality of the QML estimator under a set of regularity conditions. These conditions are discussed in Wooldridge (1994), Lee and Hansen (1994), Lumsdaine (1996), and Weiss (1986). The limiting distribution of the QML estimator is

$$\sqrt{T}(\hat{\theta}_{\mathrm{qml}} - \theta_0) \to \mathrm{N}(0, V_{\mathrm{qml}}), \tag{8}$$

where $V_{\mathrm{qml}} = A_0^{-1(\mathrm{qml})} B_0^{\mathrm{qml}} A_0^{-1(\mathrm{qml})}$ and where A^{qml} and B^{qml} are (minus) the expectation of the Hessian and the expectation of the outer product of the score respectively calculated under conditional normality. This estimator is less efficient than the ML estimator, reflecting the lack of information about the pdf. The finite-sample properties of QML and efficiency losses with respect to ML have been studied in several Monte Carlo simulations by Engle and González-Rivera (1991) and by Bollerslev and Wooldridge (1992). Newey and Steigerwald (1997) have shown that a quasi-maximum-likelihood approach with t-distributions is also feasible if an additional parameter is added. Although the derivations for this approach do not differ essentially from the ones in ordinary (normal) QML, we will not include the exact expressions for these estimators.

The third methodology that we consider is the semiparametric estimation of the parameter vector θ. In this situation, the researcher does not know the pdf of the standardized innovations but assumes that it is sufficiently smooth (Hájek and Šidák, 1967) to be approximated by a nonparametric density estimator. Semiparametric ARCH models were introduced by Engle and González-Rivera (1991), and their asymptotic properties were studied by Linton (1993), Steigerwald (1994), Drost and Klaassen (1997). The semiparametric estimator is a two-step estimator. In the first step, consistent estimates of the parameters of interest are obtained through, for example, quasi-maximum-likelihood estimation and are used to construct a nonparametric density of the standardized innovations. The second step consists of using this nonparametric density to adapt the initial estimator by a one-step Newton–Raphson improvement. The goal is to recapture the asymptotic efficiency losses due to quasi-maximum-likelihood estimation, which can be substantial when the departure of the true pdf from normality is large. On efficiency grounds, the semiparametric estimator is an intermediate estimator between the unattainable maximum-likelihood and the quasi-maximum-likelihood estimators. The semiparametric estimator is termed *adaptive* if it happens to have the same asymptotic efficiency as the maximum-likelihood estimator. The semiparametric efficiency bound depends on how informative the nuisance parameters, η, of the density are for the estimation of the parameters of interest θ. Let $S(\eta)$ be the population score vector for the nuisance parameters and $S(\theta)$ be the population score vector for the parameters of interest. The vector of parameters η is unknown and consequently the semiparametric estimator of θ cannot exploit the information contained in η. If η contains any information about θ, the efficient score for θ is found by calculating the residual vector $R(\theta)$ from the projection of $S(\theta)$ on the closure of the set of all linear combinations of $S(\eta)$, called the tangent set $\mathscr{T}$. The tangent set consists of linear combinations of $S(\eta)$ and, because the u's are random variables with mean zero and variance one, the elements of the tangent set are orthogonal to the function vector $(u_t, u_t^2 - 1)'$. Through the projection, all the variation of $S(\theta)$ due to $S(\eta)$ is removed (Newey, 1990). The residual vector $R(\theta)$ is the difference between $S(\theta)$ and the projection and, by construction, is orthogonal to this. Hence, $R(\theta)$ is the efficient score for θ and the semiparametric efficiency bound is

$$V_{\mathrm{sp}} = \left(\frac{1}{T}\sum_t \mathrm{E}[R_t(\theta_0)R_t(\theta_0)']\right)^{-1}. \tag{9}$$

We consider two cases: (i) the density g of the errors is completely unknown and (ii) the density g is known to be symmetric. In (ii) the tangent set contains only symmetric functions. In the rest of this article, expectations are implicitly taken under θ_0 and g.

