

Non-Gaussian Quasi Maximum Likelihood Estimation of GARCH Models

Lei Qi*, Dacheng Xiu and Jianqing Fan

June 15, 2010

Abstract

The non-Gaussian quasi maximum likelihood estimator is frequently used in GARCH models with intension to improve the efficiency of the GARCH parameters. However, the method is usually inconsistent unless the quasi-likelihood happens to be the true one. We identify an unknown scale parameter that is critical to the consistent estimation of non-Gaussian QMLE. As a part of estimating this unknown parameter, a two-step non-Gaussian QMLE (2SNG-QMLE) is proposed for estimation the GARCH parameters. Without assumptions on symmetry and unimodality of the distributions of innovations, we show that the non-Gaussian QMLE remains consistent and asymptotically normal, under a general framework of non-Gaussian QMLE. Moreover, it has

*Lei Qi and Dacheng Xiu are PhD candidates, Jianqing Fan is Frederick L. Moore18 Professor of Finance at Bendheim Center for Finance, Princeton University, Princeton, NJ 08544. (Email: lqi@princeton.edu, dachengx@princeton.edu, jqfan@princeton.edu) This research was supported in part by NSF Grant DMS-0704337 and DMS-0714554.

higher efficiency than the Gaussian QMLE, particularly when the innovation error has heavy tails. Two extensions are proposed to further improve the efficiency of 2SNG-QMLE. The impact of relative heaviness of tails of the innovation and quasi-likelihood distributions on the asymptotic efficiency has been thoroughly investigated. Monte Carlo simulations and an empirical study confirm the advantages of the proposed approach.

1 Introduction

Volatility has been a crucial ingredient in modeling financial time series and designing risk management and trading strategies. It is often observed that volatilities tend to cluster together. This characteristic of financial data suggests that volatilities are autocorrelated and changing over time. Engle (1982) proposed ARCH (autoregressive conditional heteroscedasticity) to model volatility dynamics by taking weighted averages of past squared forecast errors. This seminal idea led to a great richness and variety of volatility models. Among numerous generalizations and developments, GARCH model by Bollerslev (1986) has been commonly used:

$$x_t = v_t \varepsilon_t \tag{1}$$

$$v_t^2 = c + \sum_{i=1}^p \tilde{a}_i x_{t-i}^2 + \sum_{j=1}^q \tilde{b}_j v_{t-j}^2 \tag{2}$$

In this GARCH(p, q) model, variance forecast takes weighted average of not only past square errors but also historical variances. The simplicity and intuitive appeal make GARCH model, especially GARCH(1, 1), a workhorse and good start point in

many financial applications.

Earlier literature on inference from ARCH/GARCH models is based on the maximum likelihood estimation (MLE) with conditional Gaussian assumption on the innovations. Plenty of empirical evidence, however, has documented heavy-tailed and asymmetric innovation distributions of ε_t , rendering this assumption unjustified, see for instance Diebold (1988). Consequently, MLE using Student's t or generalized Gaussian likelihood functions has been introduced, see e.g. Engle and Bollerslev (1986), Bollerslev (1987), Hsieh (1989), and Nelson (1991). However, these methods may lead to inconsistent estimates if the distribution of the innovation is misspecified. Alternatively, the Gaussian MLE, regarded as a quasi maximum likelihood estimator (QMLE) may be consistent, see e.g. Elie and Jeantheau (1995), and asymptotically normal, provided that the innovation has a finite fourth moment, even if it is far from Gaussian, see Hall and Yao (2003) and Berkes et al. (2003). The asymptotic theory dates back to as early as Weiss (1986) for ARCH models, Lee and Hansen (1994) and Lumsdaine (1996) for GARCH(1, 1) with stronger conditions, and Bollerslev and Wooldbridge (1992) for GARCH(p, q) under high level assumptions.

Nevertheless, gain in robustness comes with efficiency loss. Theoretically, the divergence of Gaussian likelihood from the true innovation density may considerably increase the variance of the estimates, which thereby fails to reach the Cramér-Rao bound by a wide margin, reflecting the cost of not knowing the true innovation distribution. Engle and Gonzalez-Rivera (1991) has suggested a semiparametric procedure that can improve the efficiency of the parameter estimates up to 50% over

the QMLE based on their Monte Carlo simulations, but still incapable of capturing the total potential gain in efficiency, see also Linton (1993). Drost and Klaassen (1997) has put forward an adaptive two-step semiparametric procedure based on a re-parametrization of the GARCH(1, 1) model with unknown but symmetric error. Gonzalez-Rivera and Drost (1999) has compared its efficiency gain/loss over Gaussian QMLE and MLE. All the effort would become void if the innovation fails to have a finite fourth moment. Hall and Yao (2003) has considered the Gaussian QMLE and shown that it would converge to stable distributions asymptotically rather than a normal distribution.

The empirical reason of Gaussian QMLE's efficiency loss is that financial data are generally heavy tail distributed. The conditional normality assumption is violated. For example, Bollerslev and Wooldbridge (1992) reported that sample kurtosis of estimated residuals of Gaussian QMLE on S&P500 monthly data is 4.6, well exceeding the Gaussian kurtosis which is 3. It is therefore intuitively appealing to develop QMLE based on non-Gaussian likelihoods, especially heavy tailed likelihoods. And the efficiency loss of Gaussian QMLE can be greatly reduced by replacing the likelihoods with heavy tailed ones.

In contrast with the majority of literature focusing on Gaussian QMLE for inference, there is rather limited attention on inference using non-Gaussian QMLE. This may be partly due to the fact that the Gaussian QMLE is robust against misspecification of error distribution, while directly using non-Gaussian QMLE is not. In general a non-Gaussian QMLE does not yield consistent estimation when true error distribution deviates from the likelihood. Moreover, this inconsistency could not be

corrected even as we allow to estimate a shape parameter indexing the non-Gaussian likelihood family together with model parameters unless the true innovation density is a member of this likelihood family. Otherwise, estimating shape along with model parameters simply picks one likelihood that is “least” biased, however the bias persists. Newey and Steigerwald (1997) have considered the identification of the non-Gaussian QMLE for heteroscedastic parameters in general conditional heteroscedastic models. They have also pointed out that the scale parameter may not be identified as its true value since it is no longer a natural scale parameter for non-Gaussian densities.

A valid remedy served for non-Gaussian QMLE would be manipulating model assumptions in order to maintain consistent estimation. For example, the true innovation density is sometimes taken to be Student’s t or generalized Gaussian for granted. Alternatively, Berkes and Horváth (2004) has shown that with a different moment condition on the true innovations instead of the original $E(\varepsilon^2) = 1$, a corresponding non-Gaussian QMLE would obtain consistency and asymptotic normality. However, this moment condition $E(\varepsilon^2) = 1$ is an essential assumption which enables v_t to bear the natural interpretation of the conditional standard deviation, the notion of volatility. More importantly, moment condition is part of model specification, and it should be prior to and independent of the choice of likelihood. Changing the moment condition would not solve the robustness issue of non-Gaussian QMLE; it simply renders consistency to the correct combination of moment condition and non-Gaussian likelihood, which cannot be determined without knowing the true innovation.

Therefore, we prefer a non-Gaussian QMLE method which is robust against error misspecification, more efficient than Gaussian QMLE, independent of model assumptions, and yet practical. Such method can also well extend the usage of non-Gaussian QMLE in GARCH software packages. Current packages do include choice of likelihood as an option, for example, Student's t and generalized Gaussian. In addition the shape parameter can be specified or estimated. But as discussed before, such method is not robust against error misspecification. When running estimation, one chooses a particular likelihood family with the hope that true innovation distribution falls into such family, but typically it does not.

The main contribution of this paper is that we propose a novel two step non-Gaussian QMLE method, 2SNG-QMLE for short, which meets the desired properties. The key is the estimation of a scale adjustment parameter, denoted as η_f , for non-Gaussian likelihood to ensure the consistency of non-Gaussian QMLE under any error distributions. η_f is estimated through Gaussian QMLE in the first step; then we feed the estimated η_f into non-Gaussian QMLE in the second step. In Gaussian QMLE η_f is held constant at unity, and partly because of that this quantity has been overlooked; but in non-Gaussian QMLE η_f is no longer constant, and how much it deviates from unity measures how much asymptotic bias would incur by simply using non-Gaussian QMLE without such adjustment.

The second contribution is that we adopt a re-parameterized GARCH model (see also Newey and Steigerwald (1997) and Drost et al. (1997)) which separates the volatility scale parameter from heteroscedastic parameters. Under this new parametrization, we derive asymptotic behaviors for 2SNG-QMLE. The results show

that 2SNG-QMLE is more efficient than Gaussian-QMLE under various innovation settings, and furthermore there is a clear cut on asymptotic behaviors under the new parametrization. The heteroscedastic parameters we can always achieve $T^{\frac{1}{2}}$ asymptotic normality, whereas Gaussian QMLE has slower convergence rate when error does not have fourth moment.

The outline of the paper is as follows. Section 2 introduces the model and assumptions. Section 3 discusses the estimation procedure and derives the asymptotic results for 2SNG-QMLE. Section 4 proposes two extensions to further improve efficiency. Section 5 employs Monte Carlo simulations to verify the theoretic results. Section 6 conducts real data analysis on stock returns. Section 7 concludes. The appendix provides all the mathematical proofs.

2 The Model and Assumptions

The re-parameterized GARCH(p, q) model takes on the following parametric form:

$$x_t = \sigma v_t \varepsilon_t \tag{3}$$

$$v_t^2 = 1 + \sum_{i=1}^p a_i x_{t-i}^2 + \sum_{j=1}^q b_j v_{t-j}^2 \tag{4}$$

The model parameters are summarized in $\boldsymbol{\theta} = \{\sigma, \boldsymbol{\gamma}'\}'$, where σ is the scale parameter and $\boldsymbol{\gamma} = (\boldsymbol{a}', \boldsymbol{b}')$ is the autoregression parameter. The true parameter $\boldsymbol{\theta}_0$ is in the interior of Θ , which is a compact subset of the \boldsymbol{R}_+^{1+p+q} , satisfying $\sigma > 0$, $a_i \geq 0$, $b_j \geq 0$. We use subscript 0 to denote the value under the true model throughout the paper. The innovation $\{\varepsilon_t\}_{t=1, \dots, T}$ are i.i.d random variables with mean 0, variance 1

and unknown density $g(\cdot)$. In addition, we assume that the GARCH process $\{x_t\}$ is strictly stationary and ergodic. The elementary conditions for the stationarity and ergodicity of GARCH models have been discussed in Bougerol and Picard (1992).

We consider a parametric family of quasi likelihood $\{\eta : \frac{1}{\eta}f(\frac{\cdot}{\eta})\}$ indexed by $\eta > 0$, for any given likelihood function f . Unlike a shape parameter that is often included in a Student's t likelihood function, η is a scale parameter selected to reflect the penalty of model misspecification. More precisely, a specific quasi likelihood scaled by η_f will be used in the estimation procedure. The parameter η_f minimizes the discrepancy between the true innovation density g and an unscaled misspecified quasi likelihood in the sense of Kullback Leibler Information Distance, see e.g. White(1982). Or equivalently,

$$\eta_f = \operatorname{argmax}_{\eta > 0} E\left\{ -\log \eta + \log f\left(\frac{\varepsilon}{\eta}\right) \right\} \quad (5)$$

where the expectation is taken under the true model g .

Note that η_f here only depends on the divergence of the two densities under consideration, rendering it a universal measure of closeness irrelevant of the GARCH model. Once η_f is given, the QMLE $\hat{\boldsymbol{\theta}}$ is defined by maximizing the following quasi likelihood with this model parameter η_f :

$$L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T l_t(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T \left(-\log(\sigma v_t) + \log f\left(\frac{x_t}{\eta_f \sigma v_t}\right) \right) \quad (6)$$

Apparently, the likelihood function differs from a regular one with the additional model parameter η_f . In fact, our approach is a generalization of the Gaussian QMLE

and the MLE as illustrated in the next proposition.

