

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

October 2, 2009

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 in reasonably large T CL performs well. However, due to the estimation error introduced through nuisance parameter estimation, 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 CL method has been introduced to financial econometrics only recently by Engle, Shephard, and Sheppard (2008)¹ and as a basis for pooling information across not only time, but also cross-sectionally. In GARCH panels, this amounts to estimating the parameters of interest for all assets simultaneously, instead of individually. This is very important as the common quasi-maximum likelihood estimator (QMLE) for the GARCH model delivers poor results in samples of a few hundred observations. It is illustrated in this study that, through its information pooling mechanism, CL is potentially capable of delivering satisfactory results in such samples. However, both methods suffer from error introduced by nuisance parameter estimates. This “incidental parameter” problem has already 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, as mentioned before, QMLE based estimation of GARCH, while satisfactory in samples with 1,000s 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 will be correlated to some degree, it is implausible that all assets will be perfectly correlated. Hence, a panel of asset prices will contain 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 availability of around 250 daily observations following the break) QMLE will most likely deliver poor fitted models of conditional heteroskedasticity which, in turn, will lead to poor forecasting performance. On the other hand, CL has the potential to work well for this problem. Another application where CL can be useful is monthly hedge fund data, which consists of monthly returns on 1000s of funds and hence, is a short, wide panel.

The relevant large sample theory underlying the method used in this study has already been developed by Engle, Shephard, and Sheppard (2008) who look at large dimensional time-varying covariances. They employ CL to produce a computationally feasible estimator, where the CL is 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 het-

¹The origins of the composite likelihood method go back to at least Lindsay (1988). See Varin (2008) for a review.

eroskedasticity correctly using previously infeasible sample sizes. Furthermore, forecast comparison analysis on stock-market data from S&P100 reveals 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. This is followed by Section 4 which 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 will be 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 share a common set of parameters, $\theta = (\alpha, \beta)$, while the nuisance parameters, $\{\gamma_i\}_{i=1}^N$, are allowed to be asset-dependent. Our focus will be on fitting a very large number of univariate GARCH models.² 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 , $i = 1, \dots, N_T$ and $t = 1, \dots, T$. Moreover,

$$y_{it} = \mu_{it} + \varepsilon_{it}, \quad \mu_{it} = \mathbb{E}(y_{it} | \mathcal{F}_{t-1}), \quad (1)$$

$$\mathbb{E}[\varepsilon_{it} | \mathcal{F}_{t-1}] = 0 \quad \text{and} \quad \text{Var}[y_{it} | \mathcal{F}_{t-1}] = \text{Var}[\varepsilon_{it} | \mathcal{F}_{t-1}] \equiv \sigma_{it}^2, \quad (2)$$

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

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

²For example, this would be need for the first step of fitting a Dynamic Conditional Correlation model by Engle (2002).

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

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

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

This panel structure has many similarities with the autoregressive panels which are 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, by 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 assets returns at time t is given by $f(y_{1t}, \dots, y_{N_T t}|\mathcal{F}_{t-1})$, which we will not model, noting that knowledge of all of the N_T submodels does not deliver knowledge of $f(y_{1t}, \dots, y_{N_T t}|\mathcal{F}_{t-1})$ (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 .³ In our framework we will have time series and cross-sectional dependence in the $y_{it}|\mathcal{F}_{t-1}$.

