Fitting vast dimensional time-varying covariance models*

Robert F. Engle
Stern Business School, New York University,
44 West Fourth Street, New York, NY 10012-1126, USA
rengle@stern.nyu.edu

Neil Shephard
Oxford-Man Institute, University of Oxford,
Blue Boar Court, 9 Alfred Street, Oxford OX1 4EH, UK
&
Department of Economics, University of Oxford
neil.shephard@economics.ox.ac.uk

Kevin Sheppard
Department of Economics, University of Oxford,
Manor Road Building, Manor Road, Oxford, OX1 3UQ, UK
&
Oxford-Man Institute, University of Oxford
kevin.sheppard@economics.ox.ac.uk

September 24, 2008

Abstract

Building models for high dimensional portfolios is important in risk management and asset allocation. Here we propose a novel and fast way of estimating models of time-varying covariances that overcome an undiagnosed incidental parameter problem which has troubled existing methods when applied to hundreds or even thousands of assets. Indeed we can handle the case where the cross-sectional dimension is larger than the time series one. The theory of this new strategy is developed in some detail, allowing formal hypothesis testing to be carried out on these models. Simulations are used to explore the performance of this inference strategy while empirical examples are reported which show the strength of this method. The out of sample hedging performance of various models estimated using this method are compared.

Keywords: ARCH models; composite likelihood; dynamic conditional correlations; incidental parameters; quasi-likelihood; time-varying covariances.

*We thank Tim Bollerslev and Andrew Patton for their comments on a previous version of this paper.

1 Introduction

The estimation of time-varying covariances between the returns on hundreds of assets is a key input in modern risk management. Typically this is carried out by calculating the sample covariance matrix based on the last 100 or 250 days of data or through the RiskMetrics exponential smoother. When these covariances are allowed to vary through time using ARCH-type models, the computational burden of likelihood based fitting is overwhelming in very large dimensions, while the usual two step quasi-likelihood estimators of the dynamic parameters indexing them can be massively biased due to an undiagnosed incidental parameter problem even for very simple models. In this paper we introduce novel econometric methods which sidestep both of these issue allowing richly parameterised ARCH models to be fit in vast dimensions, which potentially can be much larger than the time series dimension.

Early work on time-varying covariances in large dimensions was carried out by Bollerslev (1990) in his constant correlation model, where the volatilities of each asset were allowed to vary through time but the correlations were time invariant. This has been shown to be empirically problematic by, for example, Tse (2000) and Tsui and Yu (1999). A survey of more sophisticated models is given by Bauwens, Laurent, and Rombouts (2006) and Silvennoinen and Terasvirta (2008), while Engle (2008a) reviews the topic.

The only econometric work that we know of which allows correlations to change through time in vast dimensions is that of RiskMetrics by J.P. Morgan released in 1994, the DECO model of Engle and Kelly (2007) and the MacGyver estimation method of Engle (2008b)[1]. Engle and Kelly (2007) assume that the correlation amongst assets changes through time but is constant across the cross-section of K assets, an assumption that allows the log-likelihood to be computed in $O(K)$ calculations, which is highly convenient. However, this equicorrelation model is quite restrictive since the diversity of correlations is often the key to risk management.

The RiskMetrics estimator of the conditional covariance matrix is parameter free and has the structure of an integrated GARCH type model but applied to outer products of daily returns. Formally this is a special case of the scalar BEKK process discussed by Engle and Kroner (1995). It has been widely used in industry and was until recently the only viable method that had been suggested which could be applied in hundreds of dimensions.

An alternative method was suggested by Engle (2008b) where he fit many pairs of bivariate estimators, governed by simple dynamics, and then took a median of these estimators. This method is known as the MacGyver estimation strategy, but it requires $O(K^2)$ calculations, is not invariant

[1]The MacGyver method is related to the estimation theory of Chen, Jacho-Chavez, and Linton (2007), which studies the theory of estimators which are average of other estimators in a cross-sectional data setup.

to reparameterisation and formalising this method in order to conduct inference is difficult. Our method has some similarities to the MacGyver strategy but is more efficient and is invariant.

A further set of papers have been written which advocate methods which can be used on moderately high dimensional problems, such as 50 assets. The first was the covariance tracking and scalar dynamics BEKK model of Engle and Kroner (1995), the second was the DCC model of introduced by Engle (2002) and studied in detailed by Engle and Sheppard (2001) — recent developments in this area include Aielli (2006), Engle (2008a) and Pesaran and Pesaran (2007). When these methods have been implemented in practice, they always use a two stage estimation strategy which removes an enormously high dimensional nuisance parameter using a method of moments estimator and then maximises the corresponding quasi-likelihood function. We will show that even if we could compute the quasi-likelihood function for these models in 100s of dimension, the incidental parameter problem causes quasi-likelihood based inference to have economically important biases in the estimated dynamic parameters.

Our approach is to construct a type of composite likelihood, which we then maximise to deliver our preferred estimator. The composite likelihood is based on summing up the quasi-likelihood of subsets of assets. Each subset yields a valid quasi-likelihood, but this quasi-likelihood is only mildly informative about the parameters. By summing over many subsets we can produce an estimator which has the advantage that we do not have to invert large dimensional covariance matrices. Further and vitally it is not effected by the incidental parameter problem. It can also be very fast — it can be $O(1)$ if needed and does not have the biases intrinsic to the usual quasi-likelihood when the cross-section is large.

A special case of our estimation strategy is used in Fast-GARCH model (Bourgoin (2002)). Fast-GARCH estimates a single univariate GARCH model for one asset, and then combines this estimate with the sample variance of the returns to fit a variance-targeted model using the method of Engle and Mezrich (1996).

The approach we advocate here can also be used in the context of more structured models, which impose stronger a priori constraints on the model. Factor models with time-varying volatility are the leading example of this, where leading papers include King, Sentana, and Wadhwani (1994), Harvey, Ruiz, and Sentana (1992), Fiorentini, Sentana, and Shephard (2004) and Chib, Nardari, and Shephard (2006). Our approach allows us to impose a factor structure on the models if this is desirable.

The structure of the paper is as follows. In Section 2 we outline the model and discuss alternative general methods for fitting time-varying covariance models. We also discuss the usual use of covariance tracking, which helps us in the optimisation of the objective functions discussed in this

paper. In Section 3 we discuss the core of the paper, where we average in different ways the results from many small dimensional "sub"-models in order to carry out inference on a large dimensional model. We show this method has a hidden incidental parameter problem and that the use of composite likelihoods largely overcomes this problem. Section 4 provides a Monte Carlo investigation comparing the finite sample properties of our estimator with the usual quasi-maximum likelihood. Section 5 illustrates our estimator on 95 components of the S&P 100, finding evidence of both qualitative and quantitative differences. We extend this analysis to cover 480 components of the S&P 500. In Section 6 we discuss some important additional topics. Section 7 concludes, while the Appendix contains some derivations and further observations of interest.

2 The model and the usual quasi-likelihood

2.1 Framework

We write a K-dimensional vector of log-returns as r_t where $t = 1, 2, ..., T$. A typical risk management model of r_t given the information available at time t is to assume:

Assumption 1

$$\mathrm{E}(r_t|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(r_t|\mathcal{F}_{t-1}) = H_t, \tag{1}$$

where $\mathcal{F}_{t-1}$ is the information available at time $t-1$ to predict r_t.

Thus r_t is a $\mathcal{F}$-martingale difference sequence with a time-varying covariance matrix. We will model how H_t depends upon the past data allowing it to be indexed by some parameters $\psi \in \Psi$. We intend to estimate ψ. For simplicity in our examples we have always used single lags in the dynamics. The extension to multiple lags is trivial but rarely used in empirical work.

Example 1 *Scalar BEKK. This puts*

$$H_t = (1 - \alpha - \beta)\,\Sigma + \alpha r_{t-1} r_{t-1}' + \beta H_{t-1}, \quad \alpha \geq 0, \quad \beta \geq 0, \quad \alpha + \beta < 1,$$

which is a special case of Engle and Kroner (1995). Typically this model is completed by setting $H_1 = \Sigma$. Hence in this model $\psi = \left(\lambda', \theta'\right)'$, where $\lambda = vech(\Sigma)$ and $\theta = (\alpha, \beta)'$.

Example 2 *Nonstationary covariances with scalar dynamics:*

$$H_t = \alpha r_{t-1} r_{t-1}' + (1 - \alpha) H_{t-1}, \quad \alpha \in [0, 1)\,.$$

A simple case of this is RiskMetrics, which puts $\alpha = 0.06$ for daily returns and 0.03 for monthly returns. Inference for this EWMA model is usually made conditional on $\lambda = vech(H_0)$, which has to be estimated, while $\theta = \alpha$ and $\psi = \left(\lambda', \theta'\right)'$.

The standard inference method is based on a Gaussian quasi-likelihood

$$\log L(\psi; r) = \sum_{t=1}^{T} l_t(\psi), \tag{2}$$

where

$$l_t(\psi) = -\frac{1}{2}\log|H_t| - \frac{1}{2} r_t{}' H_t^{-1} r_t.$$

Maximising this quasi-likelihood (2) directly in high-dimension models is difficult since

- the parameter space is typically large, which causes numerical and statistical challenges;
- each of the T inversions of H_t takes $O(K^3)$ computations per likelihood evaluation[2].

This paper will show how to side-step these two problems.

2.2 Nuisance parameters

In Example 1, Σ has to be estimated along with the dynamic parameters of interest α and β. Σ has $K(K+1)/2$ free parameters, which will be vast if K is large. Similar issues arise in a large number of multivariate models.

More abstractly we write the dynamic parameters of interest as θ and the nuisance parameters as λ whose dimension is P. Then the quasi-likelihood is

$$\log L(\theta, \lambda; r).$$

Often we can side step the optimising over λ by concentrating at some moment based estimator $\widetilde{\lambda}$.

Example 3 *For Example 1 Engle and Mezrich (1996) suggested putting* $\widetilde{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} r_t r_t'$, *then* $\widetilde{\lambda} = vech(\widetilde{\Sigma})$. *This is called covariance tracking. For Example 2 one can put* $\widetilde{H}_0 = \frac{1}{T}\sum_{t=1}^{T} r_t r_t'$ *and* $\widetilde{\lambda} = vech(\widetilde{H}_0)$.[3]

[2] In modern software packages, matrix inversion is implemented as a series of matrix multiplications. As a result, the complexity of the matrix multiplication is the dominant term when computing a matrix inverse. By direct inspection the multiplication of $K \times K$ matrices can be easily seen to be no worse than $O(K^3)$. This is because K rows must be paired with K columns, and each dot product involves K multiplications and $K-1$ additions, or $2K-1$ computations. Most common implementations are $O(K^3)$ although faster, but somewhat unstable inversions can be computed in $O(K^{\log_2 7}) \approx O(K^{2.81})$ or faster (Strassen (1969)).