Proposition 1. *If $f \propto \exp(-x^2/2)$ or $f = g$, then $\eta_f = 1$.*

Moreover, it can be implied from Newey and Steigerwald (1997) that in general, an unscaled non-Gaussian likelihood function applied in this new re-parametrization of GARCH(p, q) setting fails to identify the volatility scale parameter σ , resulting in inconsistent estimates. We show in the next section that incorporating η_f into the likelihood function facilitates the identification of the volatility scale parameter.

For convenience, we assume the following regularity conditions are always satisfied: f is twice continuously differentiable, and for any $\eta > 0$, we have $\sup_{\theta \in \Theta} E|l_t(\theta)| < \infty$, $E \sup_{\theta \in \mathcal{N}} |\nabla l_t(\theta)| < \infty$, and $E \sup_{\theta \in \mathcal{N}} |\nabla^2 l_t(\theta)| < \infty$, for some neighborhood \mathcal{N} of θ_0 .

3 Main Results

3.1 Identification

Identification is a critical condition for consistency. It requires that the expected quasi likelihood $\bar{L}_T(\theta) = E(L_T(\theta))$ has a unique maximum at the true parameter value θ_0 . To show that θ can be identified in the presence of η_f , we make the following assumptions:

Assumption 1. *A quasi likelihood of the GARCH (p, q) model is selected such that*

1. $v_t(\gamma_0) > 0$, and $v_t(\gamma)/v_t(\gamma_0)$ is not a constant if $\gamma \neq \gamma_0$.

2. The function $Q(\eta) = -\log \eta + E(\log f(\frac{\varepsilon}{\eta}))$ has a unique maximizer $\eta_f > 0$.

Note that the first point is the usual identification condition for the autoregression parameter γ_0 , and that the second requirement is the key to the identification of σ_0 .

Lemma 1. *Given Assumption 1, $\bar{L}_T(\theta)$ has a unique maximum at the true value $\theta = \theta_0$.*

The next lemma provides a few primitive sufficient conditions for the last statement of Assumption 1. The conditions given below provide a general guideline of choosing an appropriate quasi likelihood f .

Lemma 2. *Assume that f is continuously differentiable up to the second order and $h(x) = x \frac{\dot{f}(x)}{f(x)}$. Suppose that $\{\varepsilon_t\} \sim \varepsilon$ is i.i.d. with mean 0, variance 1 and a finite p^{th} moment. If, in addition,*

1. $h(x) \leq 0$.
2. $x\dot{h}(x) \leq 0$ and the equality holds if and only if $x = 0$.
3. $|h(x)| \leq C|x|^p$, and $|x\dot{h}(x)| \leq C|x|^p$, for some constant $C > 0$, and $p \geq 0$.
4. $\limsup_{x \rightarrow \infty} h(x) < -1$

then $Q(\eta)$ has a unique maximum at some point $\eta_f > 0$. Furthermore, $\eta_f > 1$ if and only if $Eh(\varepsilon) < -1$.

The last three assumptions are more general than the concavity of the function $Q(\eta)$. For some commonly used likelihood such as the family of Student's t likelihood, the concavity assumption of Q is violated. However, it still satisfies the above lemma. A few examples of families of likelihood that satisfy Lemma 2 are given below.

Remark 1. If $f = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, then $h(x) = -x^2$. If f is the standardized t_ν -distribution with $\nu > 2$, that is $f \propto (1 + \frac{x^2}{\nu-2})^{-\frac{\nu+1}{2}}$, then $h(x) = -\frac{(\nu+1)x^2}{\nu-2+x^2}$. Both cases satisfy Lemma 2 with $p = 2$. In addition, if $\log f(x) = -|x|^\beta (\frac{\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})})^{\frac{\beta}{2}} + \text{const}$, the generalized Gaussian likelihood, then $h(x) = -\beta (\frac{\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})})^{\frac{\beta}{2}} |x|^\beta$. In this case, by choosing $p = \beta$, Lemma 2 is satisfied.

3.2 The Distinction Between Gaussian and Non-Gaussian QMLE

First of all, consider the case in which η_f is given, or more directly, the true error distribution is known. The following asymptotic analysis reveals the difference between the Gaussian QMLE and the non-Gaussian one.

Theorem 1. Assume that η_f is known. Under Assumptions 1, $\hat{\theta}_T \xrightarrow{\mathcal{P}} \theta_0$, where $\hat{\theta}_T$ is the quasi likelihood estimator obtained by maximizing (6).

Next, we discuss the asymptotic normality of the QMLE. As usual, additional moment conditions are needed.

Assumption 2. Let $\mathbf{k} = (\frac{1}{\sigma}, \frac{1}{v_t} \frac{\partial v_t}{\partial \gamma})'$, and \mathbf{k}_0 be its value at $\theta = \theta_0$.

1. $0 < E(h^2(\frac{\varepsilon}{\eta_f})) < \infty$, $0 < E|\varepsilon \dot{h}(\frac{\varepsilon}{\eta_f})| < \infty$.

2. $\mathbf{M} = E(\mathbf{k}_0 \mathbf{k}_0') < \infty$.

Theorem 2. Under Assumptions 1 and 2, we have

$$T^{\frac{1}{2}}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \Sigma_1) \quad (7)$$

where $\Sigma_1 = \mathbf{M}^{-1} \frac{Eh_1^2}{(Eh_2)^2}$, $h_1(\varepsilon) = 1 + h(\frac{\varepsilon}{\eta_f})$, and $h_2(\varepsilon) = \frac{\varepsilon}{\eta_f} \dot{h}(\frac{\varepsilon}{\eta_f})$.

The moment conditions given by the first point of Assumption 2 only depend on the tail of the innovation density g and quasi likelihood f . A striking advantage of non Gaussian QMLE over its Gaussian alternative is that the former may require weaker conditions on the tail of the innovation. It is well known that the asymptotic normality of Gaussian likelihood requires a finite fourth moment. By contrast, it implies from Remark 1 that any Student's t likelihood with degree of freedom larger than 2 has a bounded moment, so that no additional moment conditions are needed other than those assumed in any GARCH model.

Moreover, it turns out that model parameter η_f has another interpretation as a bias correction for a simple non-Gaussian QMLE of the scale parameter in that $\sigma_0\eta_f$ would be reported instead of σ_0 . Therefore, the unscaled QMLE can consistently estimate σ_0 if and only if $\eta_f = 1$. Proposition 1 hence reveals the distinction in consistency between the MLE, Gaussian QMLE and the other alternatives.

In general, for an arbitrary likelihood, η_f would not equal to 1, thereby creating the popularity of the Gaussian QMLE, whose η_f is exactly 1. It is therefore necessary to incorporate this bias-correction factor η_f into non-Gaussian QMLE, which may potentially obtain a better efficiency than the Gaussian QMLE. However, as we have no prior information concerning the true innovation density, η_f is unknown. As a result, this estimator is infeasible. A promising way to resolve this issue would be to estimate η_f in the first step.

3.3 Two-Step Estimation Procedure

In order to estimate η_f , a sample on the true innovation is required. According to Proposition 1, without knowing η_f , the residuals from the Gaussian QMLE may potentially provide substitution for the true innovation sample. A two-step estimation procedure is proposed in the following. In the first step, $\hat{\eta}_f$ is obtained by maximizing (5) with estimated residuals from Gaussian quasi likelihood estimation:

$$\hat{\eta}_f = \operatorname{argmax}_{\eta} \frac{1}{T} \sum_{t=1}^T l_2(x_t, \tilde{\boldsymbol{\theta}}_T, \eta) = \operatorname{argmax}_{\eta} \frac{1}{T} \sum_{t=1}^T \left(-\log(\eta) + \log f\left(\frac{\tilde{\varepsilon}_t}{\eta}\right) \right) \quad (8)$$

where

$$\tilde{\boldsymbol{\theta}}_T = \operatorname{argmax}_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T l_1(x_t, \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T \left(-\log(\sigma v_t) - \frac{x_t^2}{2\sigma^2 v_t^2} \right) \quad (9)$$

and $\tilde{\varepsilon}_t = x_t/(\tilde{\sigma}v_t(\tilde{\gamma}))$. Next, we maximize non-Gaussian quasi likelihood with plug-in $\hat{\eta}_f$ and obtain $\hat{\boldsymbol{\theta}}_T$:

$$\hat{\boldsymbol{\theta}}_T = \operatorname{argmax}_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T l_3(x_t, \hat{\eta}_f, \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T \left(-\log(\sigma v_t) + \log f\left(\frac{x_t}{\hat{\eta}_f \sigma v_t}\right) \right) \quad (10)$$

We call $\hat{\boldsymbol{\theta}}_T$ the two step non-Gaussian QMLE, 2SNG-QMLE for short. Alternatively, this two-step procedure can be viewed as a one-step generalized methods of moments (GMM) procedure, by considering the score functions. Denote

$\tilde{\mathbf{s}}(x, \boldsymbol{\theta}, \eta, \boldsymbol{\phi}) = (s_1(x, \boldsymbol{\theta}), s_2(x, \boldsymbol{\theta}, \eta), s_3(x, \eta, \boldsymbol{\phi}))'$, where

$$s_1(x_t, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l_1(x_t, \boldsymbol{\theta}) = \mathbf{k} \left(-1 + \frac{x_t^2}{\sigma^2 v_t^2} \right) \quad (11)$$

$$s_2(x_t, \boldsymbol{\theta}, \eta) = \frac{\partial}{\partial \eta} l_2(x_t, \boldsymbol{\theta}, \eta) = -\frac{1}{\eta} \left(1 + h\left(\frac{x_t}{\eta \sigma v_t}\right) \right) \quad (12)$$

$$s_3(x_t, \eta, \boldsymbol{\phi}) = \frac{\partial}{\partial \boldsymbol{\phi}} l_3(x_t, \eta, \boldsymbol{\phi}) = -\mathbf{k} \left(1 + h\left(\frac{x_t}{\eta \sigma v_t}\right) \right) \quad (13)$$

then the estimators are obtained using GMM with identity weighting matrix:

$$(\tilde{\boldsymbol{\theta}}_T, \hat{\eta}_f, \hat{\boldsymbol{\phi}}_T) = \operatorname{argmin}_{\boldsymbol{\theta}, \eta, \boldsymbol{\phi}} \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{s}}'(x_t, \boldsymbol{\theta}, \eta, \boldsymbol{\phi}) \tilde{\mathbf{s}}(x_t, \boldsymbol{\theta}, \eta, \boldsymbol{\phi}) \quad (14)$$

so our proposed estimator is simply $\hat{\boldsymbol{\theta}}_T = \hat{\boldsymbol{\phi}}_T$.

3.4 Asymptotic Theory

Identification for the parameters $\boldsymbol{\theta}$ and η is straightforward. As in Theorem 1, the consistency thereby holds:

Theorem 3. *Given Assumption 1, $(\tilde{\boldsymbol{\theta}}_T, \hat{\eta}_f, \hat{\boldsymbol{\theta}}_T) \xrightarrow{\mathcal{P}} (\boldsymbol{\theta}_0, \eta_f, \boldsymbol{\theta}_0)$, in particular the 2SNG-QMLE $\hat{\boldsymbol{\theta}}_T$ is consistent.*

In order to obtain the asymptotic normality, we realize that a finite fourth moment for the innovation is essential in that the first step employs the Gaussian QMLE. Although alternative rate efficient estimators may be adopted to avoid moment conditions required in the first step, we prefer the Gaussian QMLE for its simplicity and popularity in practice.