In the most general case (i), for a sample of length T, the sample average vector residual is

$$\begin{aligned} R_T(\theta) &= \frac{1}{T}\sum_t R_t(\theta) \\ &= S_T^{\mathrm{ml}}(\theta) \end{aligned}$$

$$- \mathrm{E}(W_{1t}) \times \frac{1}{T}\sum_t \{[\psi_1(u_t) - \langle\psi_1, F\rangle \|F\|^{-2} F(u_t)]\}$$

$$- \mathrm{E}(W_{st}) \times \frac{1}{T}\sum_t \{[\psi_s(u_t) - \langle\psi_s, F\rangle \|F\|^{-2} F(u_t)]\}$$

$$= S_T^{\mathrm{ml}}(\theta)$$

$$- \mathrm{E}(W_t) \times \frac{1}{T}\sum_t \{[\psi(u_t) - \langle\psi, F\rangle \|F\|^{-2} F(u_t)]\}. \tag{10}$$

where $S_T^{\mathrm{ml}}(\theta)$ is given in (3). The derivation of Eq. (10) can be found along the lines in Bickel et al. (1993), see Drost et al. (1997) for the present time series set up. Note that it is quite easy to verify that $R_T(\theta)$ is indeed the efficient score. In the first place, $R_T(\theta)$ is orthogonal to the tangent set $\mathscr{T}$ and, secondly, the difference between $S_T^{\mathrm{ml}}(\theta)$ and $R_T(\theta)$ belongs to this tangent space.

In the case (ii) where the error densities are symmetric, the tangent space $\mathscr{T}$ consists of sums of all *symmetric* functions orthogonal to $F(u_t)$. In this case, the difference $S_T^{\mathrm{ml}}(\theta) - R_T(\theta)$ needs to be a symmetric function. We need to remove the non-symmetric residual of the projection of ψ_1 onto F. Hence, in the symmetric case the sample average vector residual is given by

$$R_T^{\mathrm{sym}}(\theta) = S_T^{\mathrm{ml}}(\theta)$$

$$- \mathrm{E}(W_t)\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \times \frac{1}{T}\sum_t \{[\psi(u_t) - \langle\psi, F\rangle \|F\|^{-2} F(u_t)]\}. \tag{11}$$

The conditions to show that $R_T^{\mathrm{sym}}(\theta)$ is the required efficient score are easily verified.

3. Efficiency comparisons among ML, QML, SP estimators

In this section we present the results concerning the most general time-varying location–scale model, and in the next section we particularize them for specific models. To simplify the exposition and because we work with stationary scores, we refer to one specific element of the score such that W_t becomes W, W_{1t} becomes W_1, W_{st} becomes W_s, and so on.

Observe that the expectations $\mathrm{E}(\psi F')$ and $\mathrm{E}(FF')$ can be explicitly calculated:

$$K = \langle\psi, F\rangle = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \qquad L = \|F\|^2 = \begin{pmatrix} 1 & \zeta \\ \zeta & \kappa - 1 \end{pmatrix},$$

where $\zeta = \mathrm{E}(u^3)$ and $\kappa = \mathrm{E}(u^4)$. Furthermore,

$$M = \|\psi\|^2$$

is the Fisher information in the location scale model. Note that $M - KL^{-1}K$ is positive semidefinite, since

$$M - KL^{-1}K = \|\psi - KL^{-1}F\|^2.$$

We introduce some additional notation before stating our main results. Define

$$\Pi = \|W\|^2, \quad \Lambda = \mathrm{E}(W)\mathrm{E}(W)', \quad \Sigma = \Pi - \Lambda.$$

Moreover, define for some arbitrary symmetric positive semidefinite matrix A,

$$\Pi_A = \|WA^{1/2}\|^2, \quad \Lambda_A = \mathrm{E}(W)A\mathrm{E}(W)', \quad \Sigma_A = \Pi_A - \Lambda_A.$$

By construction these matrices are positive semidefinite.

