

NUISANCE PARAMETERS, COMPOSITE LIKELIHOODS AND A PANEL OF GARCH MODELS

Cavit Pakel, Neil Shephard and Kevin Sheppard

University of Oxford

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 .

Key words and phrases: 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 (2011) 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.

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 is 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 , $i = 1, \dots, N_T$ and $t = 1, \dots, T$. We write,

$$y_{it} = \mu_{it} + \varepsilon_{it}, \quad \mu_{it} = E[y_{it}|\mathcal{F}_{t-1}], \tag{2.1}$$

$$E[\varepsilon_{it}|\mathcal{F}_{t-1}] = 0 \quad \text{and} \quad \text{Var}[y_{it}|\mathcal{F}_{t-1}] = \text{Var}[\varepsilon_{it}|\mathcal{F}_{t-1}] \equiv \sigma_{it}^2, \tag{2.2}$$

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

$$\sigma_{it}^2 = \gamma_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \quad \text{where } \gamma_i > 0, \quad \alpha, \beta \in [0, 1), \quad \alpha + \beta < 1. \tag{2.3}$$

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 $\hat{\theta} = (\hat{\alpha}, \hat{\beta})'$. It can be shown that this specification, often called variance-tracking, implies that

$$E(y_{it}^2) = \gamma_i, \tag{2.4}$$

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 $\gamma_{(N_T)}^* \equiv (\gamma_1^*, \dots, \gamma_{N_T}^*)$.

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 $f(y_{1t}, \dots, y_{N_T t}|\mathcal{F}_{t-1})$, which we do not model, noting that knowledge of all of the N_T submodels does not deliver knowledge of $f(y_{1t}, \dots, y_{N_T t}|\mathcal{F}_{t-1})$ (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 . Then the maximum likelihood estimator of θ is typically inconsistent for finite T as $N \rightarrow \infty$ 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)). 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 $\psi_i \equiv (\theta', \gamma_i)'$ and $\psi_{(N_T)} = (\theta', \gamma'_{(N_T)})'$. Then, the normal-density composite likelihood function is given by

$$\text{CL}(\psi_{(N_T)}; y) = \frac{1}{T} \sum_{t=1}^T \left\{ \frac{1}{N_T} \sum_{i=1}^{N_T} \log f(y_{it}|\mathcal{F}_{t-1}; \psi_i) \right\} = \frac{1}{T} l(\psi_{(N_T)}; y), \quad (2.5)$$

where

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

$$l_t(\psi_{(N_T)}; y_t | \mathcal{F}_{t-1}) = \sum_{i=1}^{N_T} \log f(y_{it} | \mathcal{F}_{t-1}; \psi_i).$$

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 (2.4) to obtain $\{\hat{\gamma}_i\}$, then substituted for $\{\gamma_i\}$ in (2.3), and θ is estimated using (2.5). 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 (2.4),

$$m_{N_T}(y_t, \gamma_{(N_T)}) = \begin{pmatrix} y_{1t}^2 - \gamma_1 \\ \vdots \\ y_{N_T t}^2 - \gamma_{N_T} \end{pmatrix}, \quad \text{implying } E(m_{N_T}(y_t, \gamma_{(N_T)}^*)) = 0. \quad (2.6)$$

Equation (2.6) gives the population moment condition for the nuisance parameters. For θ , the score function for the normal-density composite-likelihood function is

$$g(y_t, \theta, \gamma_{(N_T)}) = \frac{\partial}{\partial \theta} \frac{1}{N_T} \left(-\frac{1}{2} \sum_{i=1}^{N_T} \log \sigma_{it}^2 - \frac{1}{2} \sum_{i=1}^{N_T} \frac{\varepsilon_{it}^2}{\sigma_{it}^2} \right). \quad (2.7)$$