Theorem 4. Assume that $E(\varepsilon^4) < \infty$, that Assumptions 1 and 2 are satisfied. Then $(\tilde{\boldsymbol{\theta}}_T, \hat{\eta}_f, \hat{\boldsymbol{\theta}}_T)$ are jointly normal asymptotically. That is,

$$\begin{pmatrix} T^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ T^{\frac{1}{2}}(\hat{\eta}_f - \eta_f) \\ T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(\begin{pmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_G & \boldsymbol{\Pi}' & \boldsymbol{\Xi} \\ \boldsymbol{\Pi} & \Sigma_{\eta_f} & \boldsymbol{\Pi} \\ \boldsymbol{\Xi} & \boldsymbol{\Pi}' & \boldsymbol{\Sigma}_2 \end{pmatrix} \right)$$

where

$$\boldsymbol{\Sigma}_G = \frac{E(\varepsilon^2 - 1)^2}{4} \mathbf{M}^{-1} \quad (15)$$

$$\boldsymbol{\Sigma}_2 = \frac{Eh_1(\varepsilon)^2}{(Eh_2(\varepsilon))^2} \mathbf{M}^{-1} + \sigma_0^2 \left(\frac{E(\varepsilon^2 - 1)^2}{4} - \frac{Eh_1(\varepsilon)^2}{(Eh_2(\varepsilon))^2} \right) \mathbf{e}_1 \mathbf{e}_1' \quad (16)$$

$$\Sigma_{\eta_f} = \eta_f^2 E \left(\frac{\varepsilon^2 - 1}{2} - \frac{h_1(\varepsilon)}{Eh_2(\varepsilon)} \right)^2 \quad (17)$$

$$\boldsymbol{\Pi} = \frac{\eta_f \sigma_0}{2} E \left((\varepsilon^2 - 1) \left(\frac{h_1(\varepsilon)}{Eh_2(\varepsilon)} - \frac{\varepsilon^2 - 1}{2} \right) \right) \mathbf{e}_1' \quad (18)$$

$$\boldsymbol{\Xi} = \frac{E(h_1(\varepsilon) \cdot (\varepsilon^2 - 1))}{2E(h_2(\varepsilon))} \mathbf{M}^{-1} - \frac{\sigma_0^2}{2} E \left((\varepsilon^2 - 1) \left(\frac{h_1(\varepsilon)}{Eh_2(\varepsilon)} - \frac{\varepsilon^2 - 1}{2} \right) \right) \mathbf{e}_1 \mathbf{e}_1' \quad (19)$$

where \mathbf{e}_1 is a unit column vector that has the same length as $\boldsymbol{\theta}$, with the first entry one and all the rest zeros.

Before a thorough efficiency analysis of the non-Gaussian QMLE $\hat{\boldsymbol{\theta}}_T$, we first discuss the asymptotic property of $\hat{\eta}_f$. Although $\hat{\eta}_f$ is obtained using fitted residuals $\tilde{\varepsilon}_t$ in (8), the asymptotic variance of $\hat{\eta}_f$ is not necessarily worse than that using the actual innovations ε_t . In fact, with true innovation the asymptotic variance of the η_f estimator is $\eta_f^2 E h_1^2 / (E h_2)^2$. Comparing it with (17), we can find that using fitted residual improves the efficiency as long as the $|h_1 / E h_2 - (\varepsilon^2 - 1) / 2|$ is smaller

than $|h_1/Eh_2|$. This occurs when the non-Gaussian likelihood is close to Gaussian likelihood. One extreme case is choosing the same Gaussian likelihood in the second step. Then η_f exactly equals one and the asymptotic variance of $\hat{\eta}_f$ vanishes.

η_f also reveals the issue of asymptotic bias incurred by using unscaled non-Gaussian QMLE. From 10, while 2SNG-QMLE $\hat{\boldsymbol{\theta}}_T = (\hat{\sigma}_T, \hat{\boldsymbol{\gamma}}_T)$ maximizes the log-likelihood, unscaled non-Gaussian QMLE would choose estimator $(\hat{\eta}_f \hat{\sigma}_T, \hat{\boldsymbol{\gamma}}_T)$ to maximize log-likelihood without η_f in it. So for the volatility scale parameter σ it is biased exactly according to the $\hat{\eta}_f$. Such bias will propagate if using the popular original parametrization. Recall

$$\begin{aligned} x_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \tilde{c} + \sum_{i=1}^p \tilde{a}_i x_{t-i}^2 + \sum_{j=1}^q \tilde{b}_j \sigma_{t-j}^2 \end{aligned}$$

Clearly, we have $\sigma^2 a_i = \tilde{a}_i$, $b_j = \tilde{b}_j$ and $\sigma^2 = c$. Therefore, potential model misspecification would result in systematic biases in the all estimates of a_i and c if unscaled non-Gaussian MLE, such as Student's t -likelihood, is applied without introducing η_f .

3.5 Efficiency Gain over Gaussian QMLE

We compare the efficiency of three estimators of $\boldsymbol{\theta}$ using two step non-Gaussian QMLE, one step (infeasible) non-Gaussian QMLE with known η_f , and Gaussian QMLE. Their asymptotic variances are $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_G$ respectively. The difference

in asymptotic variances between the first two estimators is

$$\Sigma_2 - \Sigma_1 = \begin{pmatrix} \mu\sigma_0^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (20)$$

where

$$\mu = \frac{E(\varepsilon^2 - 1)^2}{4} - \frac{Eh_1^2}{(Eh_2)^2} \quad (21)$$

Effectively, the sign and magnitude of μ summarize the advantage of knowing η_f . μ is usually positive when the true error has heavy tails while non-Gaussian QMLE is selected to be a heavy-tailed likelihood, illustrating the loss from not knowing η_f . However, it could also be negative when the true innovation has thin tail, indicating that not knowing η_f is actually better when a heavy tail density is selected. Intuitively, this is because the two-step estimator incorporates a more efficient Gaussian QMLE into the estimation procedure. More importantly, the asymptotic variance of γ and the covariance between σ and γ are not affected by the estimation of η_f . In other words, we achieve the adaptivity property for γ : with an appropriate non-Gaussian QMLE, γ could be estimated without knowing η_f equally well as if η_f were known before.

We next compare the efficiency between Gaussian QMLE and 2SNG-QMLE. By (7) with f replaced by the Gaussian likelihood, we have

$$\Sigma_G = \frac{E(\varepsilon^2 - 1)^2}{4} \mathbf{M}^{-1}$$

It follows from Lemma 3 in the appendix that,

$$\Sigma_G - \Sigma_2 = \mu \begin{pmatrix} \sigma_0^2 \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 & -\sigma_0 \bar{\mathbf{y}}_0' \mathbf{V} \\ -\sigma_0 \mathbf{V} \bar{\mathbf{y}}_0 & \mathbf{V} \end{pmatrix} \quad (22)$$

where $\mathbf{y}_0 = v_t(\gamma_0) \frac{\partial v_t(\gamma_0)}{\partial \gamma}$, $\bar{\mathbf{y}}_0 = E(\mathbf{y}_0)$, $\mathbf{V} = \text{Var}(\mathbf{y}_0)^{-1}$ and hereby the last matrix in (22) is positive definite. Therefore as long as μ is positive, non-Gaussian QMLE is more efficient for both σ and γ .

It is well known that the financial data sets such as stock prices and exchange rates exhibit heavy tails. Therefore, if a selected likelihood has heavier tails than Gaussian density, then μ is positive, and the efficiency of the QMLE is thereby improved over Gaussian QMLE.

3.6 Efficiency Gap from the MLE

Denote the asymptotic variance of the MLE as Σ_M . By (7) with f replaced by the true likelihood g , we have

$$\Sigma_M = \mathbf{M}^{-1} \left(\int_{-\infty}^{+\infty} x^2 \frac{\dot{g}^2}{g} dx - 1 \right)^{-1} = \mathbf{M}^{-1} (E(h_g^2) - 1)^{-1}$$

where $h_g = x \frac{g(x)}{g(x)}$. The gap in asymptotic variance between 2SNG-QMLE and MLE is given by

$$\Sigma_2 - \Sigma_M = \left(\frac{E h_1^2}{(E h_2)^2} - (E(h_g^2) - 1)^{-1} \right) \mathbf{M}^{-1} + \sigma_0^2 \left(\frac{E(\varepsilon^2 - 1)^2}{4} - \frac{E h_1^2}{(E h_2)^2} \right) \mathbf{e}_1 \mathbf{e}_1'$$

An extreme case is that the selected likelihood f happens to be the true innovation density. Being unaware of it, we still apply a two-step procedure and use the estimated η_f . Therefore, the first term in (23) vanishes, but the second term remains. Consequently, $\hat{\gamma}$ reach the efficiency bounds, while the volatility scale $\hat{\sigma}$ fails, reflecting the penalty of ignorance of the true model. This example also sheds light on the fact that $\hat{\boldsymbol{\theta}}_T$ cannot obtain the efficiency bounds for all parameters unless the true underlying density and the selected likelihood are both Gaussian. This observation agrees with the comparison study in the Gonzalez-Rivera and Drost (1999) concerning the MLE and their semiparametric estimator.

3.7 The Effect of the First Step Estimation

We would like to further explore the oracle property of the estimator for heteroscedastic parameters by considering a general first step estimator. We have shown in Theorem 4 that the efficiency of the estimator for $\boldsymbol{\gamma}$ is not affected by the first step estimation of η_f using Gaussian QMLE, as if η_f were known. Therefore, we may relax the finite fourth moment requirement on the innovation error by applying another efficient estimator in the first step. On the other hand, even if the first step estimator suffers from a lower rate, it may not affect the efficiency of the heteroscedastic parameters $\boldsymbol{\gamma}$, which is always $T^{\frac{1}{2}}$ consistent and asymptotically normal.

Theorem 5. *Suppose that the first step estimator $\tilde{\boldsymbol{\theta}}$ has an influence function representation:*

$$T\lambda_T^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \lambda_T^{-1} \sum_{t=1}^T \boldsymbol{\Psi}_t(\varepsilon_t) + o_P(1)$$

with the right hand side converging to a non-degenerate distribution, and $\lambda_T \sim T^{1/\alpha}$ for some $\alpha \in [1, 2]$. Then the convergence rate for σ is also $T\lambda_T^{-1}$, while the same central limit theorem for γ as in Theorem 3 remains, that is,

$$T^{\frac{1}{2}}(\gamma - \gamma_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \frac{Eh_1^2}{(Eh_2)^2} \mathbf{V})$$

where $\mathbf{V} = \left(\text{Var}\left(\frac{1}{\nu(\gamma_0)} \frac{\partial \nu}{\partial \gamma} \middle|_{\gamma=\gamma_0}\right) \right)^{-1}$.

Theorem 5 applies to several estimators that have been discussed in the literature. For example, Hall and Yao (2003) have discussed the Gaussian QMLE with ultra heavy-tailed innovations that violate a finite fourth moment. In their analysis, λ_T is regularly varying at infinity with exponent $\alpha \in [1, 2)$. The resulting Gaussian QMLE $\tilde{\boldsymbol{\theta}}$ suffers lower convergence rates. By contrast, Drost and Klaassen (1997) have suggested an M-estimator based on the score function for logistic distribution to avoid moment conditions on the innovations. Both estimators, if applied in the first step, would not affect the efficiency of $\hat{\gamma}_T$.

4 Extensions

We discuss two ways to further improve the efficiency of 2SNG-QMLE. One is choosing the non-Gaussian likelihood from a pool of candidate likelihoods to adapt to data, the other is an affine combination of 2SNG-QMLE and Gaussian QMLE according their covariance matrix in Theorem 4 to minimize resulting estimator's asymptotic variance.