To facilitate the comparison of asymptotic variance–covariance matrices we work with the inverse of these matrices, V_{ml}^{-1}, V_{qml}^{-1}, and V_{sp}^{-1}. We focus in the absolute losses (gains). The relative losses are straightforward to derive from the absolute ones. Furthermore, the construction of the asymptotic variance–covariance matrices relies on certain regularity conditions such as those in Bollerslev and Wooldridge (1992). Essentially, these conditions require the satisfaction of uniform weak laws of large numbers and uniformly positive definiteness for (minus) the expectation of the hessian, as well as for the expectation of the outer product of the score.

Using the score function (3), and equations (4) and(5), for the maximum-likelihood estimator we can write

$$V_{\mathrm{ml}}^{-1} = \|W\psi\|^2 = \Pi_M. \tag{12}$$

For the quasi-maximum-likelihood estimator, using (7) and (8), we obtain

$$V_{\mathrm{qml}}^{-1} = \|\langle W\psi, WF\rangle \|WF\|^{-2} WF\|^2 = \Pi_K \Pi_L^{-1} \Pi_K. \tag{13}$$

For general error distributions, using (9) and(10), the semiparametric information bound can be written as

$$V_{\mathrm{sp}}^{-1} = \|W\psi - \mathrm{E}(W)[\psi - KL^{-1}F]\|^2 = \Sigma_{M-KL^{-1}K} + \Pi_{KL^{-1}K}. \tag{14}$$

The available information increases if the error distributions are known to be symmetric. The efficient score is given in (11). Thus, the semiparametric information bound is

$$V_{\mathrm{sp,sym}}^{-1} = \|W\psi - \mathrm{E}(W)[\psi - KL^{-1}F] + \mathrm{E}(W)\mathscr{I}[\psi - KL^{-1}F]\|^2$$
$$= \Sigma_{M-KL^{-1}K} + \Pi_{KL^{-1}K} + \Lambda_{\mathscr{I}[M-KL^{-1}K]\mathscr{I}}, \tag{15}$$

where $\mathscr{I}$ is the indicator matrix $\mathscr{I} = [1\ 0]'\,[1\ 0]$.

Comparisons of expressions (12)–(14), yield the following results.

Result 1. For a general error distribution, $V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1}$ is a positive-semidefinite matrix.

Proof. Comparing expressions (13) and (14),

$$V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1} = \Sigma_{M-KL^{-1}K} + \{\Pi_{KL^{-1}K} - \Pi_K \Pi_L^{-1} \Pi_K\}. \tag{16}$$

To show that the second term on the right-hand side is positive-semidefinite, observe that

$$\Pi_{KL^{-1}K} - \Pi_K \Pi_L^{-1} \Pi_K = \|WKL^{-1}F - \Pi_K \Pi_L^{-1} WF\|^2.$$

This completes the proof. □

Result 2. For a general error distribution, $V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1}$ is a positive-semidefinite matrix.

Proof. Comparing expressions (12) and (14),

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1} = \Lambda_{M-KL^{-1}K}. \quad \square \tag{17}$$

The following Result 3 can be obtained directly from Results 1 and 2. For transparency reasons, we also include a direct proof.

Result 3. For a general error distribution, $V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1}$ is a positive-semidefinite matrix.

Proof. Using the expressions for the asymptotic information matrices (12) and (13), we obtain by straightforward calculations

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1} = \Pi_{M-KL^{-1}K} + \{\Pi_{KL^{-1}K} - \Pi_K \Pi_L^{-1} \Pi_K\}. \tag{18}$$

Both terms on the right-hand side are positive semidefinite. □

If the error probability density is known to be symmetric, the efficient score in the semiparametric location–scale context is slightly different from the score for general error distributions. Results 1 and 2 have to be modified to reflect the additional information. We need to include an additional term $\Lambda_{\mathscr{I}[M-KL^{-1}K]\mathscr{I}}$. The details are left to the reader. Observe, however, that the symmetric case is not a subcase of Results 1 and 2. Knowing that the densities are symmetric yields essentially different scores.