For (2.6) and (2.7), respective sample moment conditions are given by

$$\frac{1}{T} \sum_{t=1}^T m_{N_T}(y_t, \hat{\gamma}_{(N_T)}) = 0, \quad \text{and} \quad \frac{1}{T} \sum_{t=1}^T g(y_t, \hat{\theta}, \hat{\gamma}_{(N_T)}) = 0, \quad (2.8)$$

where $\hat{\gamma}_{(N_T)}$ and $\hat{\theta}$ are appropriate estimators for $\gamma_{(N_T)}^*$ and $\theta^* \equiv (\alpha^*, \beta^*)$. Stacking (2.6) and (2.7), the population and sample moment conditions are given by

$$E \left[\tilde{g}(y_t, \theta^*, \gamma_{(N_T)}^*) \right] = E \begin{bmatrix} m_{N_T}(y_t, \gamma_{(N_T)}^*) \\ g(y_t, \theta^*, \gamma_{(N_T)}^*) \end{bmatrix} = 0,$$

and

$$\frac{1}{T} \sum_{t=1}^T \tilde{g}_{t,T}(y_t, \hat{\theta}, \hat{\gamma}_{(N_T)}) = \underset{([N_T+2] \times 1)}{0}.$$

We note that (2.8) is the first order condition for the simple optimization problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_T} \sum_{i=1}^{N_T} \log f(y_{it} | \mathcal{F}_{t-1}, \theta, \hat{\gamma}_i). \quad (2.9)$$

Equation (2.9) 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

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

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

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

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 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

$$\sqrt{T} (\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}), \quad (2.10)$$

where

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

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 $\hat{\theta}$, it makes no statement about the γ_i . To account for the nuisance parameters, a modified estimator for the score covariance is required:

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

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

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

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

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 (2.10) 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 (2.1)–(2.3). For most stock returns annual volatility is in the range 15% and 60%, so we took $\gamma_i \stackrel{i.i.d.}{\sim} U[0.02, 0.05]$. 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

$$\varepsilon_{it} = \rho_i u_t + \sqrt{1 - \rho_i^2} \tau_{it}, \quad \tau_{it}, u_t \stackrel{i.i.d.}{\sim} N(0, 1), \tag{3.1}$$

implying

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

and $\text{Cov}(\varepsilon_{it}, \varepsilon_{js}|\rho_i, \rho_j) = 0$ 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:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \left\{ \theta^{(1)}, \theta^{(2)}, \theta^{(3)} \right\} = \left\{ \begin{bmatrix} 0.02 \\ 0.97 \end{bmatrix}, \begin{bmatrix} 0.05 \\ 0.93 \end{bmatrix}, \begin{bmatrix} 0.10 \\ 0.80 \end{bmatrix} \right\}. \quad (3.2)$$

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:

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

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 of the finite sample performance of the 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 (3.2), where $T = 2,000$. Tables 3 and 4 look at the implications of varying T where

Table 1. Monte Carlo simulation results for fixed T using the three parameter sets given in (3.2): 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.

$\theta =$	(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	3.30	-0.81	0.15	-0.49	0.16	-1.20	0.007	0.019	0.010	0.017	0.020	0.050
10	0.42	-0.29	-0.08	-0.21	-0.40	-0.29	0.002	0.005	0.004	0.006	0.008	0.017
50	0.17	-0.26	-0.12	-0.19	-0.32	-0.28	0.002	0.003	0.003	0.005	0.006	0.013
100	0.08	-0.25	-0.19	-0.18	-0.29	-0.27	0.002	0.003	0.003	0.004	0.005	0.012

