Nuisance parameters, composite likelihoods and a panel of GARCH models

Cavit Pakel
Department of Economics, University of Oxford
& Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
cavit.pakel@economics.ox.ac.uk

Neil Shephard
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
neil.shephard@economics.ox.ac.uk

Kevin Sheppard
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
kevin.sheppard@economics.ox.ac.uk

August 20, 2010

Abstract

We investigate the properties of the composite likelihood (CL) method for $(T \times N_T)$ GARCH panels. The defining feature of a GARCH panel with time-series length T is that, while nuisance parameters are allowed to vary across N_T series, other parameters of interest are assumed to be common. CL pools information across the panel instead of using information available in a single series only. Simulations and empirical analysis illustrate that when T is reasonably large CL performs well. However, due to the presence of nuisance parameters, CL is subject to the "incidental parameter" problem for small T.

Keywords: ARCH models; composite likelihood; nuisance parameters; panel data.

1 Introduction

This study focuses on the application of the composite likelihood (CL) method to GARCH panels. A GARCH panel is a collection of financial time series that are characterised by time-varying volatility. The defining feature of a GARCH panel is that, while nuisance parameters are allowed to vary across series, other parameters of interest are assumed to be common for all series.

The origins of the composite likelihood method go back to at least Lindsay (1988). See Varin (2008) and Varin, Reid, and Firth (2009) for a review. The method has recently been introduced to financial econometrics by Engle, Shephard, and Sheppard (2008) as a basis for pooling information across not only time, but also cross-section. In GARCH panels, this amounts to estimating the parameters of interest for all assets simultaneously, instead of individually. This is important since the common quasi-maximum likelihood estimator (QMLE) for the GARCH model delivers poor results in samples of a few hundred observations. This study illustrates that CL is capable of delivering satisfactory results in such samples by pooling information across series, though it too suffers from error introduced by nuisance parameter estimates. This "incidental parameter" problem has been mentioned in the financial econometrics literature by Engle and Sheppard (2001), Engle, Shephard, and Sheppard (2008), and Engle (2009).

An important point in favour of CL is that QMLE-based estimation of GARCH, while satisfactory in samples with thousands of observations, is unreliable in small samples. For example, using a sample of 100 or 250 observations, the fitted GARCH is unlikely to adequately model the conditional heteroskedasticity in data. On the other hand, CL is potentially able to produce a reasonable conditional heteroskedasticity structure, even when the number of observations is very small, since it uses information contained in the whole panel. Although assets in the panel are correlated to some degree, it is implausible that all assets are perfectly correlated. Hence, a panel of asset prices contains at least as much information as a single asset does.

Forecasters often have to use a short time series or a small-T panel. A recent structural break in data is one cause. Assuming the break occurred a year ago (corresponding to the availability of around 250 daily observations following the break), parameters estimated using QMLE most likely suffer from substantial bias which, in turn, leads to poor forecasting performance. On the other hand, CL has the potential to work well in this scenario. Another application where CL can be useful is monthly hedge fund data, which consists of monthly returns on thousands of funds and hence, is a short, wide panel.

The relevant large sample theory underlying the method used here has already been developed by Engle, Shephard, and Sheppard (2008), who looked at large dimensional time-varying covariances. They employed CL to produce a computationally feasible estimator, with the CL constructed by averaging the log-likelihoods for submodels built using bivariate time series. Our study develops the GARCH panel structure using the theoretical foundations provided by Engle, Shephard, and Sheppard (2008), and employs Monte Carlo and empirical analysis to examine its properties.

Our Monte Carlo simulations demonstrate that CL is capable of modelling conditional heteroskedasticity correctly using previously infeasible sample sizes. Furthermore, forecast comparisons using stock-market data from S&P100 reveal that, even when the parameters of interest are likely to vary across the panel, CL performs well against QMLE in small-T panels. Nevertheless, as the sample size increases, information pooling loses its attractiveness, as QMLE performs well enough in long time series, and is robust to wrongly pooling information between series.

The structure of the paper is as follows. In Section 2 we introduce the GARCH panel model and the analysis by composite likelihood. In Section 3 we report results from various simulation experiments. Section 4 then provides an empirical illustration of these methods, and Section 5 draws some conclusions.

2 The GARCH panel

GARCH models are frequently used in financial econometrics. Reviews of the literature include Bollerslev, Engle, and Nelson (1994), Bauwens, Laurent, and Rombouts (2006), and Silvennoinen and Teräsvirta (2009). The focus in this paper is on a GARCH panel. The $(T \times N)$ GARCH panel is a collection of N financial time series that are assumed to have GARCH dynamics and to share a common set of parameters, $\theta = (\alpha, \beta)$, while the nuisance parameters, $\{\gamma_i\}_{i=1}^N$, are allowed to be asset-dependent (in the rest of the paper we will use $\{\cdot_i\}$ as a shorthand for $\{\cdot_i\}_{i=1}^N$). Our focus in on fitting a very large number of univariate GARCH models; for example, this would be needed for the first step of fitting a Dynamic Conditional Correlation model by Engle (2002). For simplicity of exposition we assume each time series is of length T, although in practice this is of course not necessary.

Formally, we have a panel of asset returns with T observations for each of the N_T assets. Throughout, it is assumed that the number of series in the cross-section can potentially increase with the number of observations and so N_T has the subscript T. This includes cases where there are more assets than time-series observations. Moreover, asset returns are assumed to display conditional heteroskedasticity over time and cross-sectional dependence, where y_{it} is the return on asset i at time t, $i = 1, ..., N_T$ and $t = 1, ..., T$. We write,

$$\begin{aligned} y_{it} &= \mu_{it} + \varepsilon_{it}, \qquad \mu_{it} = \mathrm{E}[y_{it}|\mathcal{F}_{t-1}], & (1) \\ \mathrm{E}\left[\varepsilon_{it}|\mathcal{F}_{t-1}\right] &= 0 \quad \text{and} \quad \mathrm{Var}\left[y_{it}|\mathcal{F}_{t-1}\right] = \mathrm{Var}\left[\varepsilon_{it}|\mathcal{F}_{t-1}\right] \equiv \sigma_{it}^2, & (2) \end{aligned}$$

where $\mathcal{F}_{t-1}$ is the historical information set at time $t-1$. As the analysis focuses on conditional variance, it is assumed that $\mu_{it} = 0$. The GARCH panel is based on the GARCH(1,1) specification given by

$$\sigma_{it}^2 = \gamma_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \quad \text{where} \quad \gamma_i > 0, \quad \alpha, \beta \in [0,1), \quad \alpha+\beta < 1. \qquad (3)$$

Here, α and β constitute the parameters of interest, while $\{\gamma_i\}$ are treated as nuisance parameters that are not of direct interest but, nevertheless, have to be estimated in order to obtain $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}\right)'$. It can be shown that this specification, often called variance-tracking, implies that

$$\mathrm{E}(y_{it}^2) = \gamma_i, \tag{4}$$

enabling the use of method of moments (MM) to estimate γ_i. Here we make the ad-hoc choice of setting $\sigma_{i0}^2 = \lfloor T^{-1/2} \rfloor \sum_{t=1}^{\lfloor T^{-1/2} \rfloor} y_{it}^2$. Finally $\{\gamma_i^*\}$, α^*, and β^* are defined as the true parameter values for $\{\gamma_i\}$, α and β, respectively, while $\gamma_{(N_T)}^* \equiv \left(\gamma_1^*, ..., \gamma_{N_T}^*\right)$.