The joint implication of Results 1, 2, and 3 is that, with the exception of the normal density, there is not any other probability density function for which the asymptotic variance–covariance matrices of the ML, QML, and SP estimators are equal. This is summarized in the following result.

Result 4. For the general time-varying location–scale model, and assuming that (i) $\mathrm{E}(W)$ has full rank, (ii) $\forall c \in \mathbb{R}^2$ $(W - \mathrm{E}(W))c \neq 0$, then

$$V_{\mathrm{ml}}^{-1} = V_{\mathrm{sp}}^{-1} = V_{\mathrm{qml}}^{-1} \tag{19}$$

if and only if the probability density function $g(.)$ is normal. Furthermore, equality between any two variance–covariance matrices implies the equality of the three matrices.

Proof. To prove sufficiency is straightforward. If the density is normal, then $K = L = M$, and (19) follows. To prove necessity, consider the following. If (17) is equal to zero, and (i) holds, then $M = KL^{-1}K$. If (16) is equal to zero and (ii) holds, then $M = KL^{-1}K$ and $\Pi_{KL^{-1}K} = \Pi_K \Pi_L^{-1} \Pi_K$. If (18) is equal to zero, then (16) and (17) are equal to zero. The equality $M = KL^{-1}K$ yields a pair of differential equations for which the only solution is the normal density. To find the solution to this system, proceed as in González-Rivera (1997). □

4. Specific models

In this section, we discuss four interesting submodels of the general time-varying location-scale model presented in Section 3. Depending upon the relation between the parameters in the mean and the parameters in the variance, we can have: (i) A pure time-varying location model. The parameters of interest are only location parameters. In this group, we may include dynamic specifications as ARMA models and general regression models. (ii) A pure time-varying scale model. There is no conditional mean and the parameters of interest are only the variance parameters. In this group we may include any type of conditional heteroscedasticity specification, such as the classic ARCH (Engle, 1982) and GARCH (Bollerslev, 1986), Exponential GARCH (Nelson, 1991), Nonlinear GARCH (González-Rivera, 1998), Asymmetric GARCH (Ding et al., 1993). (iii) Block-diagonal models, where the parameter vector can be split in two groups. One group contains the parameters in the conditional mean, and the other contains the parameters in the conditional variance, such that the conditional variance does not depend upon the parameters in the conditional mean and the conditional mean does not depend upon the parameters in the conditional variance. In this group, we may include regression models with heteroscedasticity other than GARCH, such as multiplicative heteroscedasticity, and ARMA-type models, where the conditional variance is a function of past observations. (iv) Block-triangular models, where the conditional variance depends upon the full set of mean and variance parameters but the conditional mean does not depend upon the conditional variance parameters. In this group, we may include regression models and ARMA-type models with conditional heteroscedasticity, but we require the ARCH or GARCH process to be symmetric as defined in Engle (1982), as well as symmetry of the pdf of the errors. For (iii) and (iv), we present the simplified expressions only for symmetric error distributions. In case of general error distributions, the resulting formulas are not essentially simpler than the general ones in Section 3.

4.1. Location models

In the class of pure location models, Results 1, 2, and 3 of Section 3 are greatly simplified. This is due to the fact that we do not have to differentiate the conditional variance with respect to the parameter of interest. Thus the matrix W consists of only one column, $W_t = [W_1]$. The matrices Π, Λ, and Σ contain exclusively location information. The function vector F simplifies to $F = u$ because the variance of the error distribution is not restricted to one in the location case. The matrices K, L, and M are reduced to numbers $k = 1$, $l = \mathrm{E}u^2 = \sigma^2$, and $m = \int (g'/g)^2 g$. This implies e.g. that expressions like Π_m may be written as $m\Pi$. Under a general error distribution, Results 1, 2, and 3 for the pure location model are

$$V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1} = [m - \sigma^{-2}]\Sigma,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1} = [m - \sigma^{-2}]\Lambda,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1} = [m - \sigma^{-2}]\Pi.$$