The composite likelihood function is then defined as

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

³Then the maximum likelihood estimator of θ is typically inconsistent for finite T and $N \rightarrow \infty$ 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)). However, 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 reduce the rate of convergence to rate \sqrt{T} , not \sqrt{NT} . Under those circumstances MLE will be consistent and have a simple limit theory however N relates to T (see Engle, Shephard, and Shephard (2008) for details).

$$l(\psi_{(N_T)}; \mathbf{y}) = \sum_{t=1}^T l_t(\psi_{(N_T)}; y_t | \mathcal{F}_{t-1}), \quad \text{and} \quad (6)$$

$$l_t(\psi_{(N_T)}; y_t | \mathcal{F}_{t-1}) = \sum_{i=1}^{N_T} \log f(y_{it} | \mathcal{F}_{t-1}; \psi_i), \quad \psi_i \equiv (\theta', \gamma_i), \quad \psi_{(N_T)} = (\theta', \gamma'_{(N_T)})'.$$

Estimation of $\psi_{(N_T)}$ is based on the 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).⁴

Formally, using (4),

$$m_{N_T}(y_t, \gamma_{(N_T)}) = \begin{pmatrix} y_{1t}^2 - \gamma_1 \\ \vdots \\ y_{N_T t}^2 - \gamma_{N_T} \end{pmatrix}, \quad \text{implying } E(m_{N_T}(y_t, \gamma_{(N_T)}^*)) = 0. \quad (7)$$

Equation (7) gives the population moment condition for the nuisance parameters. For θ , the score function for the normal composite-likelihood function is

$$g(y_t, \theta, \gamma_{(N_T)}) = \frac{\partial}{\partial \theta} \frac{1}{N_T} \left(-\frac{1}{2} \sum_{i=1}^{N_T} \log \sigma_{it}^2 - \frac{1}{2} \sum_{i=1}^{N_T} \frac{\eta_{it}^2}{\sigma_{it}^2} \right). \quad (8)$$

For (7) and (8), respective sample moment conditions are given by

$$\frac{1}{T} \sum_{t=1}^T m_{N_T}(y_t, \hat{\gamma}_{(N_T)}) = 0, \quad \text{and} \quad \frac{1}{T} \sum_{t=1}^T g(y_t, \hat{\gamma}_{(N_T)}, \hat{\theta}) = 0, \quad (9)$$

where $\hat{\gamma}_{(N_T)}$ and $\hat{\theta}$ are appropriate estimators for $\gamma_{(N_T)}^*$ and $\theta^* \equiv (\alpha^*, \beta^*)$. Stacking (7) and (8), the population and sample moment conditions are given by

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

We note that (9) is the first order condition for the simple optimization problem

$$\hat{\theta} = \arg \max_{\theta} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_T} \sum_{i=1}^{N_T} \log f(y_{it} | \mathcal{F}_{t-1}, \theta, \hat{\gamma}_i). \quad (10)$$

Equation (10) 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 will provide a statistically inefficient estimator for θ as it ignores dependence over individuals, employs a moment based

⁴A 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.

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 (2003). In that case, the score for the t -th observation is given by

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

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

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

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 is i.i.d. in the cross section. More formally we assume that if N increases, the cross sectional average $s_{t,N}$ does not exhibit a law of large number.

Based on this central limit result, it follows that

$$\sqrt{T} (\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}), \quad (11)$$

where

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

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

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

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

An important point of (11) 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 is generated using the specification described in (1)-(3). For most stock returns annual volatility is in the range 15% and 60%, so we draw $\gamma_i \stackrel{iid}{\sim} U[0.02, 0.05]$.⁵ For each series the γ_i are used as the initial values for the conditional variances, $h_{i,0}$. Cross-sectional dependence is generated by a single-factor model where

$$\eta_{it} = \rho_i u_t + \sqrt{1 - \rho_i^2} \tau_{it}, \quad \tau_{it}, u_t \stackrel{iid}{\sim} N(0, 1), \quad (12)$$

implying

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

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

The choice of 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 ρ_i is a constant. Engle, Shephard, and Sheppard (2008) consider a truncated normal distribution to generate ρ_i , where truncation occurs at 0.1 and 0.9. This study uses $\rho_i \sim U[0.5, 0.9]$ for all i , ensuring that the lowest and highest correlation between two assets will be equal to 0.25 and 0.81, respectively. Lastly α and β are chosen from three different alternatives which cover the range of parameter values which are found in asset data:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \left\{ \theta^{(1)}, \theta^{(2)}, \theta^{(3)} \right\} = \left\{ \begin{bmatrix} 0.02 \\ 0.97 \end{bmatrix}, \begin{bmatrix} 0.05 \\ 0.93 \end{bmatrix}, \begin{bmatrix} 0.10 \\ 0.80 \end{bmatrix} \right\}. \quad (13)$$

⁵This follows from $\sigma_D = \sqrt{\sigma_A/252}$ 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.