4.1 Optimal Choice of Likelihood

There are two distinctive edges of choosing a heavy tailed quasi likelihood over Gaussian likelihood. First, the $T^{\frac{1}{2}}$ -consistency of 2SNG-QMLE of γ no longer depends on finite fourth moment condition, but instead finite $Eh_1^2/(Eh_2)^2$. This can be easily met by, for example, choosing generalized Gaussian likelihood with $\beta \leq 1$. Second, even under finite fourth moment, heavy tailed 2SNG-QMLE has lower variance than Gaussian QMLE if true innovation is heavy tailed. A pre-specified heavy tailed likelihood can have these two advantages. However, we can adaptively choose this quasi likelihood to further improve its efficiency. This is done by minimizing the efficiency loss from MLE, which is equivalent by minimizing $Eh_1^2/(Eh_2)^2$ over certain families of heavy tailed likelihoods. We propose optimal choice of non-Gaussian likelihoods, where candidate likelihoods are from Student's t family with degree of freedom $\nu > 2$ and generalized Gaussian family with $\beta \leq 1$. Formally, for true innovation distribution g and candidate likelihood f , define

$$A(f, g) = \frac{E_g h_1^2}{E_g (h_2)^2}, \quad \text{where } h_1 = 1 + h\left(\frac{\varepsilon}{\eta_f}\right), \text{ and } h_2 = \frac{\varepsilon}{\eta_f} \dot{h}\left(\frac{\varepsilon}{\eta_f}\right) \quad (23)$$

Then the optimal likelihood is chosen from t -family and generalized Gaussian family (gg):

$$f^* = \operatorname{argmin}_{\nu, \beta} \left\{ \{A(f_\nu^t, \hat{g})\}_{\nu > 2}, \{A(f_\beta^{gg}, \hat{g})\}_{\beta \leq 1} \right\} \quad (24)$$

where \hat{g} denotes the empirical distribution of estimated residuals from Gaussian QMLE, the first step. Because this procedure of choosing likelihood is adaptive to data, it is expected that the chosen quasi likelihood results in a more efficient

2SNG-QMLE than a pre-specified one. We justify this point in simulation studies.

A 2SNG-QMLE with choosing optimal likelihood runs the following four steps: (a) Run Gaussian QMLE and obtain the estimated residuals; (b) Run optimization (24) and obtain the optimal likelihood f^* ; (c) Obtain $\hat{\eta}_f$ using f^* and estimated residuals; (d) Run 2SNG-QMLE with f^* and $\hat{\eta}_f$.

4.2 Aggregating 2SNG-QMLE and Gaussian QMLE

Another way to further improve the efficiency of 2SNG-QMLE is through aggregation. Since both Gaussian QMLE and 2SNG-QMLE are consistent, an affine combination of the two, with weights chosen according to their joint asymptotic variance, yields a consistent estimator and is more efficient than both. Define the aggregation estimator

$$\hat{\theta}_T^{\mathbf{W}} = \mathbf{W}\hat{\theta} + (\mathbf{I} - \mathbf{W})\tilde{\theta} \quad (25)$$

where \mathbf{W} is a diagonal matrix with weights $(w_1, w_2, \dots, w_{1+p+q})$ on the diagonal. From Theorem 4, the optimal weights are chosen from minimizing the asymptotic variance of each component of the aggregation estimator:

$$\begin{aligned} w_i^* &= \operatorname{argmin}_w w^2(\Sigma_2)_{i,i} + (1-w)^2(\Sigma_G)_{i,i} + 2w(1-w)\Xi_{i,i} \\ &= \frac{(\Sigma_G)_{i,i} - \Xi_{i,i}}{(\Sigma_2)_{i,i} + (\Sigma_G)_{i,i} - 2\Xi_{i,i}} \end{aligned} \quad (26)$$

It turns out that all optimal aggregation weights w_i^* are the same, which is

$$w^* = \frac{E\left(\frac{1-\varepsilon^2}{2}\left(\frac{1-\varepsilon^2}{2} + \frac{h_1}{Eh_2}\right)\right)}{E\left(\frac{1-\varepsilon^2}{2} + \frac{h_1}{Eh_2}\right)^2}. \quad (27)$$

Proposition 2. *The aggregated estimator $\hat{\boldsymbol{\theta}}_T^*$ uses optimal aggregation weights $\mathbf{W}^* = w^* \mathbf{I}$. Its asymptotic variance has diagonal terms*

$$\Sigma_{i,i}^* = \frac{(\boldsymbol{\Sigma}_2)_{i,i}(\boldsymbol{\Sigma}_G)_{i,i} - \boldsymbol{\Xi}_{i,i}^2}{(\boldsymbol{\Sigma}_2)_{i,i} + (\boldsymbol{\Sigma}_G)_{i,i} - 2\boldsymbol{\Xi}_{i,i}}, \quad i = 1, \dots, 1 + p + q. \quad (28)$$

Although estimators for σ and γ have different asymptotic properties, the optimal aggregation weights are the same: $w_i^* = w^*$. Also the weight depends only on non-Gaussian likelihood and innovation distribution, but not on GARCH model specification. The aggregated estimator $\hat{\boldsymbol{\theta}}_T^*$ always have smaller asymptotic variance than both 2SNG-QMLE and Gaussian QMLE. If data is heavy tailed, e.g., $E\varepsilon^4$ is large or equal to ∞ , from (27) it simply assigns weights approximately 1 for 2SNG-QMLE and 0 for Gaussian QMLE. In practice, after running 2SNG-QMLE with optimal choice of likelihood, we can estimate the optimal aggregation weight w^* by plugging into (27) the estimated residuals.

5 Simulation Studies

5.1 Model Free Characteristics

The scale tuning parameter η_f and the efficiency difference μ are generic characteristics of non-Gaussian likelihoods and of the true innovations, and they do not change when using another conditional heteroscedastic model. We numerically evaluate how they vary according to the non-Gaussian likelihoods and innovations.

Table 1: η_f for generalized Gaussian QMLEs (gg ,row) and innovation distributions (column)

	$gg_{0.2}$	$gg_{0.6}$	gg_1	$gg_{1.4}$	$gg_{1.8}$	gg_2	t_3	t_5	t_7	t_{11}
$gg_{0.2}$	1.000	6.237	8.901	10.299	11.125	11.416	8.128	9.963	10.483	10.885
$gg_{0.6}$	0.271	1.000	1.291	1.434	1.515	1.544	1.159	1.384	1.443	1.487
$gg_{1.0}$	0.354	0.844	1.000	1.073	1.114	1.128	0.900	1.040	1.074	1.098
$gg_{1.4}$	0.537	0.873	0.962	1.000	1.022	1.029	0.883	0.977	0.998	1.012
$gg_{1.8}$	0.811	0.952	0.981	0.993	1.000	1.002	0.946	0.985	0.991	0.997

Table 2: η_f for Student's t QMLEs (row) and innovation distributions (column)

	$t_{2.5}$	t_3	t_4	t_5	t_7	t_{11}	$gg_{0.5}$	gg_1	$gg_{1.5}$	gg_2
$t_{2.5}$	1.000	1.231	1.425	1.506	1.584	1.641	0.900	1.414	1.614	1.716
t_3	0.815	1.000	1.151	1.216	1.275	1.318	0.756	1.150	1.301	1.375
t_4	0.715	0.874	1.000	1.054	1.100	1.133	0.697	1.011	1.122	1.174
t_5	0.690	0.836	0.953	1.000	1.043	1.071	0.691	0.966	1.061	1.107
t_7	0.679	0.816	0.922	0.964	1.000	1.024	0.708	0.945	1.018	1.053
t_{11}	0.690	0.823	0.916	0.953	0.980	1.000	0.749	0.941	0.998	1.021
t_{20}	0.720	0.845	0.928	0.958	0.981	0.992	0.811	0.954	0.992	1.007
t_{30}	0.742	0.862	0.939	0.965	0.981	0.992	0.846	0.966	0.993	1.004

Table 1 and 2 show how η_f varies over generalized Gaussian likelihoods and

Student's t likelihoods with different parameters respectively. For each row, which amounts to fixing a quasi likelihood, the lighter the tails of innovation errors are, the larger η_f . Furthermore $\eta_f > 1$ for innovation errors that are lighter than the likelihood, and $\eta_f < 1$ for innovations that are heavier than the likelihood. Therefore if the non-Gaussian likelihood have heavier tails than true innovation, we should shrink the data in order for consistent estimation. On the other hand if the quasi likelihood is lighter than true innovation, we should magnify the data.

For each column (fix an innovation distribution), in most cases the heavier the tails of likelihoods are, the larger η_f , but the monotone relationship is not true for some ultra heavy tail innovations, in which cases η_f shows a “smile” dynamic. The non-monotonicity in the likelihood dimension indicates that to determine η_f one needs more information about the likelihood than just the asymptotic behavior of its tails.

Table 3 and 4 show the dependence of μ on the true innovation (column) and non-Gaussian likelihood (row). From the table we see that in most cases μ is positive, which means that non-Gaussian QMLE shows an improvement. But when heavy tailed likelihoods are applied on true innovations with moderate or thin tails, μ turns negative, which means that Gaussian QMLE performs better.

Looking at each column, by fixing the innovation distribution, non-Gaussian QMLE performs the best over Gaussian QMLE when the non-Gaussian likelihood coincides with the innovation distribution (MLE). Looking at each row, by fixing a non-Gaussian likelihood, its relative performance increases when the true innovation distributions become more heavy tailed, even after passing the MLE point where

true innovation and likelihood are the same. This is because μ is a relative measure of non-Gaussian over Gaussian, not an absolute measure for asymptotic variance. When the true innovation is heavier than the non-Gaussian likelihood, non-Gaussian QMLE does not perform as well as MLE, but Gaussian QMLE does even worse than as if the true innovation coincides with non-Gaussian likelihood. Therefore, even the absolute efficiency in terms of asymptotic variance drops for non-Gaussian QMLE, its relative performance over Gaussian QMLE actually increases.

To summarize the variation of μ , one can draw a line for distributions according to their asymptotic behavior of tails, in other words, according to how heavy their tails are, with thin tails on the left and heavy tails on the right. Then we place non-Gaussian likelihood, Gaussian likelihood and true innovation distribution onto this line. The sign and value of μ depends on where true innovation distribution is placed. (a) It is placed on the right side of non-Gaussian likelihood, then μ is positive and large. (b) Error is on the left side of Gaussian, then μ is negative and large in absolute value. (c) Error is between non-Gaussian and Gaussian, then, to which likelihood is innovation closer determines μ . This seems like a symmetric argument for Gaussian and non-Gaussian likelihood. But in financial applications we know true innovations are heavy tailed. Even the non-Gaussian likelihood may not be the innovation distribution, we still can guarantee either (a) happens or (c) happens with innovation closer to non-Gaussian likelihood. In both cases we have $\mu > 0$ and non-Gaussian QMLE is a more efficient procedure than Gaussian QMLE.

Table 3: μ for generalized Gaussian QMLEs (gg, row) and innovation distributions (column)

	$gg_{0.2}$	$gg_{0.6}$	gg_1	$gg_{1.4}$	$gg_{1.8}$	gg_2	$t_{4.5}$	t_5	t_7	t_{11}
$gg_{0.2}$	484	1.773	-0.062	-0.335	-0.416	-0.436	2.411	0.929	-0.026	-0.274
$gg_{0.6}$	482	1.978	0.195	-0.075	-0.157	-0.178	2.608	1.138	0.206	-0.030
$gg_{1.0}$	474	1.839	0.250	0.017	-0.053	-0.071	2.590	1.149	0.267	0.054
$gg_{1.4}$	443	1.424	0.209	0.040	-0.010	-0.022	2.369	1.008	0.234	0.068
$gg_{1.8}$	328	0.589	0.089	0.022	0.003	-0.002	1.588	0.596	0.114	0.032

Table 4: μ for Student's t QMLEs (row) and innovation distributions (column)

	$t_{4.5}$	t_5	t_7	t_9	t_{15}	t_{30}	$gg_{0.5}$	gg_1	$gg_{1.5}$	gg_2
$t_{2.5}$	2.534	1.045	0.071	-0.114	-0.263	-0.324	3.848	0.004	-0.296	-0.375
t_3	2.626	1.145	0.189	0.011	-0.124	-0.183	3.871	0.124	-0.158	-0.223
t_4	2.663	1.194	0.258	0.086	-0.038	-0.090	3.816	0.191	-0.067	-0.124
t_5	2.664	1.200	0.277	0.114	-0.004	-0.054	3.770	0.211	-0.031	-0.084
t_7	2.642	1.190	0.287	0.131	0.020	-0.022	3.667	0.222	-0.001	-0.051
t_{11}	2.591	1.150	0.277	0.132	0.035	-0.004	3.500	0.212	0.016	-0.025

5.2 Verification of the Asymptotic Theory

Now we verify the asymptotic formula (16)-(17). We run $N = 20000$ simulations, each generating a sample of size $T = 7000$ from a GARCH(1, 1) model. The model parameters are $\sigma_0 = 0.5$, $a_{10} = 0.35$, $b_{10} = 0.3$. The innovation errors are standardized skewed Student's t distribution with degree of freedom $\nu_0 = 7$ and degree of skewness $\lambda_0 = 0.5$, so that the left tail is heavier than the right tail. We use Student's t likelihood with degree of freedom $\nu = 4$ in non-Gaussian QMLE. We run two-step procedure to obtain the estimates $\hat{\eta}_f$ and non-Gaussian QMLE estimates $\hat{\theta}$. Figure 1 reports the standardized estimates of $(\sigma_0, a_{10}, b_{10}, \eta_f)$ compared to $N(0, 1)$.