Under the assumption $\Lambda \neq 0$ and $\Sigma \neq 0$, the three estimators have the same asymptotic distribution if and only if $m = \sigma^{-2}$. This condition is only satisfied by the class of normal distributions. The case of $\Sigma = 0$ is rather exceptional. It only happens when the derivative of the conditional mean with respect to the mean parameters is nonrandom, i.e. the i.i.d. location model. In this instance, the semiparametric estimator is as efficient as the quasi-maximum-likelihood estimator. The equality $\Lambda = 0$ can happen in several models such as ARMA models without a constant term and regression models where the regressors have zero expectation. In these models, the semiparametric estimator is as efficient as the maximum likelihood estimator for all error distributions. This property is known as adaptivity of the mean parameters.

If the error distribution is restricted to the symmetric class of densities, we always have adaptivity of the location parameters, independently of any restriction on the matrix Λ. Results 1, 2, and 3 for symmetric densities are

$$V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1} = [m - \sigma^{-2}]\Pi,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1} = 0,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1} = [m - \sigma^{-2}]\Pi.$$

4.2. Scale models

Results 1, 2, and 3 for the pure time-varying scale models are also greatly simplified. The matrix W consists of only one column, $W_t = [W_{\mathrm{s}}]$. In contrast to the location model, the vector function F remains unchanged because the error distribution has two moment restrictions. The matrices K and M reduce to the row-vector $k = (0\ 2)$ and to the real number $m = \int (1 + ug'/g)^2 g$, respectively. The matrix L is not affected. For the quasi-maximum-likelihood estimator, we find that $A_0^{\mathrm{qml}} = [2/(\kappa - 1)]B_0^{\mathrm{qmle}}$ and that its asymptotic variance–covariance matrix simplifies to $V_{\mathrm{qml}}^{-1} = [4/(\kappa - 1)]\Pi$. For general error distributions, Results 1, 2, and 3 for the pure time-varying scale model are

$$V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1} = \left[m - \frac{4}{\kappa - 1 - \zeta^2}\right]\Sigma + \frac{4\zeta^2}{(\kappa - 1)(\kappa - 1 - \zeta^2)}\Pi,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1} = \left[m - \frac{4}{\kappa - 1 - \zeta^2}\right]\Lambda,$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1} = \left[m - \frac{4}{\kappa - 1}\right]\Pi.$$

Similar to the pure location model, $\Sigma = 0$ is only possible in the i.i.d. scale model. In the pure scale models, $\Lambda \neq 0$ happens in all practical econometric situations. Under the assumption $\Sigma \neq 0$ and $\Lambda \neq 0$, adaptivity is not possible in the scale model with general error distributions. However, $V_{\mathrm{ml}}^{-1} = V_{\mathrm{sp}}^{-1}$ if and only if the following condition holds $m = 4/(\kappa - 1 - \zeta^2)$. González-Rivera (1997) has shown that this condition is satisfied by a class of symmetrized ($\zeta = 0$) square root chi-squared distributions (among which the normal density is a special case) and for a class of nonsymmetric distributions with $\zeta \neq 0$.

If $\zeta = 0$, $V_{\mathrm{ml}}^{-1} = V_{\mathrm{sp}}^{-1} = V_{\mathrm{qml}}^{-1}$ for the set of distributions described in González-Rivera (1997). If $\zeta \neq 0$, the semiparametric estimator is always more efficient than the quasi-maximum-likelihood estimator. In other words, there is no density for which the asymptotic distributions of the three estimators are the same.

Note that adding a symmetry condition to the set of error distributions does not increase the information of the semiparametric estimator in the pure scale model. Consequently, the previous conclusions remain for the symmetric case.