$\theta =$	(0.02, 0.97)		(0.05, 0.93)		(0.10, 0.80)		(0.02, 0.97)		(0.05, 0.93)		(0.10, 0.80)	
N_T	Bias						MCSD					
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
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: 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.

3.2 Simulation results

All results in the following analysis 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. The statistics are defined as follows:

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

where Z is the number of replications, $\hat{\alpha}_z$ and $\hat{\beta}_z$ are the estimates for replication z , $z = 1, \dots, Z$ and $\hat{\kappa}_z \in \{\hat{\alpha}_z, \hat{\beta}_z\}$. $\hat{\sigma}_{\hat{\kappa},z}^2$ 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 to evaluate the finite sample performance of asymptotic distribution for the CL based upon $\hat{\sigma}_{\hat{\kappa},z}^2$. All results are calculated for 95% confidence intervals.

Tables 1 and 2 present the results for the three parameter values in (13), where $T = 2,000$. A second analysis looks at the implications of varying T where $T \in \{100, 250, 500, 1000, 2000\}$, the results of which are given in Tables 3 and 4.⁶ In all cases, results for $N_T = 1$ are also provided, which corresponds to using QMLE instead of CL.

Simulation results show that when $T = 2,000$, CL generally leads to low average bias across all parameter values, with the highest average bias being equal to .88% for $\hat{\alpha}$ and $-.40\%$ for $\hat{\beta}$. In contrast, the average bias due to QMLE reaches levels as high as 3.34%. An interesting observation

⁶This second analysis is conducted for $\theta^{(2)}$ only, due to space restrictions.

$\alpha = 0.02 \quad \beta = 0.97$								
N_T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$RMSE_{\hat{\alpha}}$	$RMSE_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
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
$\alpha = 0.05 \quad \beta = 0.93$								
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
$\alpha = 0.10 \quad \beta = 0.80$								
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 (RMSE) 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.

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 $\theta^{(1)}$, the change in bias when N_T increases from 50 to 100 is equal to .01%. Moreover, taking $N_T = 1$ as a reference, when the panel size increases by a factor of 10 (to $N_T = 10$) the change in bias is equal to .52% while increasing the panel size 100-fold (to $N_T = 100$) leads to a reduction by .56%. This shows that the speed of decline falls with N_T . Table ?? also suggest that there are substantial gains in shifting from time-series (QMLE) to panel (CL) structure, in terms of both the average bias and sample standard deviation.

Sample standard deviations 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 fall in MCSDs is not large enough to imply that the speed of convergence in finite samples is equal to $\sqrt{TN_T}$ 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 further beyond 100 will not lead to substantial decreases in MCSD. These results are all in accordance with the asymptotic theory in Engle, Shephard, and Sheppard (2008).

Table ?? shows that, for CL, 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 non-vanishing bias as $N_T \rightarrow \infty$, since in some cases there is a slight difference between MCSD and RMSE suggesting 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

	$N_T=1$ (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: average biases for $\hat{\alpha}$ and $\hat{\beta}$ in percentages and Monte Carlo standard deviations ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$). $\alpha = 0.05$, $\beta = 0.93$. 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.

of focusing on the series individually. Also, CI statistics are very satisfactory, ranging between 92% and 97% across all parameter sets and cross-section sizes.