In practice we have also found that when estimating models of dimension 100 or more then great care needs to be taken with the numerical precision of the calculation of the inverse and determinant in (2) in order to achieve satisfactory results when optimising over ψ.

[3] When we use quasi-likelihood estimation to determine α in the EWMA model a significant problem arises when K is large for $\widetilde{\alpha}$ will be forced to be small in order that the implied H_t has full rank — for a large α and large K will imply H_t is singular. This feature will dominate other ones and holds even though element by element the conditional covariance matrix will very poorly fit the data.

We then maximise to deliver the m-profile[4] quasi-likelihood estimator (MMLE)

$$\widetilde{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\theta, \widetilde{\lambda}; r).$$

When K is small compared to T then inference can be thought of as a two stage GMM problem, whose theory is spelt out in, for example, Newey and McFadden (1994) and Engle and Sheppard (2001). All this is well known.

Unfortunately when K is large the dimension of λ is also large, and so estimating λ can mean $\widetilde{\theta}$ is thrown far from its true value. This generic statistical characteristic has been known since the work of, for example, Neyman and Scott (1948) and Nickell (1981). There are some hints that this might be a problem in the multivariate volatility literature. Engle and Sheppard (2001) report that for their DCC models, which we will discuss in Section 3.7, some of their quasi-likelihood based estimated dynamic parameters seem biased when K is moderately large in Monte Carlo experiments.

2.3 Empirical illustration

Here we estimate the models given in Examples 1 and 2 (and the DCC model discussed in Section 3.7) using data for all companies at one point listed on the S&P 100, plus the index itself, over the period January 1, 1997 until December 31, 2006 taken from the CRSP database. This database has 124 companies although 29, for example Google, have one or more periods of non-trading, (e.g. prior to IPO or subsequent to an acquisition). Selecting only the companies that have returns throughout the sample reduced this set to 95 (+1 for the index). This means $T = 2,516$ and $K \leq 96$. To allow K to increase, which allows us to assess the sensitivity to K, we set the first asset as the market and the other assets are arranged alphabetically by ticker[5]. The results for fitting the two models using $\widetilde{\lambda}$ are given in Example 3. The estimated θ parameters from an expanding cross-section of assets are contained in Table 1.

The empirical results suggest the increasing K destroys the MMLE as $\widetilde{\alpha}$ falls dramatically as K increases. These results will be confirmed by detailed simulation studies in Section 4 which produce the same results by simulating BEKK or DCC models and then estimating them using MMLE techniques. In addition Section 5 suggests the MMLE parameter values when $K = 96$ are poor when judged using a simple economic criteria.

These results are reinforced by an empirical study based exactly the same type of database, but now based on the corresponding components of the S&P 500. Including the index this produces a

[4]Although at first sight $l(\theta, \widehat{\lambda})$ looks like a profile likelihood, it is not as $\widehat{\lambda}$ is not a maximum quasi-likelihood estimator but an attractive moment estimator. Hence we call it a moment based profile likelihood, or m-profile likelihood for short. This means $\widehat{\theta}$ is typically less efficient than the maximum quasi-likelihood estimator.

[5]For stocks that changed tickers during the sample, the ticker on the first day of the sample was used

	S&P 100 Components						S&P 500 Components				
	Scalar BEKK		EWMA	DCC			Scalar BEKK		DCC	
K	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$	K	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
5	.0189	.9794	.0134	.0141	.9757	5	.0261	.9715	.0101	.9823
10	.0125	.9865	.0103	.0063	.9895	25	.0080	.9909	.0030	.9908
25	.0081	.9909	.0067	.0036	.9887	50	.0055	.9932	.0018	.9882
50	.0056	.9926	.0045	.0022	.9867	100	.0034	.9934	.0015	.9524
96	.0041	.9932	.0033	.0017	.9711	250	.0015	.9842	.0020	.5561
						480	.0032	.5630	.0013	.2556

Table 1: *Parameter estimates from a covariance targeting scalar BEKK, EWMA (estimating H_0) and DCC using maximum m-profile likelihood (MMLE). Based upon a real database built from daily returns from 95 companies plus the index from the S&P100, from* 1997 *until* 2006. The same analysis is also reported on 480 components from the S&P 500 over the same time period.

dataset with $K = 480$. The results in Table 1 show dramatic distortions — where the estimated β also crash towards zero as K increases.

We now turn to our preferred estimator which allows K to have any relationship to T, yielding consistency as $T \to \infty$. In particular, the estimator will work even when K is larger than T.

3 The main idea: composite-likelihood

3.1 Many small dimensional models

To progress it is helpful to move the return vector r_t into a data array $Y_t = \{Y_{1t}, ..., Y_{Nt}\}$ where Y_{jt} is itself a vector containing small subsets of the data (there is no requirement for the Y_{jt} to have common dimensions)

$$Y_{jt} = S_j r_t,$$

where S_j as non-stochastic selection matrix. In our context the leading example is where we look at all the unique "pairs" of data

$$\begin{aligned} Y_{1t} &= (r_{1t}, r_{2t})', \\ Y_{2t} &= (r_{1t}, r_{3t})', \\ &\vdots \\ Y_{\frac{K(K-1)}{2}t} &= (r_{K-1t}, r_{Kt})', \end{aligned}$$

where $N = K(K-1)/2$. Our model (1) trivially implies

$$\mathrm{E}(Y_{jt}|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(Y_{jt}|\mathcal{F}_{t-1}) = H_{jt} = S_j H_t S_j'. \tag{3}$$

Then a valid quasi-likelihood can be constructed for ψ off the j-th subset

$$\log L_j(\psi) = \sum_{t=1}^{T} l_{jt}(\psi), \quad l_{jt}(\psi) = \log f(Y_{jt}; \psi)$$

where

$$l_{jt}(\psi) = -\frac{1}{2}\log|H_{jt}| - \frac{1}{2}Y_{jt}'H_{jt}^{-1}Y_{jt}.$$

This quasi-likelihood will have information about ψ but more information can be obtained by averaging[6] the same operation over many submodels

$$c_t(\psi) = \frac{1}{N}\sum_{j=1}^{N}\log L_{jt}(\psi).$$

Of course if the $\{Y_{1t}, ..., Y_{Nt}\}$ were independent this would be the exact likelihood — but this will not be the case for us. Such functions, based on "submodels" or "marginal models", are call composite likelihoods (CLs), following the nomenclature introduced by Lindsay (1988)[7]. See Varin (2008) for a review. Summing over the time series we have the sample CL function

$$CL(\psi) = \sum_{t=1}^{T} c_t(\psi).$$

For fixed N, as T increases to infinity, the estimator which maximises this CL, written $\widehat{\psi}$, has well known asymptotic properties as the CL is a particular form of a quasi-likelihood. Appropriate references include Cox (1961), Eicker (1967), White (1982) and Gallant and White (1988).

Evaluation of $c_t(\psi)$ costs $O(N)$ calculations. In the case where all distinct pairs are used this means the CL costs $O(K^2)$ calculations — which is distinctively better than the $O(K^3)$ implied by (2). One can also use the subset of contiguous pairs $\{r_{jt}, r_{j+1t}\}$, which would be $O(K)$, or an economically motivated selection like the so called "beta CL" discussed in Section 6.3 which is also $O(K)$ and is based on using all pairs involving the market index returns. Some of the computational considerations are illustrated in Table 2 which show some computational times for a problem based on modelling up to 480 assets. A detailed discussion of this will be given in Section 5.4.

An alternative is to choose only $O(1)$ pairs, which is computationally faster. It is tempting to randomly select N pairs and make inference conditional on the selected pairs as the selection is

[6]It may make sense to also define the weighted CL

$$\frac{1}{N}\sum_{j}^{N} w_{jt}\log L_{jt}(\theta),$$

where $w_{j,t}$ are non-negative weights determined by the economic importance of the subset of assets, e.g. making the weights proportional to the geometric average of the asset's market value. The weights can be allowed to vary through time, but this variation should depend at time t solely on functions of $\mathcal{F}_{t-1}$. This weighting add little complexity to the asymptotic theory of the weighted CL.

[7]This type of marginal analysis has appeared before in the non-time series statistics literature. An early example is Besag (1974) in his analysis of spatial processes, more recently it was used by Fearnhead (2003) in bioinformatics, deLeon (2005) on grouped data, Kuk and Nott (2000) and LeCessie and van Houwelingen (1994) for correlated binary data. This type of objective function is sometimes call CL methods, following the term introduced by Lindsay (1988), or "subsetting methods". See Varin and Vidoni (2005). Cox and Reid (2003) discusses the asymptotics of this problem in the non-time series case.

	MMLE	All Pairs		Contiguous Pairs	
K	m-profile	maximum	m-profile	maximum	m-profile
5	24s	1.4s	0.1s	0.6s	0.2s
25	46s	31s	2.1s	2.3s	0.2s
50	2m 10s	2m 11s	10s	5.1s	0.5s
100	1h 50m	14m 51s	39s	16s	0.8s
250	15h 11m	1h 0m	4m 7s	29s	1.6s
480	85h 33m	3h 39m	18m 6s	50s	4.5s

Table 2: *CPU time required to estimate a covariance targeting scalar BEKK on the assets of the S&P 500. All models were estimated on a 2.5GHz Intel Core 2 Quad.*

strongly exogenous. We will see in a moment that the efficiency loss of using only $O(1)$ subsets compared to computing all possible pairs can be extremely small.

Using a CL reduces the computational challenges in fitting very large dimensional models. We now turn our attention to the statistical implications.

3.2 Many small dimensional nuisance parameters

We now make our main assumption that

$$c_t(\psi) = \frac{1}{N}\sum_{j=1}^{N} \log L_{jt}(\theta, \lambda_j),$$

that is it is possible to write the CL in terms of the common finite dimensional θ and then a vector of parameters λ_j which is specific to the j-th pair. Our interest is in estimating θ and so the λ_j are nuisances. As N increases then so does the number of nuisance parameters. This type of assumption appeared, outside the CL, first in the work of Neyman and Scott (1948), which has been highly influential in econometrics[8]. In that literature this is sometimes named a stratified model with a stratum of nuisance parameters and can be analysed by using two-index asymptotics, e.g. Barndorff-Nielsen (1996).

3.3 Parameter space

For the j-th submodel we have the common parameter θ and nuisance parameter λ_j. The joint model (1) may imply there are links across the λ_j.