Standardization is done by first subtracting the estimates by the true value, and then dividing by the theoretical asymptotic standard deviation according to Theorem 4. All plots confirm the validity of asymptotic variance formula (16)-(17).

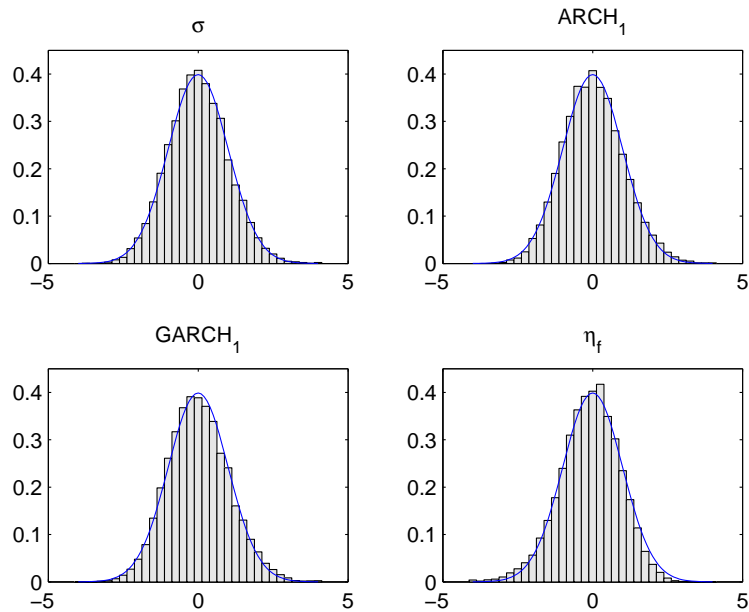


Figure 1: Histogram of standardized 2SNG-QMLE and standard normal pdf (solid line). Normalization is done by first subtracting the estimates by the true value, and then dividing by the theoretical asymptotic standard deviation suggested by our theory.

5.3 Comparison with Gaussian QMLE and MLE

We compare the efficiency of 2SNG-QMLE, Gaussian QMLE and MLE under various innovation error distributions. We don't perform optimal choice of quasi likelihood in 2SNG-QMLE, instead fix the quasi likelihood to be Student's t distribution with

degree of freedom 4. The simulation is conducted on a GARCH(1, 1) model with true parameters $(\sigma_0, a_{1,0}, b_{1,0}) = (0.5, 0.35, 0.3)$. For innovation errors we use Student's t and generalized Gaussian distributions of various degrees of freedoms to generate data. For each type of innovation distribution, we run $N = 1000$ simulations each with $T = 3000$ samples. Tables 5 and 6 reports the relative efficiencies of these three estimators in terms of ratios of sample variances and MSEs. The first ratio, Gaussian/2SNG, indicates how 2SNG-QMLE outperforms (underperforms) Gaussian QMLE. The second ratio 2SNG/MLE indicates how far 2SNG-QMLE is from efficiency bound.

In Table 5 the innovation distributions range from thin-tailed t_{20} (approximately Gaussian) to heavy-tailed $t_{2.5}$. Biases are small so standard deviations and RMSEs are nearly the same. For the first two thin-tailed cases, t_{20} and t_{15} , Gaussian QMLE outperforms 2SNG-QMLE by a small margin. For all other cases 2SNG-QMLE outperforms Gaussian QMLE. In heavy tailed cases t_6 and t_5 , 2SNG-QMLE performs nearly as well as MLE, and reduces standard deviations by 15% to 60% from Gaussian QMLE. In ultra-heavy tail cases (t_4 , t_3 and $t_{2.5}$), since fourth moment no longer exists, Gaussian QMLE is not $T^{\frac{1}{2}}$ -consistent, and its estimation precision quickly deteriorates, sometimes to an intolerable level. In contrast 2SNQ-QMLE using t_4 likelihood does not require finite fourth moment for $T^{\frac{1}{2}}$ -consistent $a_{1,0}$ and $b_{1,0}$, so standard deviations for $a_{1,0}$ and $b_{1,0}$ are still nearly equal to MLE. Standard deviations of σ_0 are now larger than MLE, but still significantly smaller than Gaussian QMLE.

In Table 6, the innovations innovations range from thin tailed gg_4 to heavy tailed

Table 5: Student's t innovations simulation

Innov. dist.	Comparing methods	Ratio of variances			Ratio of MSEs		
		σ_0	$a_{1,0}$	$b_{1,0}$	σ_0	$a_{1,0}$	$b_{1,0}$
t_{20}	G./2SNG	0.929	0.901	0.936	0.929	0.898	0.936
	2SNG/MLE	1.092	1.122	1.089	1.091	1.126	1.089
t_{15}	G./2SNG	0.942	0.960	0.961	0.939	0.948	0.960
	2SNG/MLE	1.112	1.121	1.087	1.114	1.131	1.087
t_9	G./2SNG	1.115	1.186	1.108	1.118	1.185	1.109
	2SNG/MLE	1.109	1.022	1.020	1.019	1.023	1.020
t_7	G./2SNG	1.216	1.260	1.186	1.217	1.266	1.186
	2SNG/MLE	1.036	1.024	1.031	1.037	1.026	1.031
t_6	G./2SNG	1.355	1.528	1.302	1.355	1.552	1.303
	2SNG/MLE	1.	1.022	1.	1.	1.022	1.
t_5	G./2SNG	1.526	2.495	1.405	1.547	2.530	1.409
	2SNG/MLE	1.025	1.001	1.015	1.027	1.001	1.015
t_4	G./2SNG	2.074	7.244	1.847	2.125	7.478	1.858
	2SNG/MLE	1.065	1.	1.	1.071	1.	1.
t_3	G./2SNG	2.687	31.40	2.535	2.850	33.26	2.580
	2SNG/MLE	1.235	1.	1.	1.264	1.	1.
$t_{2.5}$	G./2SNG	1.960	93.91	2.649	2.051	101.5	2.664
	2SNG/MLE	2.371	1.037	1.062	2.625	1.037	1.062

$gg_{0.4}$. For innovation with $gg_{1.2}$ and heavier, 2SNG-QMLE starts to outperform Gaussian QMLE. In all cases, the Student t_4 2SNG-QMLE performs very close to MLE as indicated by standard deviations. In comparison, Gaussian QMLE's performance deteriorates as tails grow heavier, particularly in $gg_{0.6}$ and $gg_{0.4}$, although in these cases the fourth moments are finite.

Table 6: generalized Gaussian innovations simulation

Innov. dist.	Comparing methods	Ratio of variances			Ratio of MSEs		
		σ_0	$a_{1,0}$	$b_{1,0}$	σ_0	$a_{1,0}$	$b_{1,0}$
gg_4	G./2SNG	0.743	0.742	0.769	0.748	0.736	0.771
	2SNG/MLE	1.705	1.843	1.571	1.696	1.886	1.566
$Gauss.$	G./2SNG	0.811	0.717	0.850	0.808	0.706	0.850
	2SNG/MLE	1.233	1.395	1.176	1.238	1.416	1.176
$gg_{1.2}$	G./2SNG	1.045	1.007	1.019	1.047	1.006	1.016
	2SNG/MLE	1.076	1.113	1.070	1.076	1.117	1.071
gg_1	G./2SNG	1.091	1.210	1.073	1.090	1.201	1.073
	2SNG/MLE	1.084	1.120	1.074	1.086	1.130	1.074
$gg_{0.8}$	G./2SNG	1.258	1.736	1.237	1.239	1.689	1.229
	2SNG/MLE	1.082	1.022	1.044	1.096	1.068	1.048
$gg_{0.6}$	G./2SNG	1.653	2.623	1.526	1.663	2.650	1.527
	2SNG/MLE	1.089	1.135	1.061	1.100	1.144	1.061
$gg_{0.4}$	G./2SNG	1.951	4.619	1.772	1.958	4.760	1.764
	2SNG/MLE	1.170	1.204	1.095	1.191	1.210	1.098

5.4 Ultra-Heavy Tail Case

Here we compare the efficiency when innovation are transformations from stable- α distributions. Index α ranges from 1.9 down to 1.1, and the distributions are transformed such that they do not have fourth moments but have 2nd moment to be unity. Furthermore they are asymmetric and not unimodal. Since distribution functions are not explicit, MLE is difficult to obtain. Table 7 compares the performance between 2SNG-QMLE with optimally chosen quasi likelihood and Gaussian QMLE. We still use the GARCH(1,1) model with true parameters $(\sigma_0, a_{1,0}, b_{1,0}) = (0.5, 0.35, 0.3)$, and run $N = 2000$ simulation with $T = 3000$. The candidate quasi likelihoods are Student's t distributions with DoF from 20 to 2.5, and generalized Gaussian distributions

with DoF from 4 to 0.4.

Table 7: Stable innovations simulation

Innov. dist.	Comparing methods	Ratio of variances			Ratio of MSEs		
		σ_0	$a_{1,0}$	$b_{1,0}$	σ_0	$a_{1,0}$	$b_{1,0}$
$\alpha = 1.9$	G./NG-opt	1.266	1.446	1.205	1.285	1.470	1.215
$\alpha = 1.7$	G./NG-opt	2.502	5.072	2.175	2.551	5.301	2.178
$\alpha = 1.5$	G./NG-opt	5.381	148.9	4.004	5.605	154.8	3.954
$\alpha = 1.3$	G./NG-opt	9.774	499.1	6.911	10.10	524.5	6.868
$\alpha = 1.1$	G./NG-opt	16.08	1313	10.19	16.94	1445	9.960

Gaussian QMLE deteriorates as tails grow heavier (smaller α). In particular for $a_{1,0}$, it produces many large estimates, making substantial biases upward and intolerable standard deviation levels. In contrast, 2SNG-QMLE shows little sample bias and small standard deviations. It also shows that as innovations grows heavier, 2SNG-QMLE delivers smaller standard deviations.

For $\alpha = 1.9$ case, among 2000 simulations, the algorithm chooses Student's t quasi likelihoods for 1977 times, and Gaussian likelihoods 23 times. Among the chosen Student's t likelihoods, the degrees of freedom spread out from 5 to 20, and mostly concentrate on 6, 7, 9 and 12. For the rest four cases, all chosen quasi likelihoods are Student's t likelihoods. In case $\alpha = 1.7$, the degrees of freedom concentrate on 3 and 4, with a small fraction of 5. In $\alpha = 1.5$, around 1650 simulations choose $t_{2.5}$, the rest choose t_3 . In $\alpha = 1.3$ and $\alpha = 1.1$, all chosen quasi likelihoods are $t_{2.5}$, the most heavy tailed candidate.