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	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
$\alpha = 0.02 \quad \beta = 0.97$								
1	0.007	0.019	0.030	0.107	0.007	0.021	0.922	0.928
10	0.002	0.005	0.002	0.005	0.002	0.005	0.936	0.948
50	0.002	0.003	0.002	0.003	0.002	0.004	0.938	0.937
100	0.002	0.003	0.002	0.003	0.002	0.004	0.943	0.941
$\alpha = 0.05 \quad \beta = 0.93$								
1	0.010	0.017	1.710	12.900	0.010	0.018	0.933	0.945
10	0.004	0.006	0.004	0.006	0.004	0.007	0.938	0.952
50	0.003	0.005	0.003	0.005	0.003	0.005	0.941	0.937
100	0.003	0.004	0.003	0.004	0.003	0.005	0.946	0.943
$\alpha = 0.10 \quad \beta = 0.80$								
1	0.020	0.050	0.020	0.052	0.020	0.051	0.933	0.924
10	0.008	0.017	0.008	0.017	0.008	0.018	0.954	0.944
50	0.006	0.013	0.006	0.013	0.006	0.013	0.968	0.954
100	0.005	0.012	0.006	0.012	0.005	0.012	0.970	0.954

$T \in \{100, 250, 500, 1,000, 2,000\}$ (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 low average bias across all parameter values, with the highest average bias being 0.42% for $\hat{\alpha}$ and -0.29% 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

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.

	$N_T = 1$ (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	1.410	-15.900	-23.600	-15.200	-26.700	-16.200	-27.200	-16.500
250	11.300	-10.600	-1.380	-3.500	-3.460	-2.710	-3.610	-2.610
500	5.340	-4.570	-0.157	-1.150	-0.650	-0.976	-0.695	-0.955
1,000	2.370	-1.530	-0.004	-0.484	-0.221	-0.406	-0.306	-0.393
2,000	0.534	-0.561	-0.076	-0.225	-0.050	-0.185	-0.082	-0.181
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	0.069	0.227	0.030	0.169	0.023	0.163	0.022	0.160
250	0.043	0.186	0.014	0.052	0.010	0.028	0.010	0.021
500	0.026	0.109	0.009	0.016	0.006	0.011	0.006	0.010
1,000	0.016	0.042	0.006	0.010	0.004	0.007	0.004	0.007
2,000	0.010	0.019	0.004	0.006	0.003	0.004	0.003	0.004

0.01%. Moreover, taking $N_T = 1$ as a reference, when the panel size increases to $N_T = 10$ the change in bias is 0.52%, while when the size is increased to $N_T = 100$, bias is reduced by 0.56%. 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 non-vanishing bias as $N_T \rightarrow \infty$, since in some cases there is a slight difference between MCSD and

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 ($R_{\hat{\alpha}}$ and $R_{\hat{\beta}}$) 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.

T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
$N_T=1$ (QMLE)																
100	0.069	0.227	0.111	2.250	0.069	0.272	0.551	0.863	0.030	0.169	0.072	1.670	0.032	0.220	0.791	0.832
250	0.043	0.186	0.187	7.750	0.043	0.210	0.842	0.866	0.014	0.052	0.017	0.129	0.014	0.061	0.911	0.920
500	0.026	0.109	0.028	0.143	0.026	0.117	0.889	0.905	0.009	0.016	0.009	0.018	0.009	0.019	0.931	0.937
1,000	0.016	0.042	0.018	0.057	0.016	0.044	0.909	0.916	0.006	0.010	0.006	0.010	0.006	0.011	0.934	0.945
2,000	0.010	0.019	0.011	0.021	0.010	0.019	0.923	0.932	0.004	0.006	0.004	0.006	0.004	0.007	0.940	0.948
$N_T=50$																
100	0.023	0.163	0.283	5.150	0.026	0.222	0.786	0.759	0.022	0.160	0.078	5.810	0.026	0.222	0.769	0.718
250	0.010	0.028	0.010	0.026	0.010	0.037	0.912	0.897	0.010	0.021	0.010	0.026	0.010	0.032	0.908	0.877
500	0.006	0.011	0.006	0.011	0.006	0.014	0.930	0.926	0.006	0.010	0.006	0.011	0.006	0.014	0.934	0.926
1,000	0.004	0.007	0.004	0.007	0.004	0.008	0.933	0.936	0.004	0.007	0.004	0.007	0.004	0.007	0.932	0.938
2,000	0.003	0.004	0.003	0.005	0.003	0.005	0.942	0.952	0.003	0.004	0.003	0.004	0.003	0.004	0.947	0.950
$N_T=100$																