This panel structure has many similarities with the autoregressive panels commonly used in economics and statistics. Reviews of that literature include Arellano and Honore (2001) and Diggle, Liang, and Zeger (1994). We know of only Engle and Mezrich (1996) and Bauwens and Rombouts (2007) as previous studies on GARCH panels.

Conventionally, estimation of θ can be conducted individually for each asset, using QMLE. However, this only utilises information available in a single time series. What is preferable in this situation (where all assets share a common θ) is to estimate θ by pooling all information available in the panel. This is made possible by CL.

2.1 Estimation procedure

Let $f(y_{it}|\mathcal{F}_{t-1}; \theta, \gamma_i)$ be the conditional density for y_{it}. The joint density specification for all asset returns at time t is given by $f(y_{1t}, ..., y_{N_T t}|\mathcal{F}_{t-1})$, which we do not model, noting that knowledge of all of the N_T submodels does not deliver knowledge of $f(y_{1t}, ..., y_{N_T t}|\mathcal{F}_{t-1})$ (the conditional copula is entirely unspecified) unless the individual components are conditionally independent.

This model is indexed by some common parameters θ and individual effects γ_i. This type of assumption appeared first in the influential work of Neyman and Scott (1948). Recent papers on the analysis of this setup include Barndorff-Nielsen (1996), Lancaster (2000), and Sartori (2003). In those papers, stochastic independence is assumed over i and t. Then the maximum likelihood estimator of θ is typically inconsistent for finite T as $N \to \infty$ and needs, when T increases, $N = o(T^{1/2})$ for standard distributional results to hold with rate of convergence $\sqrt{NT}$ (see Sartori (2003)). In our time series situation we are content to allow T to be large, while the important cross-sectional dependence implied by CL amongst the individual quasi likelihoods reduces the rate of convergence to $\sqrt{T}$, not $\sqrt{NT}$. Under those circumstances the m-composite likelihood estimator is consistent and has a simple limit theory however N relates to T (see Engle, Shephard, and Sheppard (2008) for details). In our framework we have both time-series and cross-sectional dependence in the $y_{it}|\mathcal{F}_{t-1}$.

Define $\psi_i \equiv \left(\theta', \gamma_i\right)'$ and $\psi_{(N_T)} = (\theta', \gamma'_{(N_T)})'$. Then, the normal-density composite likelihood function is given by

$$\begin{aligned} CL(\psi_{(N_T)}; y) &= \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{N_T}\sum_{i=1}^{N_T}\log f(y_{it}|\mathcal{F}_{t-1};\psi_i)\right\} = \frac{1}{T} l(\psi_{(N_T)}; y), \quad \text{where} \quad (5) \\ l(\psi_{(N_T)}; y) &= \sum_{t=1}^{T} l_t(\psi_{(N_T)}; y_t|\mathcal{F}_{t-1}), \quad \text{and} \\ l_t(\psi_{(N_T)}; y_t|\mathcal{F}_{t-1}) &= \sum_{i=1}^{N_T}\log f(y_{it}|\mathcal{F}_{t-1};\psi_i). \end{aligned}$$

Estimation of $\psi_{(N_T)}$ is based on a two-step estimation procedure. First, $\{\gamma_i\}$ are estimated using method of moments estimation based on (4) to obtain $\{\hat{\gamma}_i\}$. Then, these are substituted for $\{\gamma_i\}$ in (3), and θ is estimated using (5). A detailed exposition of the theory for two-step estimation is provided by Newey and McFadden (1994). There N_T is fixed so despite similarities in estimation approach, standard results do not apply to the current case.

Formally, using (4),

$$m_{N_T}(y_t, \gamma_{(N_T)}) = \begin{pmatrix} y_{1t}^2 - \gamma_1 \\ \vdots \\ y_{N_T t}^2 - \gamma_{N_T} \end{pmatrix}, \text{ implying } E(m_{N_T}(y_t, \gamma^*_{(N_T)})) = 0. \quad (6)$$

Equation (6) gives the population moment condition for the nuisance parameters. For θ, the score function for the normal-density composite-likelihood function is

$$g(y_t, \theta, \gamma_{(N_T)}) = \frac{\partial}{\partial\theta}\frac{1}{N_T}\left(-\frac{1}{2}\sum_{i=1}^{N_T}\log\sigma_{it}^2 - \frac{1}{2}\sum_{i=1}^{N_T}\frac{\varepsilon_{it}^2}{\sigma_{it}^2}\right). \quad (7)$$

For (6) and (7), respective sample moment conditions are given by

$$\frac{1}{T}\sum_{t=1}^{T} m_{N_T}(y_t, \hat{\gamma}_{(N_T)}) = 0, \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} g(y_t, \hat{\gamma}_{(N_T)}, \hat{\theta}) = 0, \quad (8)$$

where $\hat{\gamma}_{(N_T)}$ and $\hat{\theta}$ are appropriate estimators for $\gamma^*_{(N_T)}$ and $\theta^* \equiv (\alpha^*, \beta^*)$. Stacking (6) and (7), the population and sample moment conditions are given by

$$\mathrm{E}\left[\tilde{g}(y_t, \theta^*, \gamma^*_{(N_T)})\right] = \mathrm{E}\begin{bmatrix} m_{N_T}(y_t, \gamma^*_{(N_T)}) \\ g(y_t, \theta^*, \gamma^*_{(N_T)}) \end{bmatrix} = 0, \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T}\tilde{g}_{t,T}(y_t, \hat{\theta}, \hat{\gamma}_{(N_T)}) = \underset{([N_T+2]\times 1)}{0.}$$

We note that (8) is the first order condition for the simple optimization problem

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_T}\sum_{i=1}^{N_T}\log f(y_{it}|\mathcal{F}_{t-1}, \theta, \hat{\gamma}_i). \quad (9)$$

Equation (9) is based on an *m-profile* composite likelihood function, formed by ignoring the potential dependence in the data across individuals; it is an m-profile version as we have plugged the moment based estimator of γ_i into the composite likelihood. This provides a statistically inefficient estimator for θ as it ignores dependence over individuals, employs a moment based estimator to remove γ_i, and the submodels for $y_{it}|\mathcal{F}_{t-1}$ may really be just quasi-likelihoods and not true likelihoods.

In this setting, there are N_T moment conditions coming from the nuisance parameters and two moment conditions coming from the score vector. An important observation is that for each asset in the panel, there is a nuisance parameter estimation.

2.2 Large sample distribution

If we ignore the estimation of the nuisance parameters, then this is just a time-series extension of the analysis of Cox and Reid (2004). In that case, the score for the t-th observation is given by

$$s_{t,N} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \log f\left(y_{it}|\mathcal{F}_{t-1,i};\theta\right)}{\partial\theta},$$

which is a triangular array martingale difference sequence. We assume that it obeys a central limit theorem,

$$\frac{1}{T}\sqrt{T}\sum_{t=1}^{T} s_{t,N} \xrightarrow{d} N(0,\mathcal{I}), \quad \text{where} \quad \mathcal{I} = p\lim\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{Var}\left(s_{t,N}|\mathcal{F}_{t-1,N}\right)\right].$$

Here N can increase with T, but we assume that $\mathcal{I}$ is positive definite. The latter assumption is not trivial; for example, it would not be expected if the data are i.i.d. in the cross section. More formally, we assume that if N increases the cross sectional average $s_{t,N}$ does not obey a law of large numbers.