6 Empirical Work

We run a simple GARCH(1, 1) model on Citigroup stock daily return from January 03, 2008 to Jan 15, 2010. There are 514 trading days in the data. We report the estimated parameters using old parametrization. The Gaussian QMLE estimates for $(c, \tilde{a}, \tilde{b})$ is $(0.6522, 0.2205, 0.7793)$. Clearly data shows high degree of persistence in that $\tilde{a} + \tilde{b} \approx 1$. The 2SNG-QMLE chooses $gg_{1.2}$ as optimal likelihood, and the estimates for model parameters and η_f are $(0.7689, 0.2075, 0.7728)$ and 1.0458, respectively. Since $\hat{\eta}_f$ deviates from 1 about 4.6%, there would be a significant bias if we run $gg_{1.2}$ QMLE without scale adjustment.

On the other hand, even a non-Gaussian QMLE allowing to estimate shape of quasi likelihood cannot guarantee consistency. In fact, such method only picks one likelihood in some distribution family that is ‘least’ biased for the data, but bias due to misspecification of innovation distribution remains. We perform unscaled generalized Gaussian QMLE with shape estimation. The estimated shape is $\hat{\beta} = 1.305$, which is close to the 1.2, the shape of optimal likelihood in 2SNG-QMLE. We fix shape 1.305 and run 2SNG-QMLE again, $\hat{\eta}_f$ is still 1.033. This means even allowing to estimate the shape in quasi-likelihood, unscaled non-Gaussian QMLE still incurs a 3.3% bias.

7 Conclusion

This paper regards on GARCH model estimation when innovation distribution is unknown, and it questions the efficiency issue of Gaussian QMLE and consistency

issue of currently used non-Gaussian QMLE. It proposed the 2SNG-QMLE to tackle both issues. The first step runs a Gaussian QMLE whose purpose is to identify the scale tuning parameter, η_f . The second step runs a non-Gaussian QMLE to estimate model parameters. The quasi likelihood f used in second step can be a pre-specified heavy tailed likelihoods, properly scaled by η_f . It can also be chosen from a pool of candidate distributions in order to adapt different characteristics of unknown innovation distribution.

The asymptotic theory of 2SNG-QMLE does not depend on any symmetric or uni-modal assumptions of innovations. By adopting a different parametrization proposed by Newey and Steigerwald (1997), and incorporating η_f , 2SNG-QMLE improves the estimation efficiency from Gaussian QMLE. We and show that the asymptotic behavior of 2SNG-QMLE can be broken down to two parts. For the heteroscedastic parameters γ , 2SNG-QMLE is always $T^{\frac{1}{2}}$ -consistent and asymptotically normal, whereas $T^{\frac{1}{2}}$ consistency of Gaussian QMLE relies on finite fourth moment assumption. When $E\varepsilon_t^4 < \infty$, 2SNG-QMLE outperforms Gaussian QMLE in term of smaller asymptotic variance, provided that innovation distribution is reasonably heavy tailed, which is common for financial data. For the scale part σ , 2SNG-QMLE is not always $T^{\frac{1}{2}}$ -consistent, but simulation shows that the estimation for σ is usually equally well as heteroscedastic parameters, γ . We also run simulation to compare the performance of Gaussian QMLE, 2SNG-QMLE and MLE. In most cases 2SNG-QMLE shows an edge and is close to MLE.

One possible generalization of 2SNG-QMLE is to linearly combine candidate quasi likelihoods in the second step. Instead of choosing a single likelihood, the

log-likelihood objective in the second step is a weighted average of candidate log-likelihoods. The weights are chosen adaptively to optimize the asymptotic variance. By such combination efficiency, it will cover more dimensions of innovation distributions, and the efficiency will be further improved.

A Appendix section

A.1 Proof of Lemma 1

Proof.

$$\begin{aligned}
E_t(l_t(\boldsymbol{\theta})) &= Q\left(\frac{\eta_f \sigma v_t(\boldsymbol{\gamma})}{\sigma_0 v_t(\boldsymbol{\gamma}_0)}\right) - \log \sigma_0 v_t(\boldsymbol{\gamma}_0) + \log \eta_f \\
&\leq Q(\eta_f) - \log \sigma_0 v_t(\boldsymbol{\gamma}_0) + \log \eta_f \\
&= E_t(l_t(\boldsymbol{\theta}_0))
\end{aligned}$$

By Assumption 1, the inequality holds with positive probability. Therefore, by iterated expectations, $\bar{L}_T(\boldsymbol{\theta}) < \bar{L}_T(\boldsymbol{\theta}_0)$. \square

A.2 Proof of Lemma 2

Proof. Given regularity conditions, we have $\dot{Q}(\eta) = -\frac{1}{\eta}E(1 + h(\frac{\varepsilon}{\eta}))$. Denote $H(\eta) = E(1 + h(\frac{\varepsilon}{\eta}))$. $\ddot{Q}(\eta) = \frac{1}{\eta^2}H(\eta) - \frac{1}{\eta}\dot{H}(\eta)$, where $\dot{H}(\eta) = -\frac{1}{\eta^2}E(\varepsilon \dot{h}(\frac{\varepsilon}{\eta}))$, for any $\eta > 0$. Because $E(\varepsilon \dot{h}(\frac{\varepsilon}{\eta})) < 0$, so $\dot{H}(\eta) > 0$. Next, $\lim_{\eta \rightarrow +\infty} H(\eta) = 1$, since

$$\lim_{\eta \rightarrow +\infty} |H(\eta) - 1| = \lim_{\eta \rightarrow +\infty} |E(h(\frac{\varepsilon}{\eta}))| \leq \lim_{\eta \rightarrow +\infty} \frac{E|\varepsilon|^p}{\eta^p} \rightarrow 0$$

On the other hand, by Fatou's lemma along with 1 and 4, we have

$$\limsup_{\eta \rightarrow 0+} H(\eta) = \limsup_{\eta \rightarrow 0+} E(1 + h(\frac{\varepsilon}{\eta})) \leq 1 + E(\limsup_{\eta \rightarrow 0+} h(\frac{\varepsilon}{\eta})) < 0$$

then $\lim_{\eta \rightarrow +\infty} H(\eta) = 1$, $\limsup_{\eta \rightarrow 0+} H(\eta) < 0$, and $\dot{H}(\eta) > 0$. Hence, there exists a unique constant $\eta_f \in (0, \infty)$ such that $H(\eta_f) = 0$, hence $\dot{Q}(\eta_f) = 0$, and $\ddot{Q}(\eta_f) < 0$.

This concludes the proof. \square

A.3 Proof of Theorem 1

Proof. The proof is similar to Elie and Jeantheau (1995) by verifying the conditions given in Pfanzagl (1969). \square

A.4 Proof of Theorem 2

Proof. Let $\rho_t(\boldsymbol{\theta}) = (\sigma v_t)^{-2} \mathbf{k}$ and $\sigma_t(\boldsymbol{\theta}) = \sigma v_t(\boldsymbol{\gamma})$. Define the vector-valued function $\boldsymbol{\psi}$ as

$$\boldsymbol{\psi}(\omega, \boldsymbol{\theta}) = \frac{\partial L_T}{\partial \boldsymbol{\theta}} = -\frac{1}{T} \sum_{t=1}^T \left(1 + \frac{\dot{f}(\frac{x_t}{\eta_f \sigma_t})}{f(\frac{x_t}{\eta_f \sigma_t})} \frac{x_t}{\eta_f \sigma_t} \right) \mathbf{k}$$

For convenience, we consider the parameters ranging within a local neighborhood of the true values as in Hall and Yao (2003). This simplification may not be critical, given that the estimator is proved to be consistent. By Taylor expansion,

$$\sigma_t(\boldsymbol{\theta})^2 = \sigma_t(\boldsymbol{\theta}_0)^2 + \mathbf{A}_t(\boldsymbol{\theta}_0)'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \mathbf{R}_{1t}(\boldsymbol{\theta})\sigma_t(\boldsymbol{\theta}_0)^2 \quad (29)$$

$$\boldsymbol{\rho}_t(\boldsymbol{\theta}) = \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \mathbf{B}_t(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \mathbf{R}_{2t}(\boldsymbol{\theta})\sigma_t(\boldsymbol{\theta}_0)^{-2} \quad (30)$$

where $\mathbf{R}_{1t}(\boldsymbol{\theta})$ and $\mathbf{R}_{2t}(\boldsymbol{\theta})$ are an r -vector and $r \times r$ matrix, and $r = 1 + p + q$.

On the other hand,

$$\begin{aligned} h\left(\frac{x_t}{\eta_f \sigma_t}\right) &= h\left(\frac{\varepsilon_t \sigma_t(\boldsymbol{\theta}_0)}{\eta_f \sigma_t(\boldsymbol{\theta})}\right) \\ &= h\left(\frac{\varepsilon_t}{\eta_f}\right) - \frac{\varepsilon_t}{\eta_f} \dot{h}\left(\frac{\varepsilon_t}{\eta_f}\right) \sigma_t(\boldsymbol{\theta}_0)^2 \boldsymbol{\rho}_t(\boldsymbol{\theta}_0)'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \mathbf{R}_{3t}(\boldsymbol{\theta}) \end{aligned} \quad (31)$$

where $\mathbf{R}_{3t}(\boldsymbol{\theta})$ is an r -vector.

It has been shown in Hall and Yao (2003) that for $\mathbf{R}_t(\boldsymbol{\theta}) = \mathbf{R}_{1t}(\boldsymbol{\theta})$, $\mathbf{R}_{2t}(\boldsymbol{\theta})$ and $\mathbf{R}_{3t}(\boldsymbol{\theta})$, component-wise,

$$P\left(T^{-1} \sum_{t=1}^T \sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \leq \xi} |\mathbf{R}_t(\boldsymbol{\theta})| \leq C\right) \longrightarrow 1 \quad (32)$$

with ξ sufficiently small. Therefore, we can rewrite the equation (29) as

$$\begin{aligned} \mathbf{0} &= \sum_{t=1}^T \left(1 + h\left(\frac{\varepsilon_t}{\eta_f}\right)\right) \sigma_t(\boldsymbol{\theta}_0)^2 \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sum_{t=1}^T \left(-\frac{\varepsilon_t}{\eta_f} \dot{h}\left(\frac{\varepsilon_t}{\eta_f}\right) \sigma_t(\boldsymbol{\theta}_0)^4 \boldsymbol{\rho}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0)\right. \\ &\quad \left.+ \left(1 + h\left(\frac{\varepsilon_t}{\eta_f}\right)\right) (\mathbf{A}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sigma_t(\boldsymbol{\theta}_0)^2 \mathbf{B}_t(\boldsymbol{\theta}_0)) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 T \mathbf{R}(\boldsymbol{\theta})\right) \end{aligned}$$

where

$$P\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\xi} |\mathbf{R}(\boldsymbol{\theta})| \leq C\right) \longrightarrow 1 \quad (33)$$

Note that

$$\begin{aligned} & E\left((1+h(\frac{\varepsilon_t}{\eta_f}))(\mathbf{A}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sigma_t(\boldsymbol{\theta}_0)^2 \mathbf{B}_t(\boldsymbol{\theta}_0))\right) \\ = & E\left((\mathbf{A}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sigma_t(\boldsymbol{\theta}_0)^2 \mathbf{B}_t(\boldsymbol{\theta}_0)) E_t\left(1+h(\frac{\varepsilon_t}{\eta_f})\right)\right) \\ = & E\left((\mathbf{A}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sigma_t(\boldsymbol{\theta}_0)^2 \mathbf{B}_t(\boldsymbol{\theta}_0))\right) E\left(1+h(\frac{\varepsilon_t}{\eta_f})\right) \\ = & \mathbf{0} \end{aligned} \quad (34)$$