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 $T \in \{100, 250, 500, 1,000, 2,000\}$. 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

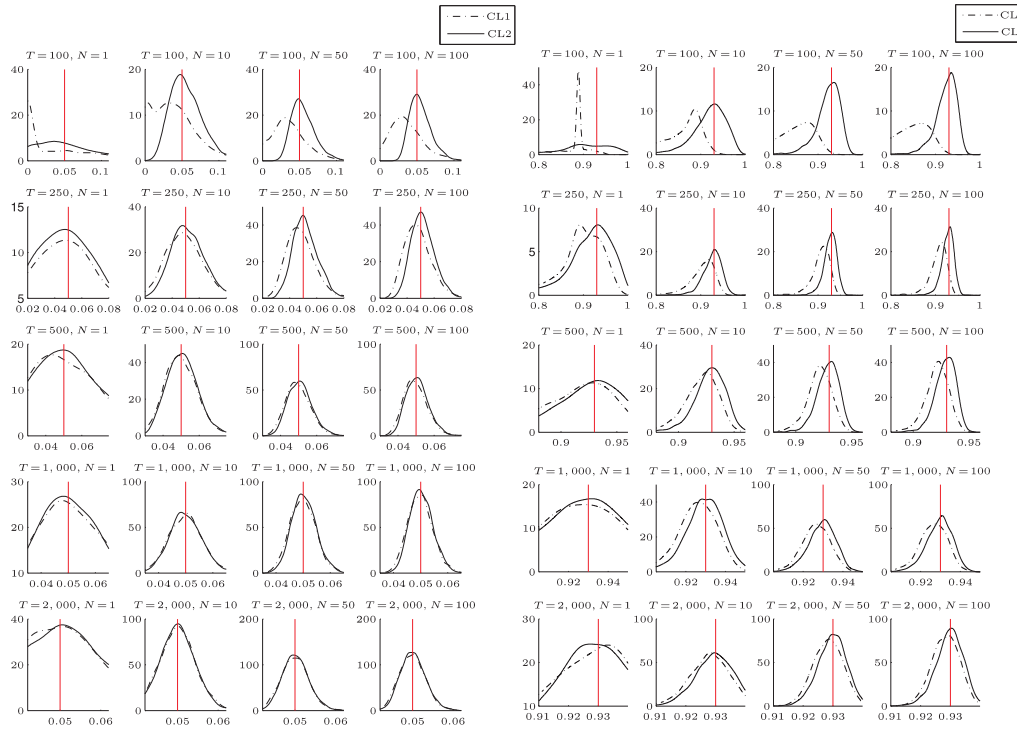


Figure 1. Sample distribution graphs for $\hat{\alpha}$ (left) and $\hat{\beta}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. CL1 gives the sample distribution for the case with nuisance parameter estimation, whereas CL2 is the sample distribution for the case where nuisance parameter 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.

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. In this study, median is used as the blend function (*MG-Median*).

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

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

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 study also analyses the implications of ignoring $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ when $\hat{\alpha}$ 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

$$L_{t+\tau} \left(Y_{t+\tau}, \hat{Y}_{t+\tau} \right). \tag{4.1}$$

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

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

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

$$H_0 : \text{E} \left[L_{t+\tau}(Y_{t+\tau}, f_t(\hat{\beta}_{1t})) - L_{t+\tau}(Y_{t+\tau}, g_t(\hat{\beta}_{2t})) \mid \mathcal{G}_t \right] = 0, \tag{4.2}$$

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 (4.2), 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 et al. (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 3 March 2000 to 12 January 2008 which provides 2,200 observations on each stock. 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 one-step ahead forecasts. To cover a variety of cases, different in-sample sizes, m , are considered (in-sample corresponds to the part of data which is used for estimation of the parameters. Then, $n = T - m$, where T is the total sample size, gives the size of the out-of-sample which is the portion

of the sample that is being forecast). Three different comparisons are analysed: CL vs MG-Median, CL vs QMLE and MG-Median vs 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.