[8]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 j and t. Then the maximum likelihood estimator of θ is typically inconsistent for finite T and $N \to \infty$ and needs, when T increases, $N = o(T^{1/2})$ for standard distributional results to hold (Sartori (2003)) with rate of convergence $\sqrt{NT}$. 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 $\log L_{jt}(\theta, \lambda_j)$ will be shown to reduce the rate of convergence to rate $\sqrt{T}$, not $\sqrt{NT}$. Under those circumstances we will see the MCLE will be consistent and have a simple limit theory however N relates to T.

Example 4 *The scalar BEKK model of Example 1*

$$Y_{1t} = (r_{1t}, r_{2t})', \quad Y_{2t} = (r_{2t}, r_{3t})',$$

then

$$\lambda_1 = (\Sigma_{11}, \Sigma_{21}, \Sigma_{22})', \quad \lambda_2 = (\Sigma_{22}, \Sigma_{32}, \Sigma_{33})'.$$

Hence, the joint model implies there are common elements across the λ_j.

As econometricians we may potentially gain by exploiting these links in our estimation. An alternative, is to be self-denying and to never use these links, even if they exist in the data generating process. The latter means the admissible values are

$$(\lambda_1, \lambda_2, ..., \lambda_N) \in \Lambda_1 \times \Lambda_2 \times ... \times \Lambda_N, \tag{4}$$

i.e. they are variation-free (e.g. Engle, Hendry, and Richard (1983)).

In the context of CLs imposing variation freeness on inference has great conceptual virtues for it allows the estimation to be carried out for λ_j based solely on $Y_{j1}, ..., Y_{jT}$ and the common structure determined by θ. Of course, this approach risks efficiency loss — but not bias. Throughout our paper we will impose variation-free on our estimation strategy (of course inference will be agnostic to it). Our experiments, not reported here, which have used the cross-submodel constraints indicate the efficiency loss in practice of this is tiny when N is large.

Remark 1 *This variation-free structure requires that* λ_j *is identified using the* j*-th submodel's likelihood, given knowledge of* θ*. For many models this will be the case, e.g. an unstructured* Σ *in a scalar BEKK model. If a factor model is impose on* Σ *however, some care needs to be taken that the* $\dim(Y_{jt})$ *is larger than the dimension of the factor.*

3.4 Estimators

Our estimation strategy can be generically stated as solving

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \log L_{jt}(\widehat{\theta}, \widehat{\lambda}_j),$$

where $\widehat{\lambda}_j$ solves for each j

$$\sum_{t=1}^{T} g_{jt}(\widehat{\theta}, \widehat{\lambda}_j) = 0.$$

Here g_{jt} is a $\dim(\lambda_j)$-dimensional moment constraint so that for each j and θ there exists a single $\lambda_{j\theta}$ which solves

$$\mathrm{E}\left\{g_{jt}(\theta, \lambda_{j\theta})\right\} = 0, \quad t = 1, 2, ..., T.$$

This structure has some important special cases.

Example 5 *The maximum CL estimator (MCLE) follows from writing*

$$g_{jt}(\theta, \lambda_j) = \frac{\partial \log L_{jt}(\theta, \lambda_j)}{\partial \lambda_j},$$

so

$$\widehat{\lambda}_j(\theta) = \underset{\lambda_j}{\operatorname{argmax}} \sum_{t=1}^{T} \log L_{jt}(\theta, \widehat{\lambda}_j),$$

which means

$$\frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \log L_{jt}(\theta, \widehat{\lambda}_j)$$

is the profile CL which $\widehat{\theta}$ maximises.

Example 6 *Suppose $G_{jt} = G_{jt}(Y_{jt})$ and*

$$g_{jt}(\theta, \lambda_j) = G_{jt} - \lambda_j, \quad \textit{where} \quad \mathrm{E}(G_{jt}) = \lambda_j,$$

then

$$\widehat{\lambda}_j = \frac{1}{T} \sum_{t=1}^{T} G_{jt}.$$

We call the resulting $\widehat{\theta}$ a m-profile CL estimator (MMCLE).

3.5 Consistency of $\widehat{\theta}$

3.5.1 Statement of the result

In this subsection we will give general conditions under which $\widehat{\theta}$ will be consistent.

Theorem 1 . *Assume the following conditions hold.*

1. *Θ and Λ_j are compact.*

2. *For each $\theta \in \Theta$ there exists a pseudo-true value $\lambda_{j\theta}^* \in \Lambda_j$ which uniquely solves*

$$\mathrm{E}\left\{ g_{jt}\left(\theta, \lambda_{j\theta}^*\right)\right\} = 0.$$

3. *The non-nuisance parameter version of the composite likelihood delivers a consistent estimator, i.e.*

$$\arg\sup_{\theta \in \Theta} \frac{1}{T N_T} \sum_{t=1}^{T} \sum_{j=1}^{N_T} l_{jt}(\theta, \lambda_{j\theta}^*) \xrightarrow{p} \theta^*. \tag{5}$$

4. *For every j and t, $l_{jt}(\theta, \lambda_j)$ is continuously differentiable in $\lambda_j \in \Lambda_j$.*

5. *Define*

$$c_{t,T}(r_t) = \frac{1}{N_T} \sum_{j=1}^{N_T} \sup_{\theta \in \Theta, \lambda_j \in \Lambda_j} \left| \frac{\partial l_{jt}(\theta, \lambda_j)}{\partial \lambda_j} \right|.$$

Assume that

$$\frac{1}{T} \sum_{t=1}^{T} c_{t,T}(r_t)$$

satisfies a weak law large number as $T \to \infty$.

6. *Write $\lambda = \left(\lambda_1', \lambda_2', ..., \lambda_{N_T}'\right)'$ and assume that*

$$\sup_{\theta \in \Theta} \max_{j \in \{1,2,...,N_T\}} \left| \widehat{\lambda}_{j\theta} - \lambda_{j\theta}^* \right| = o_p(1). \tag{6}$$

Then

$\widehat{\theta} \xrightarrow{p} \theta^*$.

Proof. Given in the Appendix.

3.5.2 Discussion

There are two major points to be made about the assumptions in this theorem. First when there are no incidental parameters the composite likelihood takes the form of

$$\frac{1}{T} \sum_{t=1}^{T} l_{t,T}(\theta), \quad \text{where} \quad l_{t,T}(\theta) = \frac{1}{N_T} \sum_{j=1}^{N_T} l_{jt}(\theta, \lambda_{j\theta}^*)$$

which is simply an array version of a standard quasi-likelihood. Hence general quasi-likelihood theory applies to this in a straightforward way and raises no new issues — see White (1994) and Cox and Reid (2003).

Second the condition (6) means that $\sup_{\theta \in \Theta} \max_j \left| \widehat{\lambda}_{j\theta} - \lambda_{j\theta} \right| \xrightarrow{p} 0$ goes to zero as T increases even though N_T can increase with T. For the scalar BEKK and DECC models

$$\sup_{\theta \in \Theta} \left\| \widehat{\lambda}_\theta - \lambda_\theta \right\|_\infty \leq \sup_{\theta \in \Theta} \left\| \widehat{\Sigma}_\theta - \Sigma_\theta \right\|_\infty$$

the long run unconditional covariance of the asset returns. Recall that for a p-dimensional vector z that $\|z\|_\infty = \max\{|z_1|, |z_2|, ..., |z_p|\}$.

In the MMCLE case the dependence on θ can be dropped and we get

$$\sup_{\theta \in \Theta} \left\| \widehat{\lambda}_\theta - \lambda_\theta \right\|_\infty \leq \left\| \widehat{\Sigma} - \Sigma \right\|_\infty, \quad \widehat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} r_t r_t'.$$

Fan, Zhang, and Yu (2008) have studied $\left\|\widehat{\Sigma}-\Sigma\right\|_\infty$ under a variety of regularity assumptions. In particular they show that when r_t and $r_t r_t'$ each have an autoregressive representation, driven by a martingale error terms (plus some additional technical conditions about the memory of these processes) that

$$\left\|\widehat{\Sigma}-\Sigma\right\|_\infty = O_p\left(\sqrt{\frac{\log K}{T}}\right)$$

which shows the size of the problem only impacts the estimator at a logarithmic rate. This implies that the consistency of composite likelihood is largely immune from problems of high dimensionality. This may be a rather conservative result for our Monte Carlo results suggest that K has no impact on the consistency of $\widehat{\theta}$ — proving this result is however beyond this paper.

Finally, trivially, Assumption 4 will hold if r_t is ergodic by an array law of large numbers.

3.6 Central limit theorem for $\widehat{\theta}$

We now turn to some distributional results for this class of estimator, which will be followed by a detailed Monte Carlo study. Throughout our asymptotics will have $T \to \infty$ while the cross-sectional dimension N_T can potentially increase with T.

Theorem 2 . *Assume $\widehat{\theta}$ is consistent. Throughout all functions are evaluated at $(\theta^*, \lambda_{j\theta}^*)$. We assume l_{jt} is twice continuously differentiable and g_{jt} once continuously differentiable. We first define some terms:*

$$\begin{aligned}
F_{j,T} &= \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\lambda_{jt}'}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\lambda_{jt}'}\right)^{-1}, \\
Z_{t,T} &= \frac{1}{N_T}\sum_{j=1}^{N_T}\left(\frac{\partial l_{jt}}{\partial\theta'} - F_{j,T}g_{jt}\right), \\
D_{j,\theta\theta,T} &= \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\partial\theta'}\right) - F_{j,T}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\theta'}\right), \\
D_{\theta\theta,T} &= \frac{1}{N_T}\sum_{j=1}^{N_T} D_{j,\theta\theta,T}.
\end{aligned}$$

Then assume the following:

1. *θ^* is an interior point of Θ.*
2. *$\lambda_{j\theta}^*$ is an interior point of Λ_j.*
3. *If T is large then over all j the smallest eigenvalue of*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\lambda_{jt}'}$$

is bounded above zero.

4. *That as* $T \to \infty$

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} Z_{t,T} \xrightarrow{d} N(0, \mathcal{I}_{\theta\theta}). \tag{7}$$

We also assume that $\mathcal{I}_{\theta\theta}$ *has diagonal elements which are bounded from above and* $\mathcal{I}_{\theta\theta} > 0$.

5. *That as* $T \to \infty$

$$D_{\theta\theta,T} \xrightarrow{p} \mathcal{D}_{\theta\theta} > 0,$$

where $\mathcal{D}_{\theta\theta}$ *is invertible.*

Then

$$\sqrt{T}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N(0, \mathcal{D}_{\theta\theta}^{-1}\mathcal{I}_{\theta\theta}\mathcal{D}_{\theta\theta}^{-1}).$$

Proof. Given in the Appendix.

The most important assumption we will need to produce this results is that $\mathcal{I}_{\theta\theta}$ has diagonal elements which are bounded from above and $\mathcal{I}_{\theta\theta} > 0$. Intuitively it means the average score does not exhibit a law of large numbers in the cross-section.