Based on the normal limit, it follows that

$$\sqrt{T}\left(\widehat{\theta}-\theta\right) \xrightarrow{d} N(0,\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}), \tag{10}$$

where

$$\mathcal{J} = p\lim\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\frac{\partial s_{t,N}}{\partial\theta'}\Big|\mathcal{F}_{t-1,N}\right],$$

assuming that $\mathcal{J} > 0$. Notice that $\mathcal{J}$ is approximately the average of Hessians of a randomly chosen submodel at a random time $\partial^2 \log f(y_{it};\psi)/\partial\theta\partial\theta'$, and that the CLT is only for $\widehat{\theta}$, it makes no statement about the γ_i. To account for the nuisance parameters, a modified estimator for the score covariance is required:

$$z_{t,N} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\partial \log f\left(y_{it}|\mathcal{F}_{t-1,i};\psi\right)}{\partial\theta} + \left[\sum_{t=1}^{T}\frac{\partial^2 \log f\left(y_{it}|\mathcal{F}_{t-1,i};\psi\right)}{\partial\theta\,\partial\gamma_i}\right]\left(y_{it}^2-\gamma_i\right)\right\},$$

$$\frac{1}{T}\sqrt{T}\sum_{t=1}^{T} z_{t,N} \xrightarrow{d} N(0,\tilde{\mathcal{I}}), \quad \text{where} \quad \tilde{\mathcal{I}} = p\lim\left[\frac{1}{T}\sum_{t=1}^{T}\text{Var}\left(z_{t,N}|\mathcal{F}_{t-1,N}\right)\right].$$

Here, $z_{t,N}$ is different from $s_{t,N}$ in that it contains a correction term given by

$$\left[\sum_{t=1}^{T}\frac{\partial^2 \log f\left(y_{it}|\mathcal{F}_{t-1,i};\psi\right)}{\partial\theta\,\partial\gamma_i}\right]\left(y_{it}^2-\gamma_i\right).$$

The first term accounts for the influence of estimating γ_i on estimating θ. As such, if θ and γ_i are orthogonal, then there is no such influence and the correction term disappears. The second term can be related to estimation of γ_i by the method of moments. If the data provide an accurate estimate of γ_i, then this term is small, making the correction term small as well. The correction term may also be small, even when the data yield a very inaccurate estimate of γ_i, if γ_i and θ are nearly orthogonal.

An important point of (10) is that the rate of convergence of the estimator is not improved by having a cross-section. Instead the cross-section influences the size of $\mathcal{I}$, but its impact is limited. For a more detailed exposition of the related large sample theory, see Engle, Shephard, and Sheppard (2008).

In practice, to make inference we need estimators for $\tilde{\mathcal{I}}$ and $\mathcal{J}$. An estimator for $\mathcal{J}$ can be obtained by evaluating the Hessian at sample observations. $\tilde{\mathcal{I}}$ on the other hand requires the use of a HAC estimator. Examples of such estimators are provided by Newey and West (1987) and Andrews (1991).

3 Simulation analysis

3.1 The setting

The asset panel was generated using the specification described in (1)-(3). For most stock returns annual volatility is in the range 15% and 60%, so we took $\gamma_i \overset{iid}{\sim} U\left[0.02, 0.05\right]$. This is suggested by $\sigma_D = \sqrt{\sigma_A/252}$, where σ_D and σ_A are daily and annual volatility, respectively. For an annual volatility of 15%, daily volatility according to this method is 0.0244, while for 60% the daily volatility is 0.0488. For each series the γ_i were used as the initial values for the conditional variances, $h_{i,0}$. Cross-sectional dependence was generated by a single-factor model where

$$\varepsilon_{it} = \rho_i u_t + \sqrt{1-\rho_i^2}\tau_{it}, \quad \tau_{it}, u_t \overset{iid}{\sim} N(0,1), \tag{11}$$

implying

$$\mathrm{E}\left(\varepsilon_{it}|\rho_i\right)=0, \quad \mathrm{Var}\left[\left.\begin{pmatrix}\varepsilon_{it}\\ \varepsilon_{jt}\end{pmatrix}\right|\rho_i,\rho_j\right]=\begin{bmatrix}1 & \rho_i\rho_j\\ \rho_i\rho_j & 1\end{bmatrix} \ \forall\ i\neq j \text{ and } \forall t,$$

and $\mathrm{Cov}(\varepsilon_{it},\varepsilon_{js}|\rho_i,\rho_j)=0$ for all $t\neq s$ and all i,j.

The choice of the ρ_i in a way that ensures neither perfect correlation nor independence can be done in various ways. A restrictive option is to assume that the ρ_i are equal. Engle, Shephard, and Sheppard (2008) considered a truncated normal distribution to generate the ρ_i, where truncation occurs at 0.1 and 0.9. Our study used $\rho_i\sim U\,[0.5,0.9]$ for all i, ensuring that the lowest and highest correlation between two assets were 0.25 and 0.81, respectively. Lastly, α and β were chosen from three alternatives that cover the range of parameter values found in asset data:

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\in\left\{\theta^{(1)},\theta^{(2)},\theta^{(3)}\right\}=\left\{\begin{bmatrix}0.02\\ 0.97\end{bmatrix},\begin{bmatrix}0.05\\ 0.93\end{bmatrix},\begin{bmatrix}0.10\\ 0.80\end{bmatrix}\right\}. \tag{12}$$

3.2 The results

All results are based on 2,500 replications. Average biases of estimates and their Monte Carlo standard deviations (MCSD) are amongst obvious criteria for comparison. To investigate whether the theoretical large sample properties of CL hold in finite samples, asymptotic standard deviation (ASD) and root mean squared error (RMSE) statistics are also provided:

$$\begin{aligned} MCSD \ &: \ \bar{\sigma}_{\hat{\kappa}}=\sqrt{\frac{1}{Z}\sum\nolimits_{z=1}^{Z}\left(\hat{\kappa}_z-\frac{1}{Z}\sum\nolimits_{z=1}^{Z}\hat{\kappa}_z\right)^2},\\ ASD \ &: \ \hat{\sigma}_{\hat{\kappa}}=\frac{1}{Z}\sqrt{\sum\nolimits_{z=1}^{Z}\hat{\sigma}^2_{\hat{\kappa},z}},\\ RMSE \ &: \ \mathcal{R}_{\hat{\kappa}}=\sqrt{\frac{1}{Z}\sum\nolimits_{z=1}^{Z}\left(\hat{\kappa}_z-\kappa\right)^2}, \end{aligned}$$

where Z is the number of replications, $\hat{\alpha}_z$ and $\hat{\beta}_z$ are the estimates for replication z, $z=1,...,Z$, and $\hat{\kappa}_z\in\{\hat{\alpha}_z,\hat{\beta}_z\}$. $\hat{\sigma}^2_{\hat{\kappa},z}$ is the estimated asymptotic variance for $\hat{\kappa}_z$. ASD serves as an average measure of the asymptotic standard deviation across all replications. In addition, coverage rates of sample confidence interval statistics (CI) are provided as a further measure of the finite sample performance of the asymptotic distribution for the CL based upon $\hat{\sigma}^2_{\hat{\kappa},z}$. All results are calculated for 95% confidence intervals.