Therefore, it may be proved from the ergodic theorem that

$$T^{-1} \sum_{t=1}^T (1+h(\frac{\varepsilon_t}{\eta_f}))(\mathbf{A}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) + \sigma_t(\boldsymbol{\theta}_0)^2 \mathbf{B}_t(\boldsymbol{\theta}_0)) \longrightarrow \mathbf{0} \quad (35)$$

$$T^{-1} \sum_{t=1}^T \frac{\varepsilon_t}{\eta_f} \dot{h}(\frac{\varepsilon_t}{\eta_f}) \sigma_t(\boldsymbol{\theta}_0)^4 \boldsymbol{\rho}_t(\boldsymbol{\theta}_0)' \boldsymbol{\rho}_t(\boldsymbol{\theta}_0) \longrightarrow ME\left(\frac{\varepsilon_t}{\eta_f} \dot{h}(\frac{\varepsilon_t}{\eta_f})\right) \quad (36)$$

Hence, we have

$$\begin{aligned} & (ME\left(\frac{\varepsilon_t}{\eta_f} \dot{h}(\frac{\varepsilon_t}{\eta_f})\right) + o_P(1))(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \mathbf{R}(\boldsymbol{\theta}) \\ = & T^{-1} \sum_{t=1}^T (1+h(\frac{\varepsilon_t}{\eta_f})) \mathbf{k}_0 \end{aligned} \quad (37)$$

where $o_P(1)$ does not depend on $\boldsymbol{\theta}$. It may be proved from martingale central limit

theorem that

$$T^{-1/2} \sum_{t=1}^T (1 + h(\frac{\varepsilon_t}{\eta_f})) \mathbf{k}_0 \xrightarrow{\mathcal{L}} N\left(0, \mathbf{M}E(1 + h(\frac{\varepsilon_t}{\eta_f}))^2\right) \quad (38)$$

then it follows from the same argument as in Hall and Yao (2003) that

$$\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 = O_p(T^{-1/2}) \quad (39)$$

and

$$(\mathbf{M}E\left(\frac{\varepsilon_t}{\eta_f} \dot{h}\left(\frac{\varepsilon_t}{\eta_f}\right)\right) + o_P(1))(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = T^{-1} \sum_{t=1}^T (1 + h(\frac{\varepsilon_t}{\eta_f})) \mathbf{k}_0 \quad (40)$$

Thus,

$$T^{1/2}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \mathbf{M}^{-1} \frac{E\left((1 + h(\frac{\varepsilon_t}{\eta_f}))^2\right)}{\left(E\left(\frac{\varepsilon_t}{\eta_f} \dot{h}\left(\frac{\varepsilon_t}{\eta_f}\right)\right)\right)^2}\right) \quad (41)$$

□

A.5 Proof of Proposition 1

Proof. Define the likelihood ratio function $G(\eta) = E(\log(\frac{\frac{1}{\eta}f(\frac{\varepsilon}{\eta})}{f(\varepsilon)}))$. Suppose $G(\eta)$ has no local extremal values. And since $\log(x) \leq 2(\sqrt{x} - 1)$,

$$\begin{aligned} E(\log(\frac{\frac{1}{\eta}f(\frac{\varepsilon}{\eta})}{f(\varepsilon)})) &\leq 2E(\sqrt{\frac{\frac{1}{\eta}f(\frac{\varepsilon}{\eta})}{f(\varepsilon)}} - 1) = 2 \int_{-\infty}^{+\infty} \sqrt{\frac{1}{\eta}f(\frac{x}{\eta})f(x)}dx - 2 \\ &\leq - \int_{-\infty}^{+\infty} (\sqrt{\frac{1}{\eta}f(\frac{x}{\eta})} - \sqrt{f(x)})^2 dx \\ &\leq 0 \end{aligned}$$

The equality holds if and only if $\eta = 1$. Therefore, $\eta = 1$ is the unique maximum of $Q(\eta)$. □

A.6 Proof of Theorem 4

In order to show the asymptotic normality, we first list some notations and derive a lemma. For convenience, we denote $\mathbf{y}_0 = \frac{1}{v_t(\gamma_0)} \frac{\partial v_t(\gamma_0)}{\partial \gamma}$ and $\bar{\mathbf{y}}_0 = E(\mathbf{y}_0)$, so $\mathbf{k}_0 = (\frac{1}{\sigma_0}, \mathbf{y}_0')'$ and $\bar{\mathbf{k}}_0 = E\mathbf{k}_0 = (\frac{1}{\sigma_0}, \bar{\mathbf{y}}_0)'$. Also, let $\mathbf{M} = E(\mathbf{k}_0\mathbf{k}_0')$, $\mathbf{N} = \bar{\mathbf{k}}_0\bar{\mathbf{k}}_0'$ and $\mathbf{V} = \text{Var}(\mathbf{y}_0)^{-1}$. All the expectations above are taken under the true density g .

Lemma 3. *The following claims hold:*

1. *The inverse of \mathbf{M} in block expression is*

$$\mathbf{M}^{-1} = \begin{pmatrix} \sigma_0^2(1 + \bar{\mathbf{y}}_0'\mathbf{V}\bar{\mathbf{y}}_0) & -\sigma_0\bar{\mathbf{y}}_0'\mathbf{V} \\ -\sigma_0\mathbf{V}\bar{\mathbf{y}}_0 & \mathbf{V} \end{pmatrix}; \quad (42)$$

$$2. \bar{\mathbf{k}}_0' \mathbf{M}^{-1} = \sigma_0 \mathbf{e}_1', \bar{\mathbf{k}}_0' \mathbf{M}^{-1} \mathbf{k}_0 = \bar{\mathbf{k}}_0' \mathbf{M}^{-1} \bar{\mathbf{k}}_0 = 1;$$

$$3. \mathbf{M}^{-1} \mathbf{N} \mathbf{M}^{-1} = \mathbf{M}^{-1} \mathbf{N} \mathbf{M}^{-1} \mathbf{N} \mathbf{M}^{-1} = \sigma_0^2 \mathbf{e}_1 \mathbf{e}_1', \text{ where } \mathbf{e}_1 \text{ is a unit column vector that has the same length as } \boldsymbol{\theta}, \text{ with the first entry one and all the rest zeros.}$$

Proof. The proof uses Moore-Penrose pseudo inverse described in Ben-Israel and Greville (2003). Observe that

$$\mathbf{M} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \text{Var}(\mathbf{y}_0) \end{pmatrix} + \bar{\mathbf{k}}_0 \bar{\mathbf{k}}_0'. \quad (43)$$

Use the technique of Moore-Penrose pseudo inverse,

$$\mathbf{M}^{-1} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \text{Var}(\mathbf{y}_0) \end{pmatrix}^+ + \mathbf{H} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} + \mathbf{H} \quad (44)$$

where \mathbf{H} is formed by the elements below:

$$\begin{aligned} \beta &= 1 + \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 \\ \mathbf{w} &= (\sigma_0^{-1}, \mathbf{0})' \\ \mathbf{m} &= \mathbf{w} \\ \mathbf{v} &= (0, \bar{\mathbf{y}}_0' \mathbf{V})' \\ \mathbf{n} &= \mathbf{v} \end{aligned}$$

$$\begin{aligned}
\mathbf{H} &= -\frac{1}{\|\mathbf{w}\|^2} \mathbf{v} \mathbf{w}' - \frac{1}{\|\mathbf{m}\|^2} \mathbf{m} \mathbf{n}' + \frac{\beta}{\|\mathbf{w}\|^2 \|\mathbf{m}\|^2} \mathbf{m} \mathbf{w}' \\
&= \sigma_0^2 \begin{pmatrix} 1 + \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 & -\sigma_0^{-1} \bar{\mathbf{y}}_0' \mathbf{V} \\ -\sigma_0^{-1} \mathbf{V} \bar{\mathbf{y}}_0 & \mathbf{0} \end{pmatrix}
\end{aligned}$$

So (42) is obtained by plugging \mathbf{H} into (44). The rest two points of the lemma can be obtained by simple matrix manipulation. \square

Next we return to the proof of Theorem 4.

Proof. According to Theorem 3.4 in Newey and McFadden (1986), $(\tilde{\boldsymbol{\theta}}_T, \hat{\eta}, \hat{\boldsymbol{\theta}}_T)$ are jointly $T^{\frac{1}{2}}$ -consistent and asymptotic normal. The asymptotic variance matrix is

$$\mathbf{G}^{-1} E(\tilde{\mathbf{s}}(\boldsymbol{\theta}_0, \eta_f, \boldsymbol{\theta}_0) \tilde{\mathbf{s}}(\boldsymbol{\theta}_0, \eta_f, \boldsymbol{\theta}_0)') \mathbf{G}'^{-1}. \quad (45)$$

where $\mathbf{G} = E(\nabla \tilde{\mathbf{s}}(x_t, \boldsymbol{\theta}_0, \eta_f, \boldsymbol{\theta}_0))$. View this matrix as 3×3 blocks, with asymptotic variances of $(\tilde{\boldsymbol{\theta}}_T, \hat{\eta}, \hat{\boldsymbol{\theta}}_T)$ on the first, second and third diagonal blocks. We now calculate the second and the third diagonal blocks. The expect Jacobian matrix G can be decomposed into

$$\mathbf{G} = E \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \mathbf{s}_1(x_t, \boldsymbol{\theta}_0) & \mathbf{0} & \mathbf{0} \\ \nabla_{\boldsymbol{\theta}} s_2(x_t, \boldsymbol{\theta}_0, \eta_f) & \nabla_{\eta} s_2(\boldsymbol{\theta}_0, \eta_f) & \mathbf{0} \\ \mathbf{0} & \nabla_{\eta} \mathbf{s}_3(x_t, \eta_f, \boldsymbol{\theta}_0) & \nabla_{\phi} \mathbf{s}_3(x_t, \eta_f, \boldsymbol{\theta}_0) \end{pmatrix}$$

Denote the corresponding blocks as G_{ij} , $i, j = 1, 2, 3$. Direct calculation yields

$$\begin{aligned}
\mathbf{G}_{11} &= -2\mathbf{M} \\
\mathbf{G}_{21} &= \frac{1}{\eta_f} E\left(\dot{h}\left(\frac{\varepsilon}{\eta_f}\right) \frac{\varepsilon}{\eta_f}\right) \bar{\mathbf{k}}_0' \\
G_{22} &= \frac{1}{\eta_f^2} E\left(\dot{h}\left(\frac{\varepsilon}{\eta_f}\right) \frac{\varepsilon}{\eta_f}\right) \\
\mathbf{G}_{32} &= \mathbf{G}_{21}' \\
\mathbf{G}_{33} &= E\left(\dot{h}\left(\frac{\varepsilon}{\eta_f}\right) \frac{\varepsilon}{\eta_f}\right) \mathbf{M}
\end{aligned}$$

The second diagonal block depends on the second row of \mathbf{G}^{-1} and $\tilde{\mathbf{s}}(x_t, \boldsymbol{\theta}_0, \eta_f, \boldsymbol{\theta}_0)$.