4.3. Block-diagonal models

In these models, the parameter vector is partitioned in two subsets, mean and variance parameters, such that the conditional mean does not depend on the variance parameters and the conditional variance does not depend on the mean parameters. The matrix W is partitioned as follows:

$$W = \begin{bmatrix} W_{\mathrm{l}}^{\mathrm{m}} & 0 \\ 0 & W_{\mathrm{s}}^{\mathrm{v}} \end{bmatrix},$$

where the superindexes 'm' and 'v' account for 'mean' and 'variance', respectively. We restrict our attention to symmetric error distributions. Denote the diagonal elements of the diagonal matrix M by m_{l} and m_{s} (the off-diagonal elements are zero because of symmetry) and let the block-diagonal matrix Π have an upper-left block Π_{l} and a lower-right block Π_{s}. Results 1, 2, and 3 for block-diagonal models with symmetric error distributions are

$$V_{\mathrm{sp}}^{-1} - V_{\mathrm{qml}}^{-1} = \begin{pmatrix} [m_{\mathrm{l}} - 1]\Pi_{\mathrm{l}} & 0 \\ 0 & [m_{\mathrm{s}} - \frac{4}{\kappa - 1}]\Sigma_{\mathrm{s}} \end{pmatrix},$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{sp}}^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & [m_{\mathrm{s}} - \frac{4}{\kappa - 1}]\Lambda_{\mathrm{s}} \end{pmatrix},$$

$$V_{\mathrm{ml}}^{-1} - V_{\mathrm{qml}}^{-1} = \begin{pmatrix} [m_{\mathrm{l}} - 1]\Pi_{\mathrm{l}} & 0 \\ 0 & [m_{\mathrm{s}} - \frac{4}{\kappa - 1}]\Pi_{\mathrm{s}} \end{pmatrix}.$$

Note that the location parameters can be adaptively estimated, just as in the pure location problem with symmetric error distributions. Adaptivity of the scale parameters will not hold. The asymptotic efficiency of the three estimators will be identical, i.e. $V_{\mathrm{ml}}^{-1} = V_{\mathrm{sp}}^{-1} = V_{\mathrm{qml}}^{-1}$, if and only if $m_{\mathrm{l}} = 1$ and $m_{\mathrm{s}} = 4/(\kappa - 1)$. These two conditions are *jointly* satisfied only by the normal distribution.

4.4. Block-triangular models

In these models, the conditional variance depends on the parameters of the conditional mean but the conditional mean does not depend on the parameters of the conditional variance. We restrict our attention to symmetric densities. The matrix W is of the following form:

$$W = \begin{bmatrix} W_{\mathrm{l}}^{\mathrm{m}} & W_{\mathrm{s}}^{\mathrm{m}} \\ 0 & W_{\mathrm{s}}^{\mathrm{v}} \end{bmatrix}.$$

Furthermore, if the conditional variance is a symmetric process in the innovations of the mean model, it can be shown that $\mathrm{E}(W_{\mathrm{s}}^{\mathrm{m}} W_{\mathrm{s}}^{\mathrm{v}\prime}) = 0$ (Theorem 4 in Engle, 1982). Examples where this orthogonality condition is satisfied are the classical symmetric ARCH and GARCH models. In these cases, the matrices Π_K and Π_L are block-diagonal, but the equality $\Pi_{KL^{-1}K} - \Pi_K \Pi_L^{-1} \Pi_K = 0$ does not hold anymore. Nevertheless, the lower-right block of this matrix is zero, and only for the *scale* parameters do we obtain the same comparisons as those of the block-diagonal models, i.e.

$$V_{\mathrm{sp}_{(s)}}^{-1} - V_{\mathrm{qml}_{(s)}}^{-1} = \left[m_{\mathrm{s}} - \frac{4}{\kappa - 1} \right] \Sigma_{\mathrm{s}},$$

$$V_{\mathrm{ml}_{(s)}}^{-1} - V_{\mathrm{sp}_{(s)}}^{-1} = \left[m_{\mathrm{s}} - \frac{4}{\kappa - 1} \right] \Lambda_{\mathrm{s}},$$

$$V_{\mathrm{ml}_{(s)}}^{-1} - V_{\mathrm{qml}_{(s)}}^{-1} = \left[m_{\mathrm{s}} - \frac{4}{\kappa - 1} \right] \Pi_{\mathrm{s}}.$$

4.5. Numerical efficiency losses

For the particular models of the previous sections we calculate the relative efficiency loss of the QML estimator with respect to the ML estimator.