In order to implement this theory we have to estimate $\mathcal{D}_{\theta\theta}$ and $\mathcal{I}_{\theta\theta}$. The former can be estimated by $D_{\theta\theta,T}$ where we evaluate the functions at estimates rather than the true parameter points. The small dimensional $\mathcal{I}_{\theta\theta}$ is estimated by using a HAC estimator (e.g. Andrews (1991)) applied to $\{Z_{t,T}\}$. Notice the dimension of $Z_{t,T}$ does not vary with N_T.

Example 7 *In the case where* $\widehat{\lambda}_j$ *is a moment estimator, Example 6, then* $g_{jt} = G_{jt} - \lambda_j$, *so*

$$F_{j,T} = -\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\lambda_{jt}'}\right).$$

Remark 2 *We can directly see the effect on the efficiency of this procedure of* N_T *by studying (7). Then for large* N_T

$$\begin{aligned}\operatorname{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{t,T}\right) &= \frac{1}{N_T^2}\sum_{j=1}^{N_T}\operatorname{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{j,t,T}\right) + \frac{1}{N_T^2}\sum_{j\neq k}^{N_T}\operatorname{Cov}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{j,t,T}, \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{k,t,T}\right)\\ &\simeq \frac{1}{N_T^2}\sum_{j\neq k}^{N_T}\operatorname{Cov}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{j,t,T}, \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{k,t,T}\right).\end{aligned}$$

So as N_T *increases it knocks out the variance term in* $Z_{j,t,T}$ *and this drives the gains of using the cross-sectional information. It also shows that there is no expectation, for our applications, that*

as N_T increases that this variance is driven to zero. Instead the limit

$$\mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{t,T}\right) \simeq \lim_{N\to\infty}\frac{1}{N^2}\sum_{j\neq k}^{N}\mathrm{Cov}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{j,t,T}, \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{k,t,T}\right),$$

the average covariance between randomly selected pairs

$$\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{j,t,T}, \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{k,t,T}\right\}$$

for $j \neq k$.

3.7 Extended example: DCC model

The DCC model of Engle (2002) and Engle and Sheppard (2001) allows a much more flexible time-varying covariance model than Examples 1 and 2. Write the submodel based on a pair as

$$Y_{jt} = \{r_{1jt}, r_{2jt}\}, \quad \mathrm{Cov}(Y_{jt}|\mathcal{F}_{t-1}) = \begin{pmatrix} h_{1jt}^{1/2} & 0 \\ 0 & h_{1jt}^{1/2} \end{pmatrix} R_{jt} \begin{pmatrix} h_{1jt}^{1/2} & 0 \\ 0 & h_{1jt}^{1/2} \end{pmatrix},$$

where we construct a model for the conditional variance $h_{ijt} = \mathrm{Var}(r_{ijt}|\mathcal{F}_{t-1}, \eta_{ij})$, which is indexed by the variation free parameters η_{ij}[9]. This has a log-likelihood for the $\{r_{ijt}\}$ return sequence of

$$\log E_{ijt} = -\frac{1}{2}\log h_{ijt} - \frac{1}{2}r_{ijt}^2/h_{ijt}, \quad i = 1, 2.$$

The devolatilised series is defined as

$$S_{jt} = \begin{pmatrix} h_{1jt}^{-1/2} & 0 \\ 0 & h_{1jt}^{-1/2} \end{pmatrix} \begin{pmatrix} r_{1jt} \\ r_{2jt} \end{pmatrix}, \quad \text{so} \quad \mathrm{Cov}(S_{jt}|\mathcal{F}_{t-1}) = R_{jt} = \mathrm{Cor}(Y_{jt}|\mathcal{F}_{t-1}).$$

We build a model for R_{jt} using the cDCC dynamic introduced by Aielli (2006). It is defined as

$$R_{jt} = P_{jt}^{-1/2} Q_{jt} P_{jt}^{-1/2}, \quad P_{jt} = \begin{pmatrix} Q_{11jt} & 0 \\ 0 & Q_{22jt} \end{pmatrix},$$

[9]The first step of fitting the cDCC models is to model $h_{jt} = \mathrm{Var}(r_{jt}|\mathcal{F}_{t-1})$. It is important to note that although it is common to fit standard GARCH models for this purpose, allowing the h_{jt} to depend the lagged squared returns on the j-th asset, in principle $\mathcal{F}_{t-1}$ includes the lagged information from the other assets as well — including market indices. Many of the return series exhibited large moves in volatility during this period. This large increase has been documented by, for example, Campbell, Lettau, Malkeil, and Xu (2001) and appears both in systematic volatility and idiosyncratic volatility. Initial attempts at fitting the marginal volatilities $\mathrm{Var}(r_{jt}|r_{jt-1}, r_{jt-2}, ...)$ included a wide range of "standard" ARCH family models failed residual diagnostics tests for our data.

To overcome this difficulty, a flexible components framework has been adopted which brings in a wider information set. The first component is the market volatility as defined by the index return, $\overline{r}_t = \frac{1}{K}\sum_{j=1}^{K} r_{j,t}$. The volatility was modeled using an EGARCH specification Nelson (1991),

$$\ln h_{\bullet,t} = \omega_\bullet + \alpha_\bullet|\epsilon_{\bullet,t-1} - \sqrt{2/\pi}| + \kappa_\bullet\epsilon_{\bullet,t-1} + \beta_\bullet \ln h_{\bullet,t-1}, \quad \epsilon_{\bullet,t} = \overline{r}_t h_{\bullet,t}^{-1/2}. \tag{8}$$

A second component was included for assets other than the market, resulting in a factor structure for each asset j,

$$\ln \tilde{h}_{j,t} = \omega_j + \alpha_j|\epsilon_{j,t-1} - \sqrt{2/\pi}| + \kappa_j\epsilon_{j,t-1} + \beta_j \ln h_{j,t-1}, \quad h_{j,t} = h_{\bullet,t}\tilde{h}_{j,t}, \quad \epsilon_{j,t} = r_{j,t}h_{j,t}^{-1/2}. \tag{9}$$

This two-component model was able to adequately describe the substantial variation in the level of volatility seen in this panel of returns.

where

$$Q_{jt} = \Psi_j (1 - \alpha - \beta) + \alpha P_{jt-1}^{1/2} \left(S_{jt-1} S_{jt-1}' - R_{jt-1} \right) P_{jt-1}^{1/2} + (\alpha + \beta) Q_{jt-1}, \quad \Psi_j = \begin{pmatrix} 1 & \varphi_j \\ \varphi_j & 1 \end{pmatrix}.$$

It has the virtue that if we let $S_{jt}^* = P_{jt}^{1/2} S_{jt}$, then $\text{Cov}\left(S_{jt}^* | \mathcal{F}_{t-1}\right) = P_{jt}^{1/2} R_{jt} P_{jt}^{1/2} = Q_{jt}$, and so $\frac{1}{T}\sum_{t=1}^{T} S_{jt}^* S_{jt}^{*\prime} \xrightarrow{p} \Psi_j$.

The parameters for this model are $\theta = (\alpha, \beta)'$, $\lambda_j = \left(\eta_{1j}', \eta_{2j}', \varphi_j\right)'$. The corresponding ingredients into the estimation of θ from this model is the common structure

$$\log L_{jt} = -\frac{1}{2} \log |R_{jt}| - \frac{1}{2} S_{jt}' R_{jt}^{-1} S_{jt},$$

while for the j-th submodel

$$g_{jt} = \begin{pmatrix} \frac{\partial \log E_{1jt}}{\partial \eta_{1j}} \\ \frac{\partial \log E_{2jt}}{\partial \eta_{2j}} \\ \frac{1}{T}\sum_{t=1}^{T} S_{1jt}^* S_{2jt}^* - \varphi_j \end{pmatrix}.$$

4 Monte Carlo experiments

4.1 Relative performance of estimators

Here we explore the effectiveness of three estimators of the parameters in the DCC model discussed above,

- maximum m-profile likelihood based estimator (MMLE), based on the quasi-likelihood in Section 2;
- maximum m-profile CL based estimator (MCLE), using all the pairs to construct the CL as in Section 3;
- maximum m-profile subset CL estimators (MSCLE), using contiguous pairs to construct the CL as in Section 3.

The Appendix A.3 mirrors exactly the same setup based upon the scalar BEKK model: the results are very similar for that model.

A Monte Carlo study based on $2,500$ replications has been conducted across a variety of sample sizes and parameter configurations. As in Engle and Sheppard (2001), we assume away ARCH effects by setting $\sigma_{jt}^2 = 1$. Throughout we used $T = 2,000$, K is one of $\{3, 10, 50, 100\}$ and the returns were simulated according to a cDCC model given in Section 3.7. Three choices spanning the range of empirically relevant values of the temporal dependence in the Q process were used

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0.02 \\ 0.97 \end{pmatrix}, \quad \begin{pmatrix} 0.05 \\ 0.93 \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} 0.10 \\ 0.80 \end{pmatrix}.$$

	Bias						RMSE					
	MMLE		MCLE		MSCLE		MMLE		MCLE		MSCLE	
K	α	β	α	β	α	β	α	β	α	β	α	β
$\alpha = .02, \beta = .97$												
3	.001	-.011	.001	-.012	.001	-.017	.006	.033	.007	.038	.008	.059
10	-.001	-.004	-.000	-.005	-.000	-.006	.002	.005	.002	.006	.003	.009
50	-.003	-.003	-.000	-.005	-.000	-.005	.003	.003	.001	.005	.002	.006
100	-.005	-.004	-.000	-.005	-.000	-.005	.005	.004	.001	.005	.001	.005
$\alpha = .05, \beta = .93$												
3	-.000	-.005	-.000	-.006	-.000	-.007	.008	.015	.009	.016	.011	.022
10	-.002	-.001	-.000	-.003	-.000	-.004	.003	.004	.003	.006	.005	.009
50	-.009	.003	-.001	-.003	-.001	-.003	.009	.003	.002	.004	.003	.005
100	-.014	.002	-.001	-.003	-.001	-.003	.014	.002	.002	.004	.002	.004
$\alpha = .10, \beta = .80$												
3	-.001	-.007	-.001	-.008	-.001	-.010	.016	.037	.017	.040	.019	.051
10	-.003	-.003	-.001	-.005	-.001	-.006	.006	.011	.007	.016	.009	.022
50	-.014	.000	-.001	-.004	-.001	-.004	.014	.004	.004	.009	.005	.011
100	-.024	-.003	-.001	-.004	-.001	-.004	.024	.004	.004	.008	.005	.010

Table 3: *Properties of the estimators of α and β in the cDCC model using $T = 2,000$. The estimators are: subset CL (MSCLE), full CL (MCLE), and m-profile likelihood (MMLE) estimators. Based on $2,500$ replications.*

The parameters were estimated using a constraint that $0 \leq \alpha < 1$, $0 \leq \beta < 1$, $\alpha + \beta < 1$. None of the estimators were on the boundary of the parameter space.