Now, we turn to the implications of varying both the number of assets and 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 $\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$, 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 for, when $T = 100$, which is due to the optimisation routines 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

$N_T=1$ (QMLE)								$N_T=10$								
					RMSE								RMSE			
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\alpha}$	$\hat{\beta}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\alpha}$	$\hat{\beta}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
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: 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. $\alpha = 0.05$, $\beta = 0.93$. 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. based on 2,500 replications.

small- T panels is investigated 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 γ_i for each model?

The sampling distributions of the estimators of θ for the cases with (CL1) and without (CL2) nuisance parameter estimation are presented in Figure 1. 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 ; for a given T , an increase in N_T leads to higher precision.⁷

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,

⁷However, 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 will be 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 will not lead to any improvement in efficiency.

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.

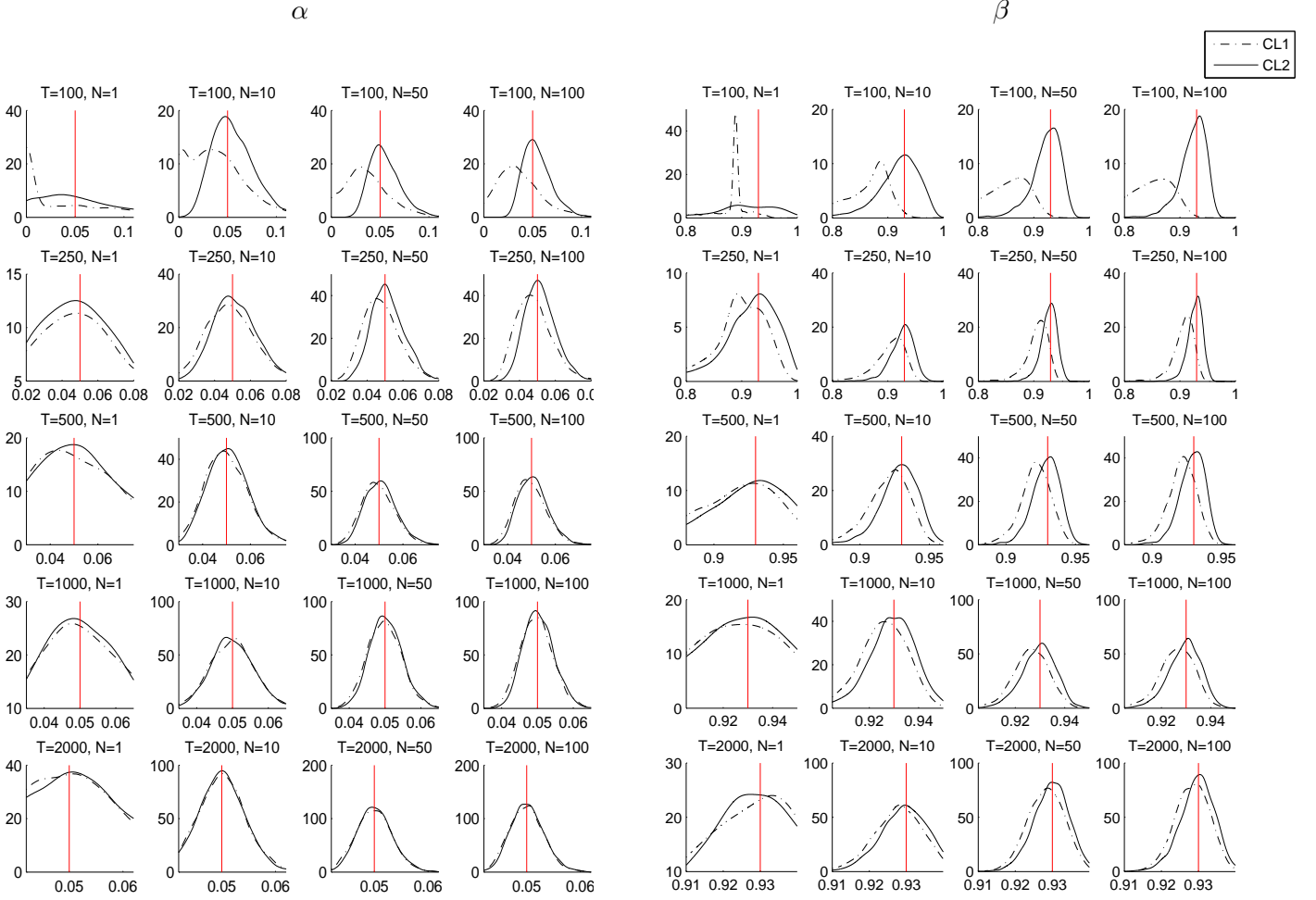


Figure 1: Sample distribution graphs for $\hat{\alpha}$ (left) and $\hat{\beta}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. CL1 gives the sample distribution for the case with nuisance parameter estimation, whereas CL2 is the sample distribution for the case where nuisance parameter 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. The vertical lines are drawn at the true value of each parameter.

It is also interesting that when QMLE is used ($N_T = 1$), even when the true nuisance parameter is known, the estimators are still performing 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 very small to adequately model conditional volatility, it is obvious that knowledge of the true value of the nuisance parameter will not help.

4 Empirical analysis

In this section, in addition to CL and QMLE, we also 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 where the highest and lowest 5% of the estimates are eliminated. The latter two blending functions serve the purpose of discarding outliers, which could otherwise introduce bias.

For the GARCH panel, $\{\hat{\theta}_i\}$ are estimated using two-step estimation. In the first step, $\{\gamma_i\}$ are obtained in the same way as for CL. In the second step, $\{\hat{\theta}_i\}$ are estimated using