Tables 1 and 2 present the results for the three parameter values in (12), where $T=2,000$. Tables 3 and 4 look at the implications of varying T where $T\in\{100,250,500,1000,2000\}$ (this second analysis is conducted for $\theta^{(2)}$ only, due to space restrictions). In all cases, results for $N_T=1$ are also provided, which corresponds to using QMLE instead of CL.

Simulation results presented in Table 1 show that when $T=2,000$, CL generally leads to

$\theta =$	$(0.02, 0.97)$		$(0.05, 0.93)$		$(0.10, 0.80)$		$(0.02, 0.97)$		$(0.05, 0.93)$		$(0.10, 0.80)$	
	Bias						MCSD					
N_T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
1	3.3%	-.81%	.15%	-.49%	.16%	-1.2%	.007	.019	.010	.017	.020	.050
10	.42%	-.29%	-.08%	-.21%	-.40%	-.29%	.002	.005	.004	.006	.008	.017
50	.17%	-.26%	-.12%	-.19%	-.32%	-.28%	.002	.003	.003	.005	.006	.013
100	.08%	-.25%	-.19%	-.18%	-.29%	-.27%	.002	.003	.003	.004	.005	.012

Table 1: Monte Carlo simulation results for fixed T using the three parameter sets given in (12): average biases for $\hat{\alpha}$ and $\hat{\beta}$ in percentages and Monte Carlo standard deviations ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$). $T = 2,000$ in all cases, while N_T gives the number of series in the cross-section. Based on $2,500$ replications.

N_T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
			$\alpha = 0.02$	$\beta = 0.97$				
1	.007	.019	.030	.107	.007	.021	.922	.928
10	.002	.005	.002	.005	.002	.005	.936	.948
50	.002	.003	.002	.003	.002	.004	.938	.937
100	.002	.003	.002	.003	.002	.004	.943	.941
			$\alpha = 0.05$	$\beta = 0.93$				
1	.010	.017	1.71	12.9	.010	.018	.933	.945
10	.004	.006	.004	.006	.004	.007	.938	.952
50	.003	.005	.003	.005	.003	.005	.941	.937
100	.003	.004	.003	.004	.003	.005	.946	.943
			$\alpha = 0.10$	$\beta = 0.80$				
1	.020	.050	.020	.052	.020	.051	.933	.924
10	.008	.017	.008	.017	.008	.018	.954	.944
50	.006	.013	.006	.013	.006	.013	.968	.954
100	.005	.012	.006	.012	.005	.012	.970	.954

Table 2: Monte Carlo simulation results: Monte Carlo standard deviation ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$), asymptotic standard deviation ($\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$), root mean squared error ($\mathcal{R}_{\hat{\alpha}}$ and $\mathcal{R}_{\hat{\beta}}$) and sample confidence interval (CI) statistics. $T = 2,000$ in all cases, while N_T gives the number of series in the cross-section. Based on $2,500$ replications.

	N_T= 1 $(QMLE)$		N_T= 10		N_T= 50		N_T= 100	
				Bias				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
100	1.41%	-15.9%	-23.6%	-15.2%	-26.7%	-16.2%	-27.2%	-16.5%
250	11.3%	-10.6%	-1.38%	-3.50%	-3.46%	-2.71%	-3.61%	-2.61%
500	5.34%	-4.57%	-.157%	-1.15%	-.650%	-.976%	-.695%	-.955%
1,000	2.37%	-1.53%	-.004%	-.484%	-.221%	-.406%	-.306%	-.393%
2,000	.534%	-.561%	-.076%	-.225%	-.050%	-.185%	-.082%	-.181%
				MCSD				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
100	.069	.227	.030	.169	.023	.163	.022	.160
250	.043	.186	.014	.052	.010	.028	.010	.021
500	.026	.109	.009	.016	.006	.011	.006	.010
1,000	.016	.042	.006	.010	.004	.007	.004	.007
2,000	.010	.019	.004	.006	.003	.004	.003	.004

Table 3: Monte Carlo simulation results for $\theta = (0.05, 0.93)$: average biases for $\hat{\alpha}$ and $\hat{\beta}$ in percentages and Monte Carlo standard deviations ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$). T and N_T give the number of observations in each time series and the number of series in the cross-section, respectively. Based on $2,500$ replications.

low average bias across all parameter values, with the highest average bias being .42% for $\hat{\alpha}$ and $-.29\%$ for $\hat{\beta}$. In contrast, the average bias due to QMLE reaches levels as high as 3.3% for $\hat{\alpha}$ and -1.2% for $\hat{\beta}$. An interesting observation is that, when $\hat{\beta}$ is concerned, there is a general tendency for the average bias to initially decrease and then plateau as N_T increases. For example, for $\theta^{(1)} = (0.02, 0.97)$, the change in bias when N_T increases from 50 to 100 is .01%. Moreover, taking $N_T = 1$ as a reference, when the panel size increases to $N_T = 10$ the change in bias is .52%, while when the size is increased to $N_T = 100$, bias is reduced by .56%. This shows that the speed of decline falls with N_T. These results also suggest that there are substantial gains in shifting from time series (QMLE) to a panel (CL) structure, in terms of both the average bias and sample standard deviation.

Sample standard deviations (MCSD) are also generally low and decrease with N_T. This is not surprising as an increase in N_T implies that there is more information to use. Moreover, the decrease in MCSDs is not large enough to imply that the speed of convergence in finite samples is $\sqrt{TN_T}$ as opposed to $\sqrt{T}$. Similar to the previous discussion for average bias, sample standard deviations exhibit a pattern of convergence to some non-zero limit. Therefore, increasing N_T beyond 100 does not lead to substantial decreases in MCSD. These results are all in accordance with the asymptotic theory in Engle, Shephard, and Sheppard (2008).

Table 2 presents further results for the same simulation exercise. The MCSD and ASD statistics for both $\hat{\alpha}$ and $\hat{\beta}$ are generally very close to each other, implying that the simulation results are in line with the relevant asymptotic theory. The RMSE statistics confirm the earlier observation of a

T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
N_T= 1 ($QMLE$)									N_T= 10							
100	.069	.227	.111	2.25	.069	.272	.551	.863	.030	.169	.072	1.67	.032	.220	.791	.832
250	.043	.186	.187	7.75	.043	.210	.842	.866	.014	.052	.017	.129	.014	.061	.911	.920
500	.026	.109	.028	.143	.026	.117	.889	.905	.009	.016	.009	.018	.009	.019	.931	.937
1,000	.016	.042	.018	.057	.016	.044	.909	.916	.006	.010	.006	.010	.006	.011	.934	.945
2,000	.010	.019	.011	.021	.010	.019	.923	.932	.004	.006	.004	.006	.004	.007	.940	.948
N_T= 50									N_T= 100							
100	.023	.163	.283	5.15	.026	.222	.786	.759	.022	.160	.078	5.81	.026	.222	.769	.718
250	.010	.028	.010	.026	.010	.037	.912	.897	.010	.021	.010	.026	.010	.032	.908	.877
500	.006	.011	.006	.011	.006	.014	.930	.926	.006	.010	.006	.011	.006	.014	.934	.926
1,000	.004	.007	.004	.007	.004	.008	.933	.936	.004	.007	.004	.007	.004	.007	.932	.938
2,000	.003	.004	.003	.005	.003	.005	.942	.952	.003	.004	.003	.004	.003	.004	.947	.950