The intercept Ψ was chosen to match the properties of the S&P 100 returns studied in the previous Section. The unconditional correlations were constructed from a single-factor model, the unconditional covariance from a strict factor model where

$$\epsilon_{i,t} = \pi_i f_t + \sqrt{1-\pi_i}\eta_{i,t} \tag{10}$$

where both f_t and $\eta_{i,t}$ have unit variance and are independent. Here π is distributed according to a truncated normal with mean 0.5, standard deviation 0.1 where the truncation occurs at ± 4 standard deviations. This means $\pi \in (0.1, 0.9)$. Obviously $\mathrm{E}(\epsilon_{i,t}|\pi_i) = 0$ and

$$\mathrm{Cov}\left\{ \begin{pmatrix} \epsilon_{i,t} \\ \epsilon_{j,t} \end{pmatrix} | \pi_i, \pi_j \right\} = \begin{pmatrix} 1 & \pi_i\pi_j \\ \pi_i\pi_j & 1 \end{pmatrix}. \tag{11}$$

so unconditionally, in the cross section, the $\epsilon_{i,t}$ and $\epsilon_{j,t}$ have a correlation of 0.25. This choice for Ψ produces assets which are all positively correlated and ensures that the intercept is positive definite for any cross-sectional dimension K.[10]

Tables 3 contains the bias and root mean square error of the estimates. The maximum m-profile likelihood (MMLE) method develops a significant bias in estimating α as K increases. This

[10]The effect of this choice of unconditional correlation was explored in other simulations. These results of these runs indicate that the findings presented are not sensitive to the choice of unconditional correlation.

	Bias						RMSE					
	MMLE		MCLE		MSCLE		MMLE		MCLE		MSCLE	
T	α	β	α	β	α	β	α	β	α	β	α	β
					$K = 10$							
100	-.021	-.161	-.011	-.141	-.009	-.218	.025	.237	.021	.221	.028	.347
250	-.006	-.018	-.002	-.021	-.002	-.026	.008	.021	.008	.026	.012	.042
500	-.003	-.005	-.001	-.008	-.001	-.009	.005	.008	.005	.011	.007	.016
1,000	-.002	-.001	-.001	-.003	-.001	-.003	.003	.004	.004	.006	.005	.009
2,000	-.001	-.000	-.000	-.002	-.000	-.002	.002	.003	.003	.004	.004	.006
					$K = 50$							
100	-.050	-.915	-.014	-.091	-.013	-.108	.050	.915	.016	.103	.018	.146
250	-.022	-.034	-.003	-.018	-.003	-.019	.022	.034	.005	.020	.006	.022
500	-.013	-.004	-.001	-.007	-.001	-.007	.013	.004	.003	.009	.004	.010
1,000	-.009	.003	-.001	-.003	-.001	-.003	.009	.003	.002	.004	.003	.005
2,000	-.006	.003	-.000	-.001	-.000	-.001	.006	.003	.001	.002	.002	.003
					$K = 100$							
100	–	–	-.014	-.090	-.014	-.098	–	–	.016	.103	.017	.121
250	-.037	-.108	-.003	-.019	-.003	-.019	.037	.109	.004	.020	.005	.021
500	-.021	-.013	-.001	-.007	-.001	-.007	.021	.013	.003	.008	.003	.009
1,000	-.014	.001	-.001	-.003	-.001	-.003	.014	.002	.002	.004	.002	.004
2,000	-.010	.004	-.000	-.001	-.000	-.001	.010	.004	.001	.002	.002	.003
					$K = 200$							
100	–	–	-.014	-.086	-.013	-.082	–	–	.016	.092	.016	.095
250	-.050	-.913	-.002	-.018	-.003	-.018	.050	.918	.004	.019	.005	.019
500	-.033	-.053	-.001	-.007	-.001	-.007	.033	.053	.002	.008	.003	.008
1,000	-.021	-.006	-.000	-.003	-.001	-.003	.022	.006	.002	.004	.002	.004
2,000	-.015	.003	-.000	-.002	-.000	-.001	.015	.003	.001	.002	.001	.002

Table 4: *Results from a simulation study for the cDCC model using the true values of* $\alpha = .05$, $\beta = .93$. *The estimators were: subset CL (MSCLE), CL (MCLE), and m-profile likelihood (MMLE) estimators. Based on* $2,500$ *replications.*

is consistent with the findings of Engle and Sheppard (2001) and our theoretical discussion given in Section 2.2.

To further examine the bias across T and K a second experiment was conducted for $K = \{10, 50, 100, 200\}$ and $T = \{100, 250, 500, 1000, 2000\}$. Only the results for the $\alpha = .05$, $\beta = .93$ parameterization are reported.

All of the estimators are substantially biased when T is very small. For any cross-section size K, the bias in the MMLE is monotonically decreasing in T. For large K, α is biased downward by 30% even when $T = 2,000$. The MCLE and MSCLE show small biases for any cross-section size as long as $T \geq 250$. Moreover, the bias does not depend on K. This experiment also highlights that the MCLE and MSCLE estimators are feasible when $T \leq K$. Results for the MMLE in the $T \leq K$ case are not reported because the estimator failed to converge in most replications.

Overall the Monte Carlo provides evidence of the MCLE has better RMSE for all cross-section sizes and parameter configurations. There seems little difference between the MCLE and MSCLE. In simulations not reported here, both estimators substantially outperform the Engle (2008b) estimator. The evidence presented here suggests MSCLE is attractive from statistical and computational viewpoints for large dimensional problems.

4.2 Efficiency gains with increasing cross-section length

Figure 1 contains a plot of the square root of the average variance against the cross-section size for the maximized MCLE and MSCLE. Both standard deviations rapidly decline as the cross-section dimension grows and the standard deviation of the MCLE is always slightly smaller than the MSCLE for a fixed cross-section size. Recall that the MCLE uses many more submodels than the MSCLE when the cross-section size is large, and so when $K = 50$ the MCLE is based on $1,225$ submodels while the MSCLE is using only 49.

This Figure shows there are very significant efficiency gains from using a CL compared to the simplest strategy for estimating θ — which is to fit a single bivariate model. The standard deviation goes down by a factor of 4 or so, which means the cross-sectional information is equivalent to increasing the time series dimension by a factor of around 16 when K is around 50.

Another interesting feature of the Figure is the expected result that as K increases the standard error of the MCLE and MSCLE estimators become very close. In the limit they will both asymptote to a value above zero — it looks like this asymptote is close to being realised by the time $K = 100$.

4.3 Performance of asymptotic standard errors

The Monte Carlo study was extended to assess the accuracy of the asymptotic based covariance estimator in Section 3.6. Data was simulated according to a cDCC model using the previously

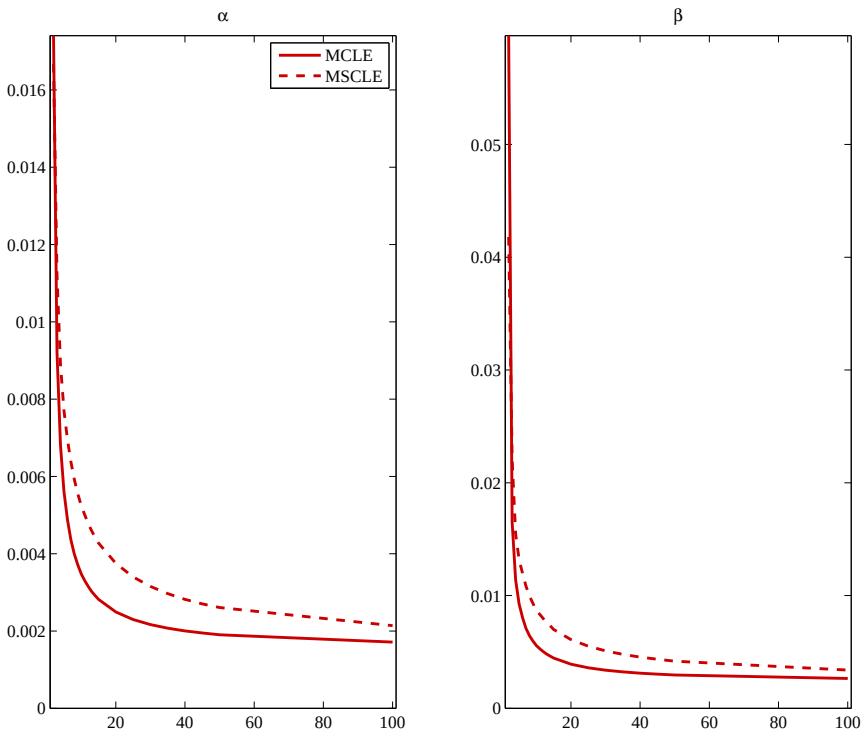

Figure 1: *Standard deviation of the CL estimators drawn against K calculated from a Monte Carlo based upon $\alpha = .05$, $\beta = .93$ using $T = 2,000$. K varies from 2 up to 100. Graphed are the results for the maximum CL estimator (MCLE) and the subset version (MSCLE) based on only contiguous submodels.*

described configuration for $\alpha = .05$, $\beta = .93$ with $T = 2,000$. The MCL estimator and the MSCL estimator, for both the maximized and m-profile strategies, were computed from the simulated data and the covariance of the parameters was estimated. This was repeated $1,000$ times and the results are presented in Table 5. The Table contains square root of the average asymptotic variance,

$$\bar{\sigma}_\alpha = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_{i,\alpha}^2} \tag{12}$$

and the standard deviation of the Monte Carlo's estimated parameters,

$$\hat{\sigma}_\alpha = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\tilde{\alpha}_i - \bar{\tilde{\alpha}}\right)^2}, \quad \bar{\tilde{\alpha}} = \frac{1}{1000}\sum_{i=1}^{1000}\tilde{\alpha}_i, \tag{13}$$

for both α and β.

The results are encouraging, except when K is tiny, the asymptotics performs quite accurately and yield a sensible basis for inference for this problem.