The level of significance for all tests is equal to 5%. Starting values of α and β for optimisation (\bar{a} and \bar{b} , respectively) are generated randomly using $\bar{a} + \bar{b} \sim U(0.5, 0.99)$ and $\bar{a}/(\bar{a} + \bar{b}) \sim U(0.01, 0.3)$. 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. 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.

Table 5 presents results of the conditional and unconditional tests, where the pool of estimates used to calculate the MG estimator (referred to as Md henceforth, as median is used as the blend function) contain cases where $\hat{\alpha} \approx 0$. As mentioned previously, we consider an estimate of α approximately equal to 0 if $\hat{\alpha} < 0.0025$. One might argue that the researcher would not hesitate to eliminate such cases from the pool, as they imply that $\hat{\beta}$ is not identified. However, it must be remembered that for more complex models, identification conditions will not be as straight-forward as in our case. In other words, the Md method has the additional difficulty that the researcher has to characterise cases where a given estimate should be eliminated from the pool. Therefore, to reflect this issue and the ad-hoc nature of MG, this analysis does not eliminate $\hat{\theta}$ where $\hat{\alpha} \approx 0$ from the pool. Results for the situation where such cases are eliminated are presented in Table 6, in order to illustrate the effect of the elimination procedure.

Table 5 reveals that both conditional and unconditional approaches exhibit similar patterns. Comparing CL and Md, it is evident that whenever their predictive abilities can be distinguished, the GW test decides in favour of CL almost all the time. Results for the same comparison in Table 6 show that excluding cases where $\hat{\alpha} \approx 0$ from the pool leads to a decrease in the number of rejections of equal predictive ability. More importantly, in this case, Md is clearly superior

to CL when $m = 250$. Clearly, excluding problematic estimates does help Md to achieve better performance. Neither CL nor QMLE enjoy such a luxury as neither is based on a pool of estimates. Without much doubt, when m is as small as 250 both CL and QMLE perform poorly in estimating the parameters of interest. However, by eliminating the very worst estimates, Md attains better performance. A parallel improvement in Md's performance against QMLE can be observed, as well. Nevertheless, it must be noted that this advantage of MG crucially depends on the ability to characterise "problematic" cases, which will be more difficult when using complex models.

Another observation for the comparison between CL and Md is that the highest number of rejections of equal predictive ability are achieved for $m \in \{500, 750\}$, while the number of rejections fall (one could claim, abruptly) when $m = 1,000$. There are two messages here. First, it can be argued that as the in-sample size increases, both methods start to perform equally well, and hence, it is more difficult to distinguish between them. Second, although it is now more difficult to distinguish between the two methods, CL still is preferred to Md (see $T = 2,200$ and $m = 1,000$).

CL's performance against QMLE is more ambiguous. Rejections in favour of CL in percentages increase as m decreases. When m is small, QMLE is not expected to perform well as the number of observations is not large enough to successfully fit conditional volatility. CL, on the other hand, utilises cross-sectional information, as well, which would explain its higher success rate when m is small. However, as m increases, CL's relative superiority over QMLE deteriorates, as QMLE has more data to estimate the parameters of interest. The advantage of using either pooling method is the ability to base estimation on a larger number of observations. However, both CL and MG have an important cost, which is assuming that parameters of interest are common to all series. CL has an additional cost, as it will suffer from the incidental parameter issue when m is small. Interestingly enough, when m is small, CL does not seem to be badly affected by the incidental parameter issue as it still performs well against QMLE. In other words, in this particular exercise, the benefits of using the panel structure outweigh the damage dealt by the incidental parameter issue and the likely violation of a restrictive assumption. It is very encouraging that, even when QMLE is expected to work well (when, for example, $m = 1,000$), CL is still preferred to QMLE more than 50% of the time.