The second row of \mathbf{G}^{-1} is

$$(-G_{22}^{-1} \mathbf{G}_{21} \mathbf{G}_{11}^{-1} \quad G_{22}^{-1} \quad \mathbf{0})$$

So the asymptotic variance of $\hat{\eta}$ is $G_{22}^{-1} E(\mathbf{q}_2 \mathbf{q}_2') G_{22}'^{-1}$, where

$$\begin{aligned}
q_2 &= -\mathbf{G}_{21} \mathbf{G}_{11}^{-1} \mathbf{s}_1(x_t, \boldsymbol{\theta}_0) + s_2(x_t, \boldsymbol{\theta}_0, \eta_f) \\
&= \frac{1}{2\eta_f} E\left(\dot{h}\left(\frac{\varepsilon}{\eta_f}\right) \frac{\varepsilon}{\eta_f}\right) \bar{\mathbf{k}}_0' \overline{\mathbf{k}_0 \mathbf{k}_0'}^{-1} (\varepsilon^2 - 1) \mathbf{k}_0 - \frac{1}{\eta_f} \left(1 + h\left(\frac{\varepsilon}{\eta_f}\right)\right) \\
&= \frac{1}{\eta_f} \left(\frac{1}{2} E h_2(\varepsilon^2 - 1) - h_1\right)
\end{aligned}$$

The last step uses the second point of Lemma 3. So (17) is obtained by plugging in the expressions for G_{22} and q_2 . Similarly, the third row of \mathbf{G}^{-1} is

$$\mathbf{G}_{33}^{-1} (\mathbf{G}_{32} G_{22}^{-1} \mathbf{G}_{21} \mathbf{G}_{11}^{-1} \quad -\mathbf{G}_{32} G_{22}^{-1} \quad \mathbf{I})$$

The asymptotic variance for $\hat{\boldsymbol{\theta}}$ is $\mathbf{G}_{33}^{-1}E(\mathbf{q}_3\mathbf{q}_3')\mathbf{G}_{33}'^{-1}$, where

$$\begin{aligned}\mathbf{q}_3 &= \mathbf{G}_{32}G_{22}^{-1}(\mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{s}_1(x_t, \boldsymbol{\theta}_0) - \mathbf{s}_2(x_t, \boldsymbol{\theta}_0, \eta_f)) + \mathbf{s}_3(x_t, \eta_f, \boldsymbol{\theta}_0) \\ &= -(1 + h(\frac{\varepsilon}{\eta_f}))(\mathbf{k}_0 - \bar{\mathbf{k}}_0) - \frac{1}{2}E(\dot{h}(\frac{\varepsilon}{\eta_f})\frac{\varepsilon}{\eta_f})\bar{\mathbf{k}}_0\bar{\mathbf{k}}_0'(\overline{\mathbf{k}_0\mathbf{k}_0'})^{-1}\mathbf{k}_0(\varepsilon^2 - 1) \\ &= -h_1(\mathbf{k}_0 - \bar{\mathbf{k}}_0) - \frac{1}{2}(Eh_2)(\varepsilon^2 - 1)\bar{\mathbf{k}}_0\end{aligned}$$

The last step uses the second point of Lemma 3. Then

$$\begin{aligned}E\mathbf{q}_3\mathbf{q}_3' &= Eh_1^2(\mathbf{M} - \mathbf{N}) + \frac{1}{4}(Eh_2)^2E(\varepsilon^2 - 1)\mathbf{N} \\ &= Eh_1^2\mathbf{M} + \left(\frac{1}{4}E(\varepsilon^2 - 1)^2 - Eh_1^2\right)\mathbf{N}\end{aligned}$$

Therefore, (16) is obtained by plugging in the expressions for \mathbf{G}_{33} , $E\mathbf{q}_3\mathbf{q}_3'$, and apply the third point of Lemma 3.

The asymptotic covariance between $\hat{\boldsymbol{\theta}}$ and $\hat{\eta}_f$ is $\mathbf{G}_{33}^{-1}E(\mathbf{q}_3q_2)G_{22}'^{-1}$, then direct calculation using the second point of Lemma 3 yields

$$\boldsymbol{\Pi} = \frac{\eta_f\sigma_0}{2}E\left((\varepsilon^2 - 1)\left(\frac{h_1}{Eh_2} - \frac{\varepsilon^2 - 1}{2}\right)\right)\mathbf{e}_1'$$

The same formula recurs in the asymptotic covariance between $\tilde{\boldsymbol{\theta}}$ and $\hat{\eta}_f$, which is $\mathbf{G}_{11}^{-1}E(\mathbf{q}_1q_2)G_{22}'^{-1}$.

Finally, the asymptotic covariance between $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ is $\mathbf{G}_{11}^{-1}E(\mathbf{q}_1\mathbf{q}_3')\mathbf{G}_{33}'^{-1}$, denoted as $\boldsymbol{\Xi}$. It implies from the third point of Lemma 3 that

$$\boldsymbol{\Xi} = \frac{E(h_1(\varepsilon^2 - 1))}{2E(h_2)}\mathbf{M}^{-1} - \frac{\sigma_0^2}{2}E\left((\varepsilon^2 - 1)\left(\frac{h_1}{Eh_2} - \frac{\varepsilon^2 - 1}{2}\right)\right)\mathbf{e}_1\mathbf{e}_1'$$

which concludes the proof. \square

A.7 Proof of Theorem 5

Proof. Following the similar idea to GMM, we may prove:

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \lambda_T T^{-\frac{1}{2}} \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{32} & \mathbf{G}_{33} \end{pmatrix} \begin{pmatrix} T \lambda_T^{-1} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ T^{\frac{1}{2}} (\hat{\eta}_f - \eta_f) \\ T^{\frac{1}{2}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \end{pmatrix} = \begin{pmatrix} \frac{1}{\lambda_T} \sum_{t=1}^T \boldsymbol{\Psi}_t(\varepsilon_t) + o_P(\mathbf{1}) \\ -\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{1}{\eta_f} (1 + h(\frac{\varepsilon_t}{\eta_f})) + o_P(1) \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T (1 + h(\frac{\varepsilon_t}{\eta_f})) \mathbf{k}_0 + o_P(\mathbf{1}) \end{pmatrix}$$

Clearly, the corresponding weighting vector for $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)$ is

$$\begin{pmatrix} \mathbf{G}_{33}^{-1} \mathbf{G}_{32} G_{22}^{-1} \lambda_T T^{-\frac{1}{2}} \mathbf{G}_{21} & -\mathbf{G}_{33}^{-1} \mathbf{G}_{32} G_{22}^{-1} & \mathbf{G}_{33}^{-1} \end{pmatrix}$$

Note that

$$\begin{aligned} \mathbf{G}_{33}^{-1} \mathbf{G}_{32} G_{22}^{-1} \lambda_T T^{-\frac{1}{2}} \mathbf{G}_{21} &= \lambda_T T^{-\frac{1}{2}} \mathbf{M}^{-1} \bar{\mathbf{k}}_0 \bar{\mathbf{k}}_0' = \lambda_T T^{-\frac{1}{2}} \sigma_0 \mathbf{e}_1 \bar{\mathbf{k}}_0' \\ -\mathbf{G}_{33}^{-1} \mathbf{G}_{32} G_{22}^{-1} &= -\sigma_0 \eta_f (E\left(h\left(\frac{\varepsilon}{\eta_f}\right) \frac{\varepsilon}{\eta_f}\right))^{-1} \mathbf{e}_1 \end{aligned}$$

thus the sub-matrices corresponding to $\boldsymbol{\gamma}$ parameter are $\mathbf{0}$ s. Therefore, the first step has no effect on the central limit theorem of $\hat{\boldsymbol{\gamma}}_T$. The result follows from Lemma 3. In terms of $\hat{\sigma}_T$, its convergence rate becomes $T \lambda_T^{-1}$. \square

A.8 Proof of Proposition 2

Proof. Denote random variables $\kappa_G = (1 - \varepsilon^2)/2$, and $\kappa_2 = h_1(\varepsilon/\eta_f)/E(h_2(\varepsilon/\eta_f))$.

We show the optimal weights for σ and γ are the same. From Lemma 3, Theorem 4 and (27), for σ , the numerator in w_1^* is

$$\begin{aligned} (\Sigma_G)_{1,1} - \Xi_{1,1} &= \sigma_0^2(1 + \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0) E \kappa_G^2 - \sigma_0^2 E \kappa_G^2 + \sigma_0^2 \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 E(\kappa_G \kappa_2) \\ &= \sigma_0^2 \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 E(\kappa_G(\kappa_G + \kappa_2)) \end{aligned}$$

The denominator in w_1^* is

$$\begin{aligned} (\Sigma_G)_{1,1} + (\Sigma_2)_{1,1} - 2\Xi_{1,1} &= \sigma_0^2(1 + \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0)(E \kappa_G^2 + E \kappa_2^2) + \sigma_0^2(E \kappa_G^2 - E \kappa_2^2) \\ &\quad - 2\sigma_0^2 E \kappa_G^2 + 2\sigma_0^2 \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 E(\kappa_G \kappa_2) \\ &= \sigma_0^2 \bar{\mathbf{y}}_0' \mathbf{V} \bar{\mathbf{y}}_0 E(\kappa_G^2 + \kappa_2^2 + 2\kappa_G \kappa_2) \end{aligned}$$

Therefore we obtain $w_1^* = E(\kappa_G(\kappa_G + \kappa_2))/E(\kappa_G + \kappa_2)^2$. Now we compute the weights corresponding to γ . For $i = 2, \dots, 1 + p + q$, let $j = i - 1$, also from (27),

$$w_i^* = \frac{\mathbf{V}_{j,j} E \kappa_G^2 + \mathbf{V}_{j,j} E(\kappa_G \kappa_2)}{\mathbf{V}_{j,j} E \kappa_G^2 + \mathbf{V}_{j,j} E \kappa_2^2 + 2\mathbf{V}_{j,j} E(\kappa_G \kappa_2)} = \frac{E(\kappa_G(\kappa_G + \kappa_2))}{E(\kappa_G + \kappa_2)^2}$$

Therefore all the optimal aggregation weights are the same. □

References

- BEN-ISRAEL, A. and GREVILLE, T. (2003). *Generalized Inverses Theory and Applications*. Springer.
- BERKES, I. and HORVÁTH, L. (2004). The efficiency of the estimators of the parameters in garch processes. *The Annals of Statistics*, **32** 633–655.
- BERKES, I., HORVÁTH, L. and KOKOSZKA, P. (2003). Garch processes: Structure and estimation. *Bernoulli*, **9** 201–227.
- BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31** 307 – 327.
- BOLLERSLEV, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, **69** 542–547.
- BOLLEVSLEV, T. and WOOLDBRIDGE, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews*, **11** 143–172.
- BOUGEROL, P. and PICARD, N. (1992). Stationarity of garch processes and of some nonnegative time series. *Journal of Econometrics*, **52** 115–127.
- DIEBOLD, F. (1988). *Empirical modeling of exchange rate dynamics*. Springer-Verlag.

- DROST, F. C. and KLAASSEN, C. A. J. (1997). Efficient estimation in semiparametric garch models. *Journal of Econometrics*, **81** 193 – 221.
- DROST, F. C., KLAASSEN, C. A. J. and WERKER, B. J. M. (1997). Adaptive estimation in time-series models. *The Annals of Statistics*, **25** 786–817.
- ELIE, L. and JEANTHEAU, T. (1995). Consistency in heteroskedastic models. *Comptes Rendus de l 'Académie des Sciences*, **320** 1255–1258.
- ENGLE, R. and BOLLERSLEV, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, **5** 1–50.
- ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, **50** 987–1007.
- ENGLE, R. F. and GONZALEZ-RIVERA, G. (1991). Semiparametric arch models. *Journal of Business and Economic Statistics*, **9** 345–359.
- GONZLEZ-RIVERA, G. and DROST, F. C. (1999). Efficiency comparisons of maximum-likelihood-based estimators in garch models. *Journal of Econometrics*, **93** 93 – 111.
- HALL, P. and YAO, Q. (2003). Inference in arch and garch models with heavy-tailed errors. *Econometrica*, **71** 285–317.
- HSIEH, D. A. (1989). Modeling heteroscedasticity in daily foreign-exchange rates. *Journal of Business and Economic Statistics*, **7** 307–317.

- LEE, S.-W. and 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** 347–370.
- NEWKEY, W. K. and MCFADDEN, D. (1986). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (R. F. Engle and D. McFadden, eds.), vol. 4 of *Handbook of Econometrics*, chap. 36. Elsevier, 2111–2245.
- NEWKEY, W. K. and STEIGERWALD, D. G. (1997). Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica*, **65** 587–599.
- PFANZAGL, J. (1969). Consistent estimation of a location parameter in the presence of an incidental scale parameter. *The Annals of Mathematical Statistics*, **40** 1353–1357.
- WEISS, A. A. (1986). Asymptotic theory for arch models: Estimation and testing. *Econometric Theory*, **2** 107–131.