	MCLE								MSCLE							
	m-profile				maximized				m-profile				maximized			
K	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$
α=.02, β=.97																
3	.010	.008	.261	.152	.009	.008	.123	.147	.008	.007	.052	.028	.009	.008	.085	.028
10	.002	.002	.004	.004	.002	.002	.004	.004	.003	.003	.008	.007	.003	.003	.008	.007
50	.001	.001	.002	.002	.001	.001	.002	.002	.002	.002	.003	.003	.002	.002	.003	.003
100	.001	.001	.002	.001	.001	.001	.002	.001	.001	.001	.002	.002	.001	.001	.002	.002
α=.05, β=.93																
3	.009	.009	.016	.015	.009	.009	.016	.015	.011	.010	.021	.019	.011	.011	.021	.019
10	.003	.003	.006	.006	.003	.004	.006	.006	.005	.005	.009	.009	.005	.005	.009	.009
50	.002	.002	.003	.003	.002	.002	.003	.003	.003	.003	.004	.004	.003	.003	.004	.004
100	.002	.002	.003	.003	.002	.002	.003	.003	.002	.002	.003	.003	.002	.002	.003	.003
α=.10, β=.80																
3	.017	.016	.041	.040	.017	.017	.040	.040	.020	.019	.052	.049	.020	.019	.053	.049
10	.007	.006	.015	.014	.007	.006	.014	.014	.009	.010	.022	.022	.010	.010	.022	.022
50	.004	.004	.008	.008	.004	.004	.008	.008	.005	.005	.011	.011	.005	.005	.011	.011
100	.003	.003	.007	.007	.003	.003	.007	.007	.004	.004	.009	.009	.004	.004	.009	.009

Table 5: *Square root of average asymptotic variance, denoted $\bar{\sigma}_\alpha$ and $\bar{\sigma}_\beta$, and standard deviation of the Monte Carlo estimated parameters, denoted $\hat{\sigma}_\alpha$ and $\hat{\sigma}_\beta$.*

5 Empirical comparison

5.1 Database

The data used in this empirical illustration is the same as used in Section 2.3. Recall this database includes the superset of all companies listed on the S&P 100, plus the index itself, over the period January 1, 1997 until December 31, 2006 taken from the CRSP database. This set included 124 companies although 29, for example Google, have one or more periods of non-trading, for example prior to IPO or subsequent to an acquisition. Selecting only the companies that have returns throughout the sample reduced this set of 95 (+1 for the index).

We will use pairs of data and look at two MMCLE estimators for a variety of models. One is based on all distinct pairs, which has $N = K(K-1)/2$. The other just looks at contiguous pairs $Y_{jt} = (r_{jt}, r_{j+1t})'$ so $N = K - 1$. The results, given in Table 6, are directly comparable with Table 1. The figures in brackets are asymptotic standard errors.

The results for the m-profile CL are reasonably stable with respect to K and they do not vary much as we move from using all pairs to a subset of them. The corresponding results for the maximum CL estimator, optimising the CL over λ, are also reported in Table 6. Again the results are quite stable with respect with K.

Estimates from the MMLE are markedly different from those of any of the CL based estimators, which largely agree with each other. The parameter estimates of the MMLE and other estimators also produced meaningfully different fits.

	m-profile					maximised			
	Scalar BEKK		EWMA	DCC		Scalar BEKK		DCC	
K	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\widetilde{\alpha}$	$\widetilde{\beta}$	$\widetilde{\alpha}$	$\widetilde{\beta}$
All Pairs									
5	.0287	.9692	.0205	.0143	.9829	.0288	.9692	.0116	.9873
	(.0081)	(.0092)	(.0037)	(.0487)	(.0846)	(.0073)	(.0082)	(.0048)	(.0056)
10	.0281	.9699	.0211	.0107	.9881	.0276	.9705	.0107	.9875
	(.0055)	(.0063)	(.0027)	(.0012)	(.0016)	(.0050)	(.0057)	(.0013)	(.0021)
25	.0308	.9667	.0234	.0100	.9871	.0327	.9646	.0102	.9866
	(.0047)	(.0055)	(.0023)	(.0009)	(.0017)	(.0043)	(.0047)	(.0010)	(.0021)
50	.0319	.9645	.0225	.0101	.9856	.0345	.9615	.0104	.9848
	(.0046)	(.0056)	(.0026)	(.0008)	(.0018)	(.0037)	(.0042)	(.0009)	(.0017)
96	.0334	.9636	.0249	.0103	.9846	.0361	.9601	.0106	.9841
	(.0041)	(.0049)	(.0019)	(.0009)	(.0019)	(.0031)	(.0034)	(.0009)	(.0018)
Contiguous Pairs									
5	.0284	.9696	.0189	.0099	.9885	.0251	.9733	.0078	.9917
	(.0083)	(.0094)	(.0037)	(.0033)	(.0045)	(.0070)	(.0079)	(.0055)	(.0059)
10	.0272	.9709	.0201	.0093	.9886	.0266	.9717	.0088	.9900
	(.0054)	(.0062)	(.0027)	(.0016)	(.0018)	(.0049)	(.0055)	(.0018)	(.0020)
25	.0307	.9668	.0227	.0089	.9889	.0315	.9660	.0088	.9894
	(.0049)	(.0056)	(.0024)	(.0011)	(.0012)	(.0044)	(.0050)	(.0012)	(.0013)
50	.0316	.9647	.0220	.0092	.9869	.0347	.9612	.0095	.9864
	(.0047)	(.0057)	(.0029)	(.0010)	(.0019)	(.0038)	(.0043)	(.0011)	(.0019)
96	.0335	.9634	.0247	.0094	.9860	.0364	.9598	.0095	.9863
	(.0043)	(.0051)	(.0020)	(.0009)	(.0014)	(.0032)	(.0035)	(.0009)	(.0012)

Table 6: *Based on the maximum m-profile and maximum CL estimator (MMCLE) using real and simulated data. Top part uses $K(K-1)/2$ pairs based subsets, the bottom part uses K-1 contiguous pairs. Parameter estimates from a covariance targeting scalar BEKK, EWMA (estimating H_0) and DCC. The real database is built from daily returns from 95 companies plus the index from the S&P100, from* 1997 *until* 2006*. Numbers in brackets are asymptotic standard errors.*

It is interesting to see how sensitive the contiguous pairs estimator is to the selection of the subset of pairs. The bottom row of Figure 2 shows the density of the estimator as we select randomly 1,000 sets of different subsets of $K-1$ pairs. We see the estimate is hardly effected at all.

To examine the fit of the models, the conditional correlations of the 95 individual stocks with the S&P 500 from the MCLE and MMLE are presented in Figure 3. Rather than present all of the series simultaneously, the figure contains the median, inter-quartile range, and the maximum and minimum. The parameter estimates from the MCLE produce large, persistent shifts in conditional correlations with the market, including a marked decrease in the conditional correlations near the peak of the technology boom in 2001. The small estimated α for MMLE produces conditional correlations which are nearly constant and exhibiting little variation even at the height of the technology bubble in 2001.

5.2 Out of sample comparison of hedging performance

To determine whether the fit from the estimators was statistically different, a simple hedging problem is considered in an out-of-sample period. The out-of-sample comparison was conducted using the first 75% of the sample: January 2, 1997 until July 1, 2002 as the "in-sample" period for parameter estimation, and July 2, 2002 until December 31, 2006 as the evaluation period. All of

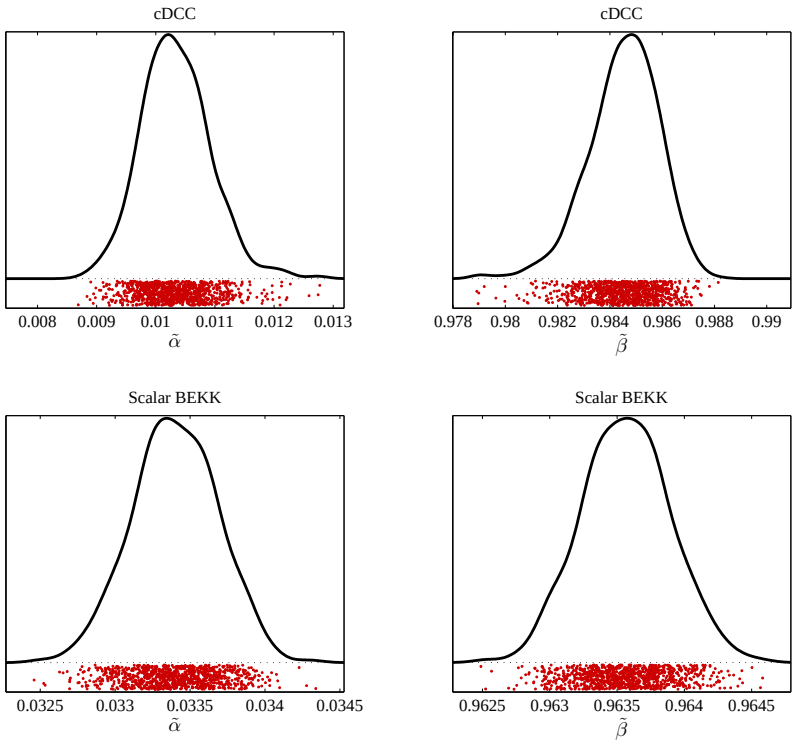

Figure 2: *Sensitivity to random selection of pairs. Density of the maximum m-profile CL estimator based on $K-1$ distinct but randomly choosen pairs. Top row are the estimators of the cDCC model and the bottom row are the corresponding estimators for the scalar BEKK.*

the parameters were estimated once and used throughout the tests.

We examined the hedging errors of a conditional CAPM where the S&P 100 index proxied for the market. Using one-step ahead forecasts, the conditional time-varying market betas were computed as

$$\widehat{\beta}_{j,t} = \frac{\widehat{h}_{j,t}^{1/2}\widehat{\rho}_{jm,t}}{\widehat{h}_{m,t}^{1/2}}, \quad j=1,2,...,K, \quad h_{j,t}=\mathrm{Var}(r_{j,t}|\mathcal{F}_{t-1}), \tag{14}$$

$$h_{j,t} = \mathrm{Var}(r_{j,t}|\mathcal{F}_{t-1}), \quad \rho_{jm,t}=\mathrm{Cor}(r_{j,t},r_{m,t}|\mathcal{F}_{t-1}) \tag{15}$$

and the corresponding hedging errors were computed as $\widehat{\nu}_{j,t} = r_{j,t} - \widehat{\beta}_{j,t} r_{m,t}$. Here $r_{j,t}$ is the return on the j-th asset and $r_{m,t}$ is the return on the market. Since all of the volatility models are identical in the DCC models in this comparison and use the same parameter estimates, all differences in the hedging errors are directly attributable to differences in the correlation forecast.

We use the Giacomini and White (2006) (GW) test to examine the relative performance of the MCLE to the MMLE. The GW test is designed to compare forecasting methods, which incorporate such things as the forecasting model, sample period and, importantly from our purposes, the estimation method employed[11].

[11]The related tests of Diebold and Mariano (1995) and West (1996) focus solely on comparing forecast models and are thus well-suited for our problem.