Table 4: Monte Carlo simulation results for $\theta = (0.05, 0.93)$: Monte Carlo standard deviation ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$), asymptotic standard deviation ($\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$), root mean squared error (RMSE) and sample confidence interval (CI) statistics. T and N_T give the number of observations in each time series and the number of series in the cross-section, respectively. Based on $2,500$ replications.

non-vanishing bias as $N_T \to \infty$, since in some cases there is a slight difference between MCSD and RMSE that suggests some very small bias. As for QMLE, although MCSD and RMSE values are close to each other, ASD is very high for $\theta^{(1)}$ and, especially, $\theta^{(2)}$. This is another point in favour of using the panel structure instead of focusing on the series individually. Also, CI statistics are very satisfactory, ranging between 92% and 97% across all cases.

Now, we turn to the implications of varying both the number of assets and the observations per asset by using $N_T \in \{1, 10, 50, 100\}$ and $T \in \{100, 250, 500, 1000, 2000\}$. Table 3 shows that average bias decreases with T. This is not unanticipated as fitted GARCH usually models the conditional heteroskedasticity dynamics much better when longer time series are used. Unsurprisingly, both $\bar{\sigma}_\alpha$ and $\bar{\sigma}_\beta$ decrease with T. Clearly, having a larger number of observations for each series delivers less biased and more efficient estimators.

Table 4 reveals that CL performs well when there are around at least 500 observations in the time series. However, in the remaining cases the large sample theory gives poor finite sample results, as reflected in the discrepancy between MCSD and ASD. The sample confidence interval statistics agree with these results. As T decreases, sample confidence intervals move further away from 95% and become more conservative. Similarly, the discrepancy between the RMSE and MCSD statistics, especially for $\hat{\beta}$, increases as T decreases, pointing to a negative correlation between average bias and sample size.

Comparing CL to QMLE, QMLE's relative performance is very poor, especially when average bias is concerned (except when $T = 100$, which is due to the optimisation routine's sensitivity to the starting values of the algorithm). Clearly, CL is preferable to QMLE, in the hypothetical situation that all series share a common set of parameters of interest. The general message of the simulation results so far is that CL performs well when $T \geq 500$. The reason for CL's biases in small-T panels is discussed next.

3.3 Nuisance parameters and estimation error

As stressed previously, CL pools all information available in the panel to form a single likelihood function. Therefore, one would intuitively expect CL to be successful even when T is small but there are indications of significant bias when $T < 250$. Is this caused by the estimation of the γ_i for each model?

Figure 1 presents sampling distributions of the estimators of θ using (i) the method of moments estimator for the nuisance parameter (CL1), and (ii) the true value of the nuisance parameter (CL2) which corresponds to infeasible estimation of the nuisance parameter. The sample distribution graphs reveal why CL performs worse when T is very small: sample distributions are not centered around α and β, and there is high dispersion. Some improvement can be observed as T increases to 250. However, $\hat{\beta}$ is prone to exhibit some mild bias even when T is high. In accordance with observations in the previous simulation study, while average bias decreases with T, an increase in N_T (for a given T) leads to higher precision. However, here, high precision is not always a desirable property. In a slightly counter-intuitive way, although higher N_T increases estimator precision, it also make a biased estimator more precise, causing more harm than good. As such, having a larger number of assets is very useful when T is very large, ensuring that the estimator is both unbiased and more efficient in the sense of having a smaller asymptotic standard deviation. It must be noted that increasing N_T beyond a certain number of assets does not lead to any improvement in efficiency.

Looking at the sample distributions of estimators without nuisance parameter estimation (CL2), it is encouraging that for both $\hat{\alpha}$ and $\hat{\beta}$ the peak of the sample distributions is always either on or very close to the real parameter value, even when $T = 100$. Similar to the previous simulations, larger T decreases bias while larger N_T leads to higher precision. Clearly, nuisance parameter estimation undermines the statistical properties of the GARCH panel model greatly when T is small. As suggested by the Associate Editor, using empirical Bayes methods in order to improve the estimation of these parameters could be beneficial, given the simulation results (see, for example, Lindsay (1985) and Liang and Tsou (1992)). Dealing with the incidental parameter problem is already the subject of another ongoing project and therefore, we do not focus on this issue further in this study.

It is also interesting that when QMLE is used ($N_T = 1$), even when the true nuisance parameter is known, the estimators still perform poorly. While nuisance parameter estimation leads to a significant bias, using true nuisance parameter causes very high dispersion. However, remembering that the real issue with QMLE is that T is too small to adequately model conditional volatility, it is obvious that knowledge of the true value of the nuisance parameter does not help.

4 Empirical analysis

In this section, in addition to CL and QMLE, we consider the MacGyver (MG) method introduced by Engle (2009). This is another information pooling method based on "blending" already available estimates of a parameter to obtain a new estimate of that parameter.

Let $\{\hat{\theta}_k\}_{k=1}^K$ be K different estimates of θ. These may be obtained by using different methods, models, or data sets. For the case at hand, N_T estimates of θ can be obtained by employing QMLE for each asset in the panel individually. These estimates are then combined using a "blend function", $b(\cdot)$, to obtain a final estimate of θ, $\hat{\theta}_{MG} = b\left(\{\hat{\theta}_k\}_{k=1}^K\right)$. Engle (2009) suggests that three obvious blend functions are the mean, median, and the mean of a trimmed set when the highest and lowest 5% of the estimates are eliminated. The latter two blending functions serve the purpose of discarding outliers that could otherwise introduce bias.

For the GARCH panel, $\hat{\theta}_i$ is estimated using two-step estimation: in the first step, $\hat{\gamma}_i$ is obtained in the same way as for CL; in the second step, $\hat{\theta}_i$ is estimated using

$$\hat{\theta}_i = \underset{\theta \in \Theta}{\text{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log f(y_{it}|\mathcal{F}_{t-1}; \theta, \hat{\gamma}_i), \qquad i = 1, ..., N_T.$$

It must be noted that there are several practical issues related to this method. First, when the sample size is not large enough, optimisation may fail and simply yield the initial values used for optimisation as the parameter estimates (the optimisation procedure used for this study starts at pre-specified starting values and searches for an optimum. If optimisation fails to find an optimum, then the starting values are given as the parameter estimates). Following Engle (2009), such cases are discarded and not used in the blend function. Furthermore, when using the GARCH specification, if $\hat{\alpha}$ is equal to zero, then $\hat{\beta}$ is not identified and has no interpretation, no matter what its value. Consequently, this analysis also ignores sets of estimates where $\hat{\alpha}$ is less than 0.0025. These issues do not occur rarely. In a simulation analysis not presented here, for 2,500 replications of a GARCH process with 100 observations in each replication, in more than 1,400 replications estimators failed to converge while around 100 replications produced $\hat{\alpha} = 0$. This particular choice of the cut-off value and the elimination of non-converging cases reflect the ad-hoc nature of MG. Nevertheless, the aim of MG is not to have a set of very good estimates, but rather to find a blend function that yields a good estimate out of a large pool of estimates.