$$\hat{\theta}_i = \arg \max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T \log f(y_{it} | \mathcal{F}_{t-1}; \theta, \hat{\gamma}_i).$$

It must be noted that there are several practical issues related to this method. First of all, when the sample size is not large enough, in some cases optimisation may fail and simply yield the initial values used for optimisation as the parameter estimates.⁸ 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. Therefore, $\hat{\beta}$ has no interpretation in such a cases, no matter what its value is. Consequently, this analysis also ignores sets of estimates where $\hat{\alpha}$ is less than 0.0025.⁹ 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 pool of a large number 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,

⁸The 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.

⁹These 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 $\hat{\alpha} = 0$.

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, which 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 empirical analysis. Moreover, the assumption of a common set of parameters for all assets is not likely to hold. Be that as it may, out of this arguably difficult situation, a highly interesting question arises: despite these issues, can the CL and MG methods indeed 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 TARARCH 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. Therefore, the loss function provides a criterion to assess how well $\hat{Y}_{t+\tau}$ predicts $Y_{t+\tau}$. Formally, the loss due to $\hat{Y}_{t+\tau}$ is defined as

$$L_{t+\tau} \left(Y_{t+\tau}, \hat{Y}_{t+\tau} \right). \quad (14)$$

Examples of loss functions used in the literature are many.¹⁰ A prominent example that will be used in this study is the QLIKE 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

¹⁰See Patton (2008) for a more detailed study of implications of using different loss functions.

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. Then, the Null Hypothesis is,

$$H_0 : E \left[L_{t+\tau}(Y_{t+\tau}, f_t(\hat{\beta}_{1t})) - L_{t+\tau}(Y_{t+\tau}, g_t(\hat{\beta}_{2t})) | \mathcal{G}_t \right] = 0, \quad (15)$$

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.

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, r_t^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.¹¹ This is important for the choice of the loss function. Hansen and Lunde (2006) show that the use of noisy proxies such as r_t^2 may lead to inconsistent ranking of volatility models, whereby the empirical ranking may not necessarily be the same as the true ranking. Patton (2009) 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 stocks¹² from S&P100 for the period between 1 April 2000 to 12 January 2008. 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

¹¹Very 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.