We consider a set of standardized probability density functions for which we compute the coefficients of skewness and kurtosis, and the Fisher information of location and scale. For a standardized Student-t with v degrees of freedom, we have that $\zeta = 0$, $\kappa = 3(v-2)/(v-4)$, $m_{\mathrm{l}} = v(v+1)/((v-2)(v+3))$, and $m_{\mathrm{s}} = 2v/(v+3)$. For a standardized Chi-square with v degrees of freedom, we

Table 1

Standardized density	Shape characteristics				Efficiency loss $(V_{\mathrm{qml}} \cdot V_{\mathrm{ml}}^{-1} - 1)\%$	
	m_l	m_s	ζ	κ	Mean par.	Variance par.
Normal	1	2	0	3	0	0
Student-t						
$v = 5$	1.25	1.25	0	9.00	25	150
$v = 8$	1.09	1.45	0	4.50	9	27
$v = 12$	1.04	1.60	0	3.75	4	10
Laplace	2	1	0	6	100	25
Chi-Square						
$v = 10$	1.67	3.33	0.89	4.20	67	167
$v = 15$	1.36	2.73	0.73	3.80	36	91
$v = 20$	1.25	2.50	0.63	3.60	25	63
$v = 30$	1.15	2.31	0.52	3.40	15	38

have that $\zeta = 2\sqrt{2/v}$, $\kappa = 3(v+4)/v$, $m_l = v/(v-4)$, and $m_s = 2v/(v-4)$. With these expressions, the efficiency loss of the QML estimator with respect to the ML estimator is straightforward to compute. In Table 1 we show some examples where the efficiency loss is quantified for the above mentioned probability density functions.

With the exception of the Laplace distribution, it can be seen that the relative efficiency loss is larger for the variance parameters than for the mean parameters. Consequently, the implementation of a semiparametric estimator has a larger pay-off in those instances in which the variance parameters are the parameters of interest.

5. Conclusions

In this paper we have quantified the asymptotic efficiency losses (gains) of the ML, QML, and SP estimators in the context of GARCH models. We have obtained a set of results for a general time-varying location–scale model. The factors that contribute to differences in efficiency among the estimators can be divided in two categories. One pertains to the parametric specifications of the conditional mean and the conditional variance. The other corresponds to the shape characteristics of the conditional density of the standardized errors, summarized in the coefficient of skewness and the coefficient of kurtosis together with the Fisher information for location and scale. We have proven that there is no probability density function, with the exception of the normal, for which the asymptotic efficiency of the three estimators is the same.

Out of the general location–scale model, we have extracted four particular models. In a pure time-varying location model, the coefficients of skewness and kurtosis, and the Fisher information for scale do not play any role in explaining efficiency differences. In a pure time-varying scale model, there is no need for the Fisher information for location, but the coefficient of skewness is important in explaining differences between the SP and QML estimators, and between the SP and the ML estimators. Surprisingly, however, skewness is irrelevant in determining efficiency differences between the ML and QML estimators. In the pure scale models with skewness equal to zero, the three estimators can have equal asymptotic efficiency for other densities than the normal. Apart from pure location and pure scale models, we have considered two more cases, block-diagonal models and block-triangular models, with symmetric density functions. In the block-diagonal models, the asymptotic variance–covariance matrices are block diagonal between the mean and variance parameters. Essentially, in these models, the efficiency comparisons reduce to those of the pure location and pure scale models together. In the block-triangular models, the asymptotic variance–covariance matrices are still block diagonal between mean and variance parameters, but it is only for the scale parameters where the efficiency comparisons reduce to those of the pure scale models.