Considering that both CL and MG are based on "pooling" information, an obvious comparison of interest is that of CL against MG. MG can be considered as a step between CL and QMLE: similar to CL, it is based on pooling information, while estimation essentially employs QMLE and not CL.

Another intriguing analysis is the comparison of the information pooling methods to QMLE, as the assumption that all assets share a common set of parameters of interest is not necessary for QMLE. CL and MG, on the other hand, crucially rely on this assumption that is likely to be violated. As far as empirical performance is concerned, what is also relevant is whether the gains from using CL and MG are worth making this restrictive assumption, even when there may be no apparent reason for it to hold.

Several points have to be mentioned. First of all, neither information pooling method is likely to explain the data perfectly, even in large-T panels. To start with, there is no guarantee that some or all of the data follows a GARCH process, although this model has been found to be very successful in practice. Moreover, the assumption of a common set of parameters for all assets is not likely to hold. Be that as it may, the question remains: despite these issues, can the CL and MG methods attain better forecasting performance through their data-pooling mechanism?

In light of these points, the questions of interest are whether pooling information in an asset panel can improve forecasting performance in samples of any size, and whether CL can have an advantage over the other methods, especially in small-T samples where QMLE is expected to perform poorly. The analysis is conducted using stock-market data from S&P100. A recent procedure due to Giacomini and White (2006) that allows comparison of different methods (such as the CL and MG methods), as opposed to different models (such as the GARCH and TARCH models), is used to test equal predictive ability and choose between methods.

4.1 Methodology

In the analysis, two competing τ-period ahead forecasts obtained at time t, $\hat{Y}_{1,t+\tau}$ and $\hat{Y}_{2,t+\tau}$, for a variable of interest, $Y_{t+\tau}$, are under scrutiny. Accuracy of forecasts are measured using loss functions. "Loss", in the forecast comparison sense, occurs due to the distance between the forecast and the true value of the variable of interest. Formally, the loss due to $\hat{Y}_{t+\tau}$ is defined as

$$L_{t+\tau}\left(Y_{t+\tau}, \hat{Y}_{t+\tau}\right). \tag{13}$$

Examples of loss functions used in the literature are many. See Patton (2008) for a more detailed study of implications of using different loss functions. A prominent example, used here, is the loss function

$$QLIKE: L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}) = \log \hat{Y}_{t+\tau} + \frac{Y_{t+\tau}}{\hat{Y}_{t+\tau}}.$$

A suitable testing framework is due to Giacomini and White (2006) (GW). Unlike the widely used Diebold-Mariano-West (DMW) framework due to Diebold and Mariano (1995) and West (1996), the GW test allows for the comparison of two different methods as opposed to two different models. The Null Hypothesis is

$$H_0 : \mathrm{E}\left[L_{t+\tau}(Y_{t+\tau}, f_t(\hat{\beta}_{1t})) - L_{t+\tau}(Y_{t+\tau}, g_t(\hat{\beta}_{2t}))|\mathcal{G}_t\right] = 0, \tag{14}$$

where $\mathcal{G}_t$ is an information set at time t and $f_t(\cdot)$ and $g_t(\cdot)$ are two (not necessarily different) forecasting models. $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$ are estimates of parameters of interest obtained by using two different methods. As evident in (14), the GW test allows for a conditional, as well as an unconditional approach. The latter compares the average performance of two forecasting methods while the former analyses whether past information can be used to predict which method will provide a better forecast for a particular date.

An important feature of volatility is that it is a latent variable and is never observed, even ex-post. Therefore, a proxy should be used for forecast comparison. In this study the squared return, y_{it}^2, is used as proxy, which is a common choice. It must however be noted that there is now a growing literature suggesting that squared returns may lead to a wrong ranking of forecasts. Instead, realised volatility is recommended as a better proxy. Very briefly, realised volatility is the sum of squared high-frequency intra-daily returns. It was formalised from an econometric viewpoint by Andersen, Bollerslev, Diebold, and Labys (2001) and Barndorff-Nielsen and Shephard (2002). See Andersen and Benzoni (2009) for a recent survey. This is important for the choice of the loss function. Hansen and Lunde (2006) show that the use of noisy proxies such as y_{it}^2 may lead to inconsistent ranking of volatility models, whereby the empirical ranking may not be the same as the true ranking. Patton (2008) extends this analysis and focuses on loss functions that are robust to the choice of the volatility proxy, in the sense that the empirical ranking implied by those loss functions are the same independent of which proxy is used. He provides a family of homogeneous and robust loss functions that contains QLIKE, as well. Furthermore, Patton and Sheppard (2009) provide a Monte Carlo analysis to compare the power of different loss functions from this family under the DMW framework using realised volatility as the proxy. Their results indicate that the QLIKE function has the best power performance. Motivated by these results, we employ QLIKE only, due to space restrictions.

4.2 Empirical results

The empirical analysis is based on the daily returns for 94 stocks from S&P100 for the period between 1 April 2000 to 12 January 2008. Data for six firms has been discarded as the stocks for these firms were not traded in part of the period considered in the analysis. These firms are Covidien, Google, Kraft Foods, Mastercard, NYSE Euronext and Philip Morris International. Data were obtained from DataStream. The analysis considers three forecast horizons: one day, five days, and ten days, where the latter two correspond to one and two working weeks, respectively. To cover a variety of cases, combinations of different in- and out-of-sample sizes are considered that span a wide range of possibilities (in-sample corresponds to the part of data which is used for estimation of the parameters while the out-of-sample is the portion of data that is being forecast; here n and m, respectively). Seven different comparisons are analysed. Three of them (I, II and III) compare CL to MG, while the remaining ones (IV, V, VI and VII) compare the two pooling methods to QMLE.

A "test function" is required for the conditional GW test. We employ a test function which consists of a constant and the previous period's loss-difference, namely, $h_t = (1, \Delta L_{m,t-1+\tau})'$ (this is the same test function used by Giacomini and White (2006). It must be mentioned that the choice of a test function could perhaps be a separate research topic as Giacomini and White (2006) explicitly mention both the importance of choosing an appropriate test function and the possible issues due to choosing an irrelevant one). Possible time-independent difference in the predictive abilities of the two methods at any point in time is reflected by the constant. Past comparisons of methods can also give an idea about their relative future performances since a method that has been superior in the past is more likely to be so in the future, as well. This is reflected by the past loss difference.

Lastly, all tests are conducted on an asset-by-asset basis; that is, comparison of predictive ability is conducted for each asset individually, using estimators obtained by the three methods. The test results are presented in Tables 5 and 6. It is an interesting idea to integrate the GW test into a pooling framework, where a single test for the whole panel is conducted; this is left for future research.

Both conditional and unconditional approaches exhibit similar patterns. An immediate observation is that CL performs distinctively better than both MG and QMLE when the in-sample size is very small ($m \leq 250$). Moreover, MG is also superior to QMLE in most of the cases when $m \leq 250$. This has two implications. First, in small-T panels where QMLE is expected to perform poorly, information pooling methods deliver better forecasting performance, suggesting that they provide better estimates in such cases. The second and more interesting implication is that CL outperforms MG, again, in small-T panels. This can be explained by the fact that CL uses all available information to obtain a single estimate while MG use information in a piece-wise fashion. Thus, even though MG blends all estimates, it is nevertheless based on individual estimates that are obtained by using limited information.