¹²Data 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.

in- and out-of-sample sizes¹³ (n and m , respectively) are considered, which span a wide range of possibilities. Seven different comparisons are analysed. Three of those (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.¹⁴ Namely, $h_t = (1, \Delta L_{m,t-1+\tau})'$. 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.¹⁵ The test results are presented in Tables 5 to 10.

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.

Considering larger samples, CL only rarely outperforms other methods. This can be related to the incidental parameter issue, which is now more costly since QMLE starts delivering better results as T increases. The performance of MG against QMLE, compared to that of CL against QMLE, is also in support of this view since MG performs less poorly 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, as this paper uses real instead of

¹³In-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.

¹⁴This 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.

¹⁵It can be 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.

“tailored” data, estimation is always more 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 increases¹⁶.

Observing the changes due to the increase in the forecasting horizon from one day to two weeks 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. However it is encouraging to observe that even in samples of size 500 or 1,000, CL and MG still deliver superior performance in a few cases. 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 loss of using the MG and CL methods is very 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.

¹⁶It 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 will be more relevant as $n \rightarrow \infty$.

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.

To conclude, the general message of this study is that the CL method has the potential to carry volatility modelling into small- T GARCH panels, where QMLE remains ineffective.

6 Acknowledgements

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

References

- Andersen, T. and L. Benzoni (2009). Realized volatility. In T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 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 J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5. Amsterdam: North-Holland. Forthcoming.
- Barndorff-Nielsen, O. E. (1996). Two index asymptotics. In A. Melnikov (Ed.), *Frontiers in Pure and Applied Probability II: Proceedings of the Fourth Russian-Finnish Symposium on Theoretical and Mathematical Statistics*, pp. 9–20. Moscow: TVP Science.
- 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 R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics, Volume 4*, pp. 2959–3038. Amsterdam: North-Holland.
- Cox, D. R. and N. Reid (2003). 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: Oxford University Press.
- 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 J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics: Papers in Honour of David F Hendry*, pp. 122–148. Oxford University Press.
- Engle, R. F. and J. Mezrich (1996). GARCH for groups. *Risk*, 36–40.
- Engle, R. F., N. Shephard, and K. K. Sheppard (2008). Fitting vast dimensional time-varying covariance models. Unpublished paper: Oxford-Man Institute, University of Oxford.
- Engle, R. F. and K. K. Sheppard (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Unpublished paper: UCSD.
- 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.
- Lindsay, B. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, pp. 221–239. Providence, RI: American Mathematical Society.
- Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics, Volume 4*, pp. 2111–2245. North-Holland.
- 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. (2009). Volatility forecast evaluation and comparison using imperfect volatility proxies. *Journal of Econometrics*. Forthcoming.
- Patton, A. J. and K. K. Sheppard (2009). Evaluating volatility forecasts. In T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 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 T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 201–229. Springer-Verlag.
- Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* 92, 1–28.
- West, K. (1996). Asymptotic inference about predictive ability. *Econometrica* 64, 1067–1084.

T	m	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	
		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, for comparison III, T=2000 and m=250, in 19 out of 94 comparisons, equal predictive ability is rejected 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).

T	m	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	
		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.

T	m	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	
		Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	49	8.16	53	1.89	50	2.00	45	6.67	39	53.85	42	28.57	37	37.84
2,000	1,000	60	41.67	47	31.91	61	29.51	49	14.29	52	36.54	57	28.07	54	38.89
1,000	500	44	9.09	22	13.64	46	6.52	30	50.00	27	70.37	31	51.61	30	70.00
2,000	250	22	72.73	16	62.50	15	60.00	68	98.53	65	100.00	72	98.61	66	98.48
1,000	250	8	37.50	21	85.71	6	66.67	38	92.11	38	94.74	38	94.74	40	90.00
500	250	3	0.00	29	100.00	3	33.33	23	91.30	26	84.62	22	81.82	25	88.00
1,000	100	4	50.00	11	90.91	3	33.33	47	93.62	48	95.83	51	88.24	49	95.92
500	100	4	50.00	19	100.00	9	77.78	30	100.00	28	100.00	24	91.67	29	96.55
250	100	2	0.00	5	20.00	2	0.00	9	77.78	9	66.67	8	87.50	10	70.00

Table 7: Conditional test results for 5-Step forecasts. See Table 5 for explanations.