These results have practical implications for the empirical researcher. A potential strategy may be to start the estimation process with a QML methodology. To recapture the efficiency losses of the QML estimator, we need to evaluate the matrix M, the coefficient of skewness and the coefficient of kurtosis of the standardized residuals. The matrix M can be estimated by nonparametric methods. The matrices Π and Λ are already estimated in the QML estimation. Straightforward application of Results 1, 2, and 3 provides the efficiency loss. This implies that even with the most inefficient estimator such as QML, the researcher can estimate the maximal efficiency bound provided by the unattainable ML estimator.

Acknowledgements

We wish to thank to the participants of the 1995 Joint Meeting of the American Statistical Association in Orlando, FL, the 1997 Econometric Society Meetings, New Orleans, LA, and seminars at UCLA and UCR. We are specially thankful to Jeffrey Wooldridge, the editor Richard Blundell, the associate editor, and two anonymous referees, for their constructive comments. The first author gratefully acknowledges the financial support of the Academic Senate of the University of California, Riverside. The usual caveat applies.

References

Basawa, I.V., Feigin, P.D., Heyde, C.C., 1976. Asymptotic properties of maximum-likelihood estimators for stochastic processes. Sankhya: The Indian Journal of Statistics 38, 259–270.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Statistical Inference for Semiparametric Models. Johns Hopkins University Press, Baltimore, MD.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Wooldridge, J.M., 1992. Quasi-maximum-likelihood estimation and inference in dynamic models with time varying covariances. Econometric Reviews 11, 143–72.

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

Drost, F.C., Klaassen, C.A.J., 1997. Efficient estimation in semiparametric GARCH models. Journal of Econometrics 81, 193–221.

Drost, F.C., Klaassen, C.A.J., Werker, B.J.M., 1997. Adaptive estimation in time series models. The Annals of Statistics 25, 786–817.

Engle, R.F., González-Rivera, G., 1991. Semiparametric ARCH models. Journal of Business and Economic Statistics 9, 345–360.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica 50, 987–1008.

González-Rivera, G., 1997. A note on adaptation in GARCH models. Econometric Reviews 16, 55–68.

González-Rivera, G., 1998. Smooth transition GARCH models. Studies in Nonlinear Dynamics and Econometrics 3(2), 61–78.

Hájek, J., Šidák, Z., 1967. Theory of Rank Tests. Academia, Prague.

Koul, H.L., Shick, A., 1997. Efficient estimation in nonlinear autoregressive time series models. Bernoulli 3, 247–277.

Lee, S.W., Hansen, B.E., 1994. Asymptotic theory for the GARCH(1, 1) quasi-maximum-likelihood estimator. Econometric Theory 10, 29–52.

Linton, O., 1993. Adaptive estimation in ARCH models. Econometric Theory 9, 539–569.

Lumsdaine, R.L., 1996. Consistency and asymptotic normality of the quasi-maximum-likelihood estimator in IGARCH(1, 1) and covariance stationary GARCH(1, 1) models. Econometrica 64, 575–596.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: a new approach. Econometrica 59, 307–346.

Newey, W.K., 1990. Semiparametric efficiency bounds. Journal of Applied Econometrics 5, 99–135.

Newey, W.K., Steigerwald, D.G., 1997. Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. Econometrica 65, 587–599.

Steigerwald, D.G., 1994. Efficient estimation of models with conditional heteroscedasticity. Working Paper. Department of Economics. University of California, Santa Barbara.

Weiss, A.A., 1986. Asymptotic theory for ARCH models: Estimation and testing. Econometric Theory 2, 107–131.

Wooldridge, J.M., 1994. Estimation and inference for dependent processes. Handbook of Econometrics, vol. 4.