In larger samples, CL only rarely outperforms other methods. This is not surprising: when T is large, QMLE models volatility adequately; CL, on the other hand, is based on the restrictive assumption that parameters of interest are common across series. Therefore, as T gets larger, it becomes harder for CL to outperform QMLE. The performance of MG against QMLE, compared to that of CL against QMLE, is also in support of this view since MG performs better than CL against QMLE when there are sufficiently many observations ($m \geq 1,000$).

It must be noted that here, the assumption that all series share a common θ is almost certainly violated. Therefore, as QMLE starts working properly with increasing number of observations, it outperforms the information pooling methods. Moreover, in practice, estimation is always problematic, putting QMLE at a higher disadvantage when samples are not sufficiently large, making it easier for MG and CL to outperform QMLE. This also explains why, in smaller samples, CL is superior to MG, which is essentially based on QMLE.

Comparing the CL and QMLE methods at $m \leq 250$, both the number of cases where equal predictive ability is rejected and the ratio of rejections in favour of CL increase with the out-of-sample size, n. This supports the validity of the test results, as Giacomini and White (2006) mention that, based on their analysis, their test has better size properties as n increases. It must be noted that Giacomini and White (2006) use different loss functions and, of course, focus on an entirely different case. However, this result is still not surprising since the asymptotic theory is more relevant as $n \to \infty$.

Our analysis of 1-week and 2-week ahead forecasts, available upon request, reveals that in all seven comparisons the number of rejections of equal predictive ability decreases. This is more pronounced for the comparison between CL and MG, where in some cases only a handful of rejections remain. This is most likely due to the increase in volatility as the forecast horizon increases. Nevertheless, the pattern of preference between two methods does not change much and CL still performs better when *both* m is very low and n is very high. This issue is less severe for the comparison of CL and MG to QMLE, and the GW test is still able to distinguish the forecasting performances of different methods.

To summarise, information pooling methods perform well in small-T panels while they fail to outperform QMLE in large-T panels. Within the information pooling methods, as expected, CL delivers satisfactory performance against MG in smaller samples, despite the incidental parameter problem.

Empirical results support the view that assuming a common set of parameters is the cost of using the information pooling methods. However, the analysis also strongly suggests that when the sample consists of a small number of observations, these costs are outweighed by the advantages of the information pooling mechanism and the problems faced by QMLE. On the other hand, when the sample is sufficiently large for QMLE to work well, empirical results confirm that the cost of assuming common parameters of interest is high.

5 Conclusion

This paper studied the theoretical and empirical properties of the composite likelihood (CL) method on the special case of GARCH panels. The MacGyver (MG) method has also been included in the empirical analysis as it is the only known alternative information pooling method.

Simulation and empirical analyses reveal that using the panel structure and CL instead of employing QMLE on a single series delivers better results. Both methods suffer from the incidental parameter problem when T is small, but the CL is much more accurate. These observations are very encouraging as they imply that CL can successfully estimate conditional volatility using panels where T is as low as 250. Furthermore, forecast comparison analysis demonstrates that even when the assets are likely to be characterised by different parameter sets, both pooling methods perform well against QMLE in small-T panels.

6 Acknowledgements

We thank Andrew Patton and Alexander D. Taylor for their comments on previous versions of this work.

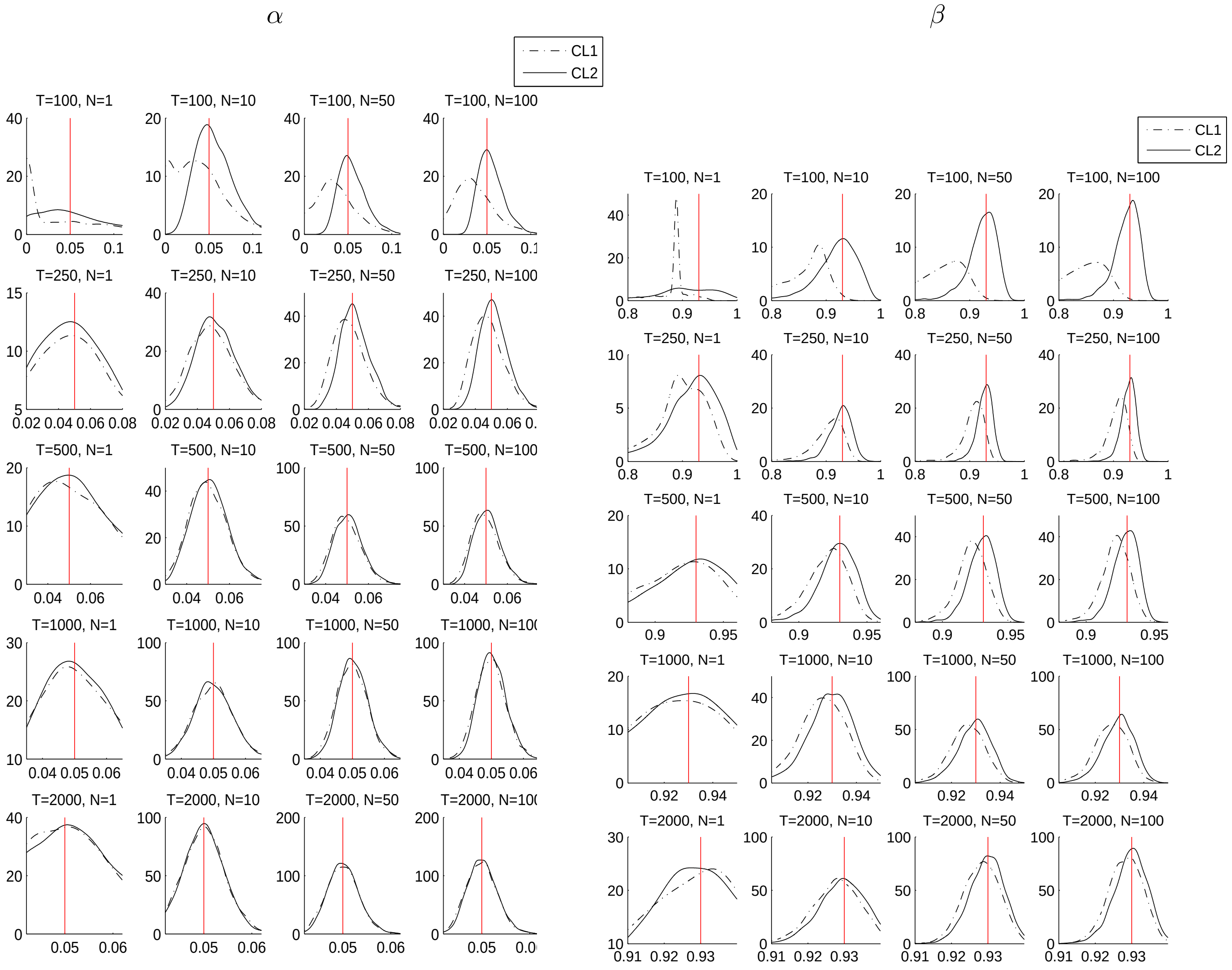

Figure 1: Sample distribution graphs for $\hat{\alpha}$ (left) and $\hat{\beta}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. CL1 gives the sample distribution for the case with nuisance parameter estimation, whereas CL2 is the sample distribution for the case where nuisance parameter estimation is by-passed. Values of $\hat{\alpha}$ and $\hat{\beta}$ are given in the horizontal axis, while respective frequencies are given in the vertical axis. Vertical lines are drawn at the true value of each parameter.