The performance of MG against QMLE is relatively worse compared to that of CL against QMLE, especially as m decreases. Comparison of Tables 5 and 6 reveals the remarkable effect of removing $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ when $\hat{\alpha} \approx 0$ from the pool. This leads to a dramatic improvement in Md's performance against QMLE at small in-sample sizes. However, it is interesting that even when such cases are

Table 5. Unconditional and conditional Giacomini-White test results at 5% level of significance. Based on one-step ahead forecasts obtained by using the GARCH panel (CL), MacGyver with median as the blend function (Md) and the Quasi Maximum Likelihood (QMLE) methods. Cases where the estimates are close to the starting values for optimisation have been excluded from the pool of estimates used by Md, while cases where $\hat{\alpha} < 0.0025$ have been retained in the pool. Based on daily returns for 94 stocks from S&P100. T gives the total sample size while m gives the in-sample size, implying that the out-of-sample size is equal to $T - m$. For each of the three comparisons, Rej gives the number of cases (out of a total of 94) where equal predictive ability is rejected. Columns 4, 6 and 8 show the percentage of these rejections in favour of CL, CL and Md, respectively. For example, when comparing CL and QMLE at $T = 2,200$ and $m = 500$, in 24 out of 94 tests of unconditional predictive ability, the hypothesis of equal predictive ability is rejected, where 75% of these rejections are in favour of CL.

Unconditional

T	m	CL vs Md		CL vs QMLE		Md vs QMLE	
		Rej	CL (%)	Rej	CL (%)	Rej	Md (%)
2,200	1,000	17	94.12	26	57.69	25	48.00
2,200	750	36	100.00	29	62.07	27	44.44
2,200	500	37	100.00	24	75.00	16	31.25
2,200	250	26	96.15	21	80.95	18	61.11

Conditional

T	m	CL vs Md		CL vs QMLE		Md vs QMLE	
		Rej	CL (%)	Rej	CL (%)	Rej	Md (%)
2,200	1,000	21	95.24	27	51.85	26	42.31
2,200	750	40	95.00	30	60.00	25	32.00
2,200	500	37	100.00	24	62.50	20	40.00
2,200	250	30	100.00	19	84.21	13	53.85

not eliminated, Md is still doing well against QMLE at $m = 250$. At first sight, it might seem strange that an estimator based on a pool of QMLE estimates, which are expected to be very poor when $m = 250$, performs distinctively better than QMLE. However, this is not surprising: the sample distributions of QMLE estimates at each point in time are so dispersed that although the estimates are individually very poor in general, their median still delivers a reasonable value, instead of $\hat{\alpha} = 0$.

We finally note that our analysis of 1-week and 2-week ahead forecasts, available upon request, exhibit a similar pattern.

In conclusion, empirical analysis results suggest that CL delivers better forecasting performance than MG at all in-sample sizes, while it stands out as a good alternative to QMLE even when m is large. Although it can be argued that

Table 6. Unconditional and conditional Giacomini-White test results at 5% level of significance. Cases where $\hat{\alpha} < 0.0025$ and the estimates are close to the starting values for optimisation have been excluded from the pool of estimates used by Md. See Table 5 for more details.