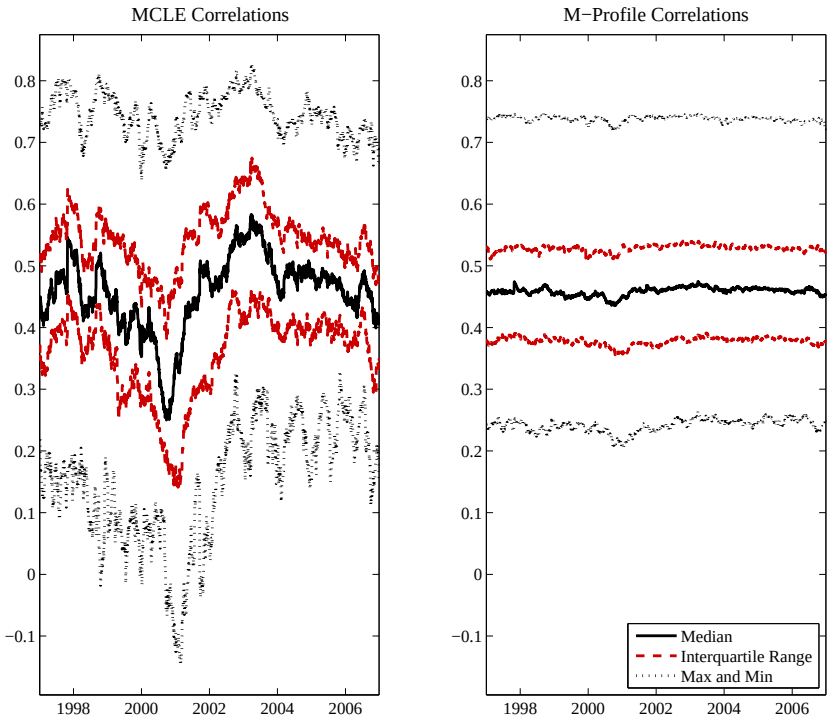

Figure 3: *How do the correlations with the market change through time? Plot of the median, interquartile range and minimum and maximum of the correlations of the 95 included S&P 100 components with the index return using the estimates produced by the maximum CL estimator (MCLE) and maximum m-profile likelihood estimator. Each day the 95 correlations were sorted to produce the necessary quantiles.*

Defining the difference in the squared hedging error

$$\widehat{\delta}_{j,t} = \left\{\widehat{\nu}_{j,t}\left(\widehat{\rho}_{j,t}^{MCLE}\right)\right\}^2 - \left\{\widehat{\nu}_{j,t}\left(\widehat{\rho}_{j,t}^{MMLE}\right)\right\}^2$$

where explicit dependence on the forecast correlation is used. If neither estimator is superior in forecasting correlations, this difference should have 0 expectation. If the difference is significantly different from zero and negative, the MCLE would be the preferred model while significant positive results would indicate favor for the MMLE. The null of

$$H_0 : \mathrm{E}\left(\widehat{\delta}_{j,t}\right) = 0$$

was tested using a t-test,

$$GW = \frac{\bar{\delta}_j}{avar\left(\sqrt{T}\bar{\delta}_j\right)} \tag{16}$$

where

$$\bar{\delta}_j = P^{-1}\sum_{t=R}^{P}\delta_{j,t}$$

Maximum

Model A	Favours A	No Decision	Favours B	Model B
DCC MCLE	37	56	2	DCC MMLE
DCC MCLE	92	3	0	DECO
DCC MCLE	28	60	7	Bivariate DCC
DCC MCLE	12	83	0	EWMA
BEKK MCLE	28	64	3	BEKK MMLE
BEKK MCLE	16	79	0	Bivariate BEKK

M-profile

Model A	Favours A	No Decision	Favours B	Model B
DCC MCLE	24	63	8	DCC MMLE
DCC MCLE	92	3	0	DECO
DCC MCLE	18	68	9	Bivariate DCC
DCC MCLE	9	82	4	EWMA
BEKK MCLE	29	65	1	BEKK MMLE
BEKK MCLE	50	44	1	Bivariate BEKK

Table 7: *Which model and estimation strategy leads to smallest hedging errors? GW t-statistics for testing the null of equal out of sample hedging performance using Giacomini-White tests with 95% critical values. 3 decisions can be made for each of the 95 single assets. Two CL methods used: maximum CL estimation and maximum m-profile CL estimation. The test can favour model A, model B or be indecisive. Table records the number of assets which fall in each of these three buckets.*

is the average loss differential. Under mild regularity conditions GW is asymptotically normal. See Giacomini and White (2006) for further details[12].

The test statistic was computed for each asset excluding the market, resulting in 95 test statistics. The results are in Table 7, which 37 series which favour the MCLE estimator compared to 2 which prefer the MMLE based estimated model. 56 are inconclusive. The corresponding results for the maximum m-profile CL estimator are 24 in favour of that estimator, 8 preferring MMLE and 63 inconclusive. We will see later that it is not a consistent pattern that maximum CL estimator performs better than the m-profiled version, although for DCC models this is the general trend.

5.3 Out of sample comparison with other models

5.3.1 Scalar BEKK

We can use the CL methods to estimate the scalar BEKK model using this database. The results are given in Table 1 and 6 — here we focus on the m-profile based estimators. The results have the same theme as before, with the estimates from the quasi-likelihood parameters yielding extreme values — in this case close to being non-responsive to the data.

The usual out of sample GW hedging error comparison is given in Table 7, which compares MMLE and MCLE. They show the CL method delivering estimators which produce smaller hedging errors than the conventional maximum m-profile likelihood technique.

[12]We also tried a heteroskedastically adjusted version of the GW test, in order to increase its power, but this had no impact.

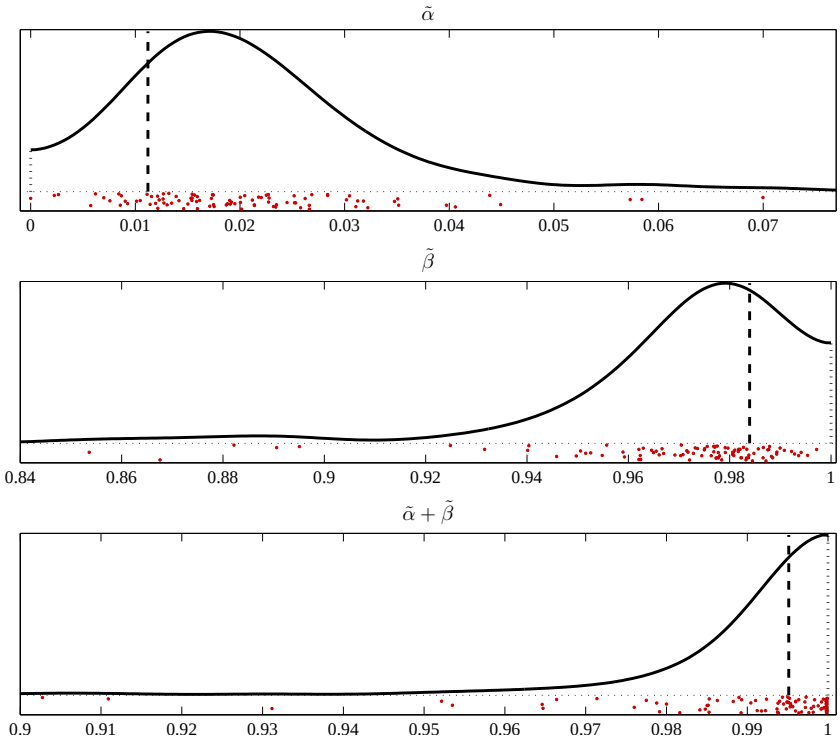

Figure 4: *Should the data be pooled across pairs? Seperately estimated α_j and β_j for each bivariate submodel for the beta-pair of the market and an individual asset. Dotted line is the CL estimator — which acts as a pooling device.*

5.3.2 Many bivariate models

An interesting way of assessing the effectiveness of the DCC model fitted by the CL method is to compare the fit to fitting a separate DCC model to each pair — that is permit θ to be different for each j. The Table 7 shows the multivariate DCC model, estimated using CL methods, performs better than fitting a different model for each pair. This is a striking result — suggesting the pooling of information is helpful in improving hedging performance.

Figure 4 shows us why the large dimensional multivariate model is so effective. This shows the estimated value of α_j and β_j for each of the j-th submodels — it demonstrates a very significant scatter. It has 22 of the estimated $\alpha_j + \beta_j$ on their unit boundary. We will see in a moment such unit root models, which are often called EWMA models, perform very poorly indeed in terms of hedging. Once in a while the estimates of $\alpha_j + \beta_j$ are pretty small.

Figure 5 shows four examples of estimated time varying correlations between a specific asset and the market, drawn for 4 specific pairs of returns we have chosen to reflect the variety we have seen in practice. The vertical dotted line indicates where we move from in sample to out of sample data. Top right shows a case where the estimated bivariate model and the fit from the highly multivariate model are very similar, both in and out of sample. The top left shows a case where

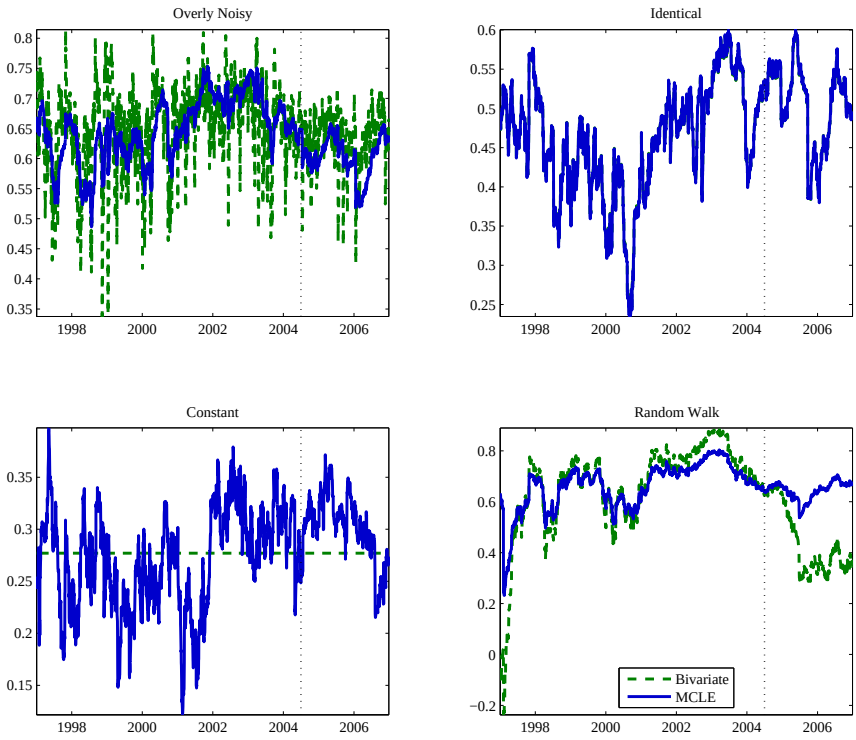

Figure 5: *Comparison of estimated conditional correlations for j-th model, including out of sampling projections, using the high dimensional model and the bivariate model. Top left looks like the bivariate model is overly noisy. Top right give results which are basically the same. Bottom left gives a constant correlation for the bivariate model, while the multivariate model is more responsive. Bottom right is a key example as we see it quite often. Here the bivariate model is basically estimated to be an EWMA, which fits poorly out of sample.*

the fitted bivariate model has too little dependence and so seems to give a fitted correlation which is too noisy. The bottom left is the flip side of this, the bivariate model delivers a constant correlation which seems very extreme. The bottom right is an example where the EWMA model is in effect imposed in the bivariate case and this EWMA model fits poorly out of sample.