		I		II		III		IV		V		VI		VII	
		CL vs MG-Mn		CL vs MG-Md		CL vs MG-Tr		CL vs QMLE		Mn vs QMLE		Md vs QMLE		Tr vs QMLE	
T	m	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	60	3.33	63	0.00	61	0.00	50	10.00	44	68.18	44	38.64	42	52.38
2,000	1,000	59	25.42	42	16.67	57	15.79	51	21.57	47	46.81	53	33.96	44	47.73
1,000	500	44	11.36	36	41.67	43	9.30	30	60.00	30	83.33	35	62.86	29	82.76
2,000	250	18	94.44	35	97.14	19	94.74	78	100.00	74	100.00	75	98.67	75	98.67
1,000	250	20	90.00	41	92.68	31	96.77	50	100.00	42	97.62	42	97.62	42	97.62
500	250	8	100.00	46	100.00	18	100.00	36	94.44	38	92.11	30	80.00	38	89.47
1,000	100	25	96.00	31	100.00	22	95.45	49	100.00	46	100.00	49	97.96	50	96.00
500	100	29	100.00	42	100.00	37	100.00	45	97.78	38	100.00	33	100.00	35	100.00
250	100	9	77.78	14	64.29	12	75.00	22	86.36	28	67.86	20	80.00	25	68.00

Table 5: Conditional test results for 1-Step forecasts. T is the total number of observations while m gives the in-sample size. Mn, Md, and Tr stand for 'mean', 'median', and 'trimmed', respectively. The total number of assets is 94. For each comparison, 'Rej' gives the number of assets for which the null hypothesis of equal predictive ability is rejected. % gives the percentage of the cases where the first method is preferred to the second method. For example, in comparison III when T=2000 and m=250, equal predictive ability is rejected in 19 out of 94 comparisons and 94.74% of these rejections are in favour of the first method, CL (or, equivalently, 5.26% of the rejections are in favour of the second method, MG-Tr).

		I		II		III		IV		V		VI		VII	
		CL vs MG-Mn		CL vs MG-Md		CL vs MG-Tr		CL vs QMLE		Mn vs QMLE		Md vs QMLE		Tr vs QMLE	
T	m	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	53	0.00	56	0.00	56	0.00	44	4.55	40	67.50	34	35.29	36	50.00
2,000	1,000	47	17.02	27	18.52	49	12.24	39	12.82	41	51.22	50	34.00	43	48.84
1,000	500	40	7.50	21	38.10	38	5.26	27	55.56	29	86.21	32	62.50	30	80.00
2,000	250	14	85.71	31	100.00	18	88.89	73	100.00	67	100.00	73	100.00	69	100.00
1,000	250	19	94.74	46	97.83	26	96.15	50	100.00	39	100.00	38	100.00	40	100.00
500	250	10	90.00	50	100.00	18	100.00	31	100.00	34	94.12	25	92.00	34	91.18
1,000	100	29	96.55	29	100.00	26	96.15	48	100.00	41	100.00	45	100.00	45	100.00
500	100	29	100.00	40	100.00	32	100.00	43	100.00	38	100.00	32	100.00	37	100.00
250	100	10	70.00	16	68.75	10	70.00	23	78.26	27	74.07	20	80.00	26	76.92

Table 6: Unconditional test results for 1-Step forecasts. See Table 5 for explanations.

References

Andersen, T. and L. Benzoni (2009). Realized volatility. In *Handbook of Financial Time Series* (Edited by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch), 555–575. Springer-Verlag.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of exchange rate volatility. *Journal of the American Statistical Association* **96**, 42–55. Correction published in 2003, volume 98, page 501.

Andrews, D.W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–858.

Arellano, M. and B. Honore (2001). Panel data models: some recent developments. In *Handbook of Econometrics*, Volume 5 (Edited by J. J. Heckman and E. Leamer), 3229–3296. North-Holland, Amsterdam.

Barndorff-Nielsen, O. E. (1996). Two index asymptotics. In *Frontiers in Pure and Applied Probability II: Proceedings of the Fourth Russian-Finnish Symposium on Theoretical and Mathematical Statistics* (Edited by A. Melnikov), 9–20. TVP Science, Moscow.

Barndorff-Nielsen, O. E. and N. Shephard (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64**, 253–280.

Bauwens, L., S. Laurent, and J. V. K. Rombouts (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics* **21**, 79–109.

Bauwens, L. and J. V. K. Rombouts (2007). Bayesian clustering of many GARCH models. *Econometric Reviews* **26**, 365–386.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994). ARCH models. In *The Handbook of Econometrics*, Volume 4 (Edited by R. F. Engle and D. McFadden), 2959–3038. North-Holland, Amsterdam.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**, 253–263.

Diggle, P. J., K. Y. Liang, and S. L. Zeger (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Engle, R. F. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20**, 339–350.

Engle, R. F. (2009). High dimensional dynamic correlations. In *The Methodology and Practice of Econometrics: Papers in Honour of David F Hendry* (Edited by J. L. Castle and N. Shephard), 122–148. Oxford University Press.

Engle, R. F. and J. Mezrich (1996). GARCH for groups. *Risk*, **9**, 36–40.

Engle, R. F., N. Shephard, and K. K. Sheppard (2008). Fitting vast dimensional time-varying covariance models. Working paper.

Engle, R. F. and K. K. Sheppard (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Working paper.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545–1578.

Hansen, P. R. and A. Lunde (2006). Consistent ranking of volatility models. *Journal of Econometrics* **131**, 97–121.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* **95**, 391–413.

Liang, K.-Y. and D. Tsou (1992). Empirical bayes and conditional inference with many nuisance parameters. *Biometrika* **79**, 261–270.

Lindsay, B. G. (1985). Using empirical partially bayes inference for increased efficiency. *The Annals of Statistics* **13**, 914–931.

Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (Edited by N. U. Prabhu), 221–239. Amercian Mathematical Society, Providence, RI.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In *The Handbook of Econometrics*, Volume 4 (Edited by R. F. Engle and D. McFadden), 2111–2245. North-Holland, Amsterdam.

Newey, W. K. and K. D. West (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–16.

Patton, A. J. (2008). Volatility forecast evaluation and comparison using imperfect volatility proxies. Forthcoming in *Journal of Econometrics.*

Patton, A. J. and K. K. Sheppard (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series* (Edited by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch), 801–838. Springer-Verlag.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Silvennoinen, A. and T. Teräsvirta (2009). Multivariate GARCH models. In *Handbook of Financial Time Series* (Edited by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch), 201–229. Springer-Verlag.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1–28.

Varin, C., N. Reid, and D. Firth (2009). An overview of composite likelihood methods. Forthcoming in *Statistica Sinica.*

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–1084.