Unconditional					
T	m	CL vs Md		Md vs QMLE	
		Rej	CL (%)	Rej	Md (%)
2,200	1,000	15	93.33	24	45.83
2,200	750	28	100.00	27	48.15
2,200	500	12	91.67	15	60.00
2,200	250	16	6.25	17	94.12

Conditional					
T	m	CL vs Md		Md vs QMLE	
		Rej	CL (%)	Rej	Md (%)
2,200	1,000	17	82.35	25	40.00
2,200	750	32	93.75	26	38.46
2,200	500	14	78.57	18	61.11
2,200	250	13	15.38	15	86.67

MG is not decisively beaten by QMLE either, it is clear that CL's performance against QMLE is superior to that of MG. This is an important result. Both pooling methods are based on the assumption that parameters of interest are common to all assets. This is almost certainly violated here. However, despite that, forecast analysis results suggest that CL still has something to offer. This can be explained by the pooling mechanism of CL which is based on utilising both temporal and cross-sectional information to obtain a single estimate. MG, on the other hand, is still based on information in the time-series only and, therefore, uses observations in a piece-wise fashion, rather than combining all observations in a single pseudo-likelihood function. Nevertheless, it must be noted that a better understanding of the empirical performance of CL requires a complete analysis which is beyond the scope of this study.

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, CL performs well against QMLE and is not beaten decisively by QMLE even when the sample size is large.

Acknowledgements

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

References

- Andersen, T. and Benzoni, L. (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., Bollerslev, T., Diebold, F. X. and Labys, P. (2001). The distribution of exchange rate volatility. *J. Amer. Statist. Assoc.* **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 Honore, B. (2001). Panel data models: some recent developments. In *Handbook of Econometrics*, **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 Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *J. Roy. Statist. Soc. Ser. B* **64**, 253-280.
- Bauwens, L., Laurent, S. and Rombouts, J. V. K. (2006). Multivariate GARCH models: a survey. *J. Appl. Econometrics* **21**, 79-109.
- Bauwens, L. and Rombouts, J. V. K. (2007). Bayesian clustering of many GARCH models. *Econometric Rev.* **26**, 365-386.
- Bollerslev, T., Engle, R. F. and Nelson, D. B. (1994). ARCH models. In *The Handbook of Econometrics*, **4** (Edited by R. F. Engle and D. McFadden), 2959-3038. North-Holland, Amsterdam.
- Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *J. Business Economic Statist.* **13**, 253-263.
- Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

- Engle, R. F. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. *J. Business Economic Statist.* **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 Mezrich, J. (1996). GARCH for groups. *Risk*, **9**, 36-40.
- Engle, R. F. and Sheppard, K. K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Working paper.
- Engle, R. F., Shephard, N. and Sheppard, K. K. (2008). Fitting vast dimensional time-varying covariance models. Working paper.
- Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545-1578.
- Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models. *J. Econometrics* **131**, 97-121.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *J. Econometrics* **95**, 391-413.
- Liang, K.-Y. and Tsou, D. (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. *Ann. Statist.* **13**, 914-931.
- Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (Edited by N. U. Prabhu), 221-239. American Mathematical Society, Providence, RI.
- Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *The Handbook of Econometrics* **4** (Edited by R. F. Engle and D. McFadden), 2111-2245. North-Holland, Amsterdam.
- Newey, W. K. and West, K. D. (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703-708.
- Neyman, J. and Scott, E. L. (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 *J. Econometrics*.
- Patton, A. J. and Sheppard, K. K. (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 Teräsvirta, T. (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. *Adv. Statist. Anal.* **92**, 1-28.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statist. Sinica* **21**, 5-42.
- West, K. (1996). Asymptotic inference about predictive ability. *Econometrica* **64**, 1067-1084.

Department of Economics, University of Oxford & Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK.

E-mail: cavit.pakel@economics.ox.ac.uk

Oxford-Man Institute, University of Oxford, Eagle House, Walton Well Road, Oxford OX2 6ED,
UK & Department of Economics, University of Oxford.

E-mail: neil.shephard@economics.ox.ac.uk

Oxford-Man Institute, University of Oxford, Eagle House, Walton Well Road, Oxford OX2 6ED,
UK & Department of Economics, University of Oxford.

E-mail: kevin.sheppard@economics.ox.ac.uk

(Received October 2009; accepted July 2010)