21

T	m	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	
		Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	49	2.04	54	0.00	52	0.00	45	4.44	39	51.28	40	25.00	34	38.24
2,000	1,000	49	34.69	34	26.47	47	21.28	43	6.98	43	34.88	49	26.53	46	36.96
1,000	500	48	12.50	22	31.82	44	6.82	30	46.67	32	62.50	35	45.71	27	66.67
2,000	250	7	85.71	9	77.78	5	100.00	56	100.00	54	100.00	60	100.00	57	100.00
1,000	250	5	40.00	20	100.00	3	100.00	30	100.00	30	100.00	29	96.55	29	100.00
500	250	5	0.00	40	100.00	1	100.00	28	85.71	32	81.25	30	70.00	31	80.65
1,000	100	0	n/a	11	100.00	3	100.00	39	100.00	40	100.00	38	100.00	41	100.00
500	100	4	50.00	28	100.00	12	83.33	32	96.88	29	96.55	28	89.29	29	96.55
250	100	3	33.33	9	44.44	5	20.00	21	57.14	18	66.67	19	63.16	18	61.11

Table 8: Unconditional test results for 5-Step forecasts. See Table 5 for explanations.

T	m	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	
		Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	41	19.51	38	2.63	40	10.00	43	9.30	29	41.38	31	22.58	28	35.71
2,000	1,000	51	58.82	41	34.15	50	42.00	44	11.36	53	28.30	50	24.00	54	29.63
1,000	500	27	11.11	16	12.50	26	7.69	26	46.15	23	65.22	24	50.00	22	63.64
2,000	250	29	93.10	13	69.23	23	86.96	55	96.36	57	89.47	58	93.10	57	92.98
1,000	250	6	33.33	15	93.33	3	33.33	21	90.48	21	90.48	22	100.00	21	90.48
500	250	1	0.00	15	100.00	3	66.67	12	100.00	13	100.00	8	100.00	11	100.00
1,000	100	1	0.00	7	85.71	3	66.67	36	94.44	41	87.80	42	88.10	41	90.24
500	100	1	100.00	10	90.00	1	100.00	20	75.00	21	71.43	16	75.00	19	73.68
250	100	0	n/a	3	33.33	0	n/a	1	100.00	1	100.00	2	100.00	1	100.00

Table 9: Conditional test results for 10-Step forecasts. See Table 5 for explanations.

T	m	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	
		Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%	Rej	%
2,000	1,500	45	6.67	53	0.00	49	0.00	49	6.12	36	41.67	41	21.95	33	33.33
2,000	1,000	50	48.00	36	36.11	48	33.33	46	8.70	49	28.57	49	14.29	49	28.57
1,000	500	46	17.39	24	20.83	48	12.50	36	27.78	32	50.00	38	34.21	32	46.88
2,000	250	16	100.00	5	80.00	10	100.00	48	100.00	46	100.00	50	100.00	47	100.00
1,000	250	7	14.29	14	92.86	2	100.00	24	100.00	23	100.00	23	100.00	23	100.00
500	250	6	0.00	26	100.00	4	100.00	25	76.00	29	68.97	28	60.71	29	68.97
1,000	100	2	0.00	11	90.91	3	33.33	35	100.00	35	100.00	34	100.00	35	100.00
500	100	4	75.00	20	100.00	5	80.00	27	88.89	29	82.76	24	79.17	27	88.89
250	100	2	50.00	7	85.71	3	33.33	23	34.78	26	30.77	23	26.09	25	32.00

Table 10: Unconditional test results for 10-Step forecasts. See Table 5 for explanations.