5.3.3 Equicorrelation model

The Engle and Kelly (2007) linear equicorrelation (DECO) model has a similar structure to the DCC type models, with each asset price process having its own ARCH model, but assumes asset returns have at each time point equicorrelation $R_t = \rho_t \iota\iota' + (1 - \rho_t) I$, while $\rho_t = \omega + \gamma u_{t-1} + \beta\rho_{t-1}$, where u_{t-1} is new information about the correlation in the devolatilised r_{t-1}. A simple approach would be to take u_{t-1} as the cross-sectional MLE of the correlation based on this simple equicorrelation model.

Table 7 compares the out of sample hedging performance of this method with the cDCC fit. We can see that cDCC is uniformly statistically preferable for this dataset.

5.3.4 RiskMetrics

The MCLE fit of the cDCC model can be compared to the RiskMetrics method given in Example 2 using the Giacomini and White (2006) t-test. The results are reported in the bottom right of Table 7, which shows that the cDCC outperforms RiskMetrics in terms of out of sample hedging errors.

5.4 Extending the empirical analysis

In this subsection we will push the previous analysis to a higher dimensions. Our database consists of the returns of all equities that appeared in the S&P 500 between January 1, 1997 and December 31, 2006 and were continuously available. This resulted in 480 unique assets, including the S&P 500 index, with $2,516$ observations of each. The data were extracted from CRSP and series were ordered alphabetically according to their ticker on the first day of the sample. Obviously around 25% of the data used in this analysis has previously appeared in the S&P 100 comparison.

As before the scalar BEKK was fitted using maximum m-profile likelihood (MMLE), maximum composite likelihood, maximum m-profiled composite likelihood, and the subset version of the two composite likelihood estimators that uses contiguous pairs. The model was estimated across $K = \{5, 25, 50, 100, 250, 480\}$. Results are presented in the top panel of Table 8.

The MMLE shows clear signs of bias as the cross-sectional dimension is increased, and for the two largest panel sizes produces volatilities that are virtually constant. When the full cross-section sample is used the smoothing coefficient β also shows a large downward bias. The composite likelihood estimates are very similar, all with $\alpha \approx .03$, $\beta \approx .96$, and the standard errors decline quickly and then modestly as K increases. For large K the difference between the contiguous and all pairs estimators is very small indeed. Further, the standard errors for the maximized composite likelihood are approximately 30% smaller in the larger K case than for the m-profile CL estimator.

In the analysis of the cDCC model, for this wider set of data the best performing univariate volatility model was the GJR-GARCH(1,1) for each margin[13]. The results for the cDCC model are presented in Table 8. The MMLE of α for the cDCC model exhibits a strong bias as the sample size increases and for $K > 250$ the β estimate is also badly affected. These estimates contrast sharply with the estimates from the maximum composite and maximum m-profile composite likelihood where $\alpha \approx .008$ and $\alpha + \beta \approx .995$.[14]

[13] In particular we fitted

$$h_{j,t} = \omega_j + \delta_j r_{j,t-1}^2 + \gamma_j r_{j,t-1}^2 I_{[r_{j,t-1}<0]} + \kappa_j h_{j,t-1}. \tag{17}$$

[14] The maximized composite likelihood was computed by jointly maximizing the correlation intercept with the dynamics parameters. The estimates from the volatility models were held at their initial estimated values.

Scalar BEKK

	MMLE		All Pairs				Contiguous Pairs			
	m-profile		maximum		m-profile		maximum		m-profile	
K	α	β	α	β	α	β	α	β	α	β
5	.0261	.9715	.0358 (.0065)	.9608 (.0074)	.0369 (.0057)	.9603 (.0065)	.0307 (.0055)	.9666 (.0063)	.0312 (.0053)	.9664 (.0061)
25	.0080	.9909	.0362 (.0066)	.9589 (.0075)	.0300 (.0062)	.9670 (.0075)	.0344 (.0058)	.9611 (.0066)	.0289 (.0055)	.9682 (.0067)
50	.0055	.9932	.0346 (.0049)	.9609 (.0056)	.0282 (.0051)	.9692 (.0062)	.0341 (.0048)	.9615 (.0055)	.0277 (.0049)	.9698 (.0059)
100	.0034	.9934	.0343 (.0038)	.9602 (.0044)	.0296 (.0046)	.9670 (.0057)	.0341 (.0038)	.9605 (.0044)	.0292 (.0045)	.9674 (.0056)
250	.0015	.9842	.0364 (.0036)	.9574 (.0042)	.0322 (.0049)	.9633 (.0064)	.0365 (.0035)	.9573 (.0041)	.0322 (.0048)	.9633 (.0063)
480	.0032	.5630	.0327 (.0030)	.9619 (.0035)	.0290 (.0041)	.9672 (.0054)	.0327 (.0029)	.9619 (.0034)	.0290 (.0040)	.9672 (.0053)

DCC

	MMLE		All Pairs				Contiguous Pairs			
	m-profile		maximum		m-profile		maximum		m-profile	
K	α	β	α	β	α	β	α	β	α	β
5	.0101	.9823	.0117 (.0090)	.9843 (.0193)	.0133 (.0041)	.9794 (.0081)	.0072 (.0038)	.9917 (.0043)	.0070 (.0033)	.9912 (.0038)
25	.0030	.9908	.0083 (.0024)	.9890 (.0055)	.0083 (.0015)	.9885 (.0031)	.0071 (.0036)	.9917 (.0048)	.0071 (.0011)	.9911 (.0016)
50	.0018	.9882	.0080 (.0014)	.9886 (.0031)	.0078 (.0010)	.9887 (.0021)	.0073 (.0015)	.9910 (.0021)	.0073 (.0010)	.9901 (.0019)
100	.0015	.9524	.0075 (.0007)	.9879 (.0016)	.0073 (.0007)	.9881 (.0015)	.0076 (.0027)	.9874 (.0061)	.0076 (.0010)	.9866 (.0028)
250	.0020	.5561	.0078 (.0007)	.9870 (.0015)	.0076 (.0007)	.9872 (.0015)	.0080 (.0010)	.9866 (.0023)	.0080 (.0016)	.9858 (.0039)
480	.0013	.2556	.0075 (.0007)	.9872 (.0015)	.0073 (.0007)	.9874 (.0016)	.0079 (.0010)	.9869 (.0021)	.0079 (.0008)	.9863 (.0020)

Table 8: *Results for fitting the Scalar BEKK model using a variety of estimators. The database is made up of the 480 components of the S&P 500, ordered alphabetically by ticker. K is the dimension of problem fitted.*

Table 2 contains the run times for each of the methods for estimating the scalar BEKK model — the simpler of the two models. The MMLE method takes around 3.5 days on the $K = 480$ problem, while for $K = 25$ the time is quite modest being under a minute. This shows the impact of the $O(K^3)$ computational load.

The composite methods can be carried out using full maximisation or m-profiling, the cost of full maximisation is non-trivial, typically leading to an increase in time by a factor of 10 compared to m-profiling, which makes the method based on all pairs slow by the time K goes much above 50. The contiguous pairs method is still reasonably fast even when $K = 480$.

When we use m-profiling the composite methods become much more rapid, with the all pairs method still being quite fast for $K = 100$ and being around 200 times faster than MMLE in that case. The contiguous pair method based on m-profiling is fast even when $K = 480$, just taking a small handful of seconds. This means it is around $68,000$ times faster than MMLE in this vast dimensional case.

6 Additional remarks

6.1 Parametric bootstrap

Having fitted the model one could compute

$$\varepsilon_t = H_t^{-1/2} r_t, \quad t = 1, 2, ..., T,$$

which, ignoring the effect of model estimation, is a $\mathcal{F}$-martingale difference sequence with $\text{Cov}(r_t|\mathcal{F}_{t-1}) = I$. Consequently we could do a parametric bootstrap off the "population" of innovations

$$\varepsilon_1, \varepsilon_2 ..., \varepsilon_T,$$

sampling from these K-dimensional random variables with replacement to produce

$$\varepsilon_1^*, \varepsilon_2^* ..., \varepsilon_T^*. \tag{18}$$

In turn then we compute

$$y_t^* = H_t^{*1/2} \varepsilon_t^*, \tag{19}$$

which in the scalar BEKK case, for example, is driven by the dynamic

$$H_t^* = (1 - \alpha - \beta)\,\Sigma + \alpha y_{t-1}^* y_{t-1}^* + \beta H_t^*, \quad H_1^* = \Sigma. \tag{20}$$

The sampling from (18)-(20) can be carried out many times, allowing us to simulate interesting quantities of interest such as a bootstrap distribution of the MMLE, MCLE and MMCLE or non-parametric forecast distributions. A disadvantage of this procedure is that step (19) costs $O(K^3)$, which will become expensive by the time K becomes 50 or more. This is not really a problem for the MMLE strategy as this already has this cost, but is disappointing for the computational fast composite strategies.

6.2 Composite bootstrap

An alternative, when we are bootstrapping the distribution of the estimator, is to bootstrap off the objective function. This is inspired by the paper by Goncalves and White (2004). To be concrete consider solely the CL estimator

$$c_t(\theta, \lambda) = \frac{1}{N} \sum_{j=1}^{N} \log L_{jt}(\theta, \lambda_j).$$

Then we construct the "population" of CL functions

$$c_1(\theta, \lambda), c_2(\theta, \lambda), \ldots, c_T(\theta, \lambda),$$

which we sample with replacement to produce

$$c_1^*(\theta,\lambda), c_2^*(\theta,\lambda), ..., c_T^*(\theta,\lambda).$$

We then samples these functions with replacement to produce a bootstrap sample of the CL function

$$c^*(\theta,\lambda) = \sum_{t=1}^{T} c_t^*(\theta,\lambda).$$

We then maximise this with respect to θ and λ in the usual way.

This strategy has the advantage that its computational cost is $O(N)$, the same as a single composite estimation, which is pretty fast, at least in the contiguous estimator case. The asymptotic justification of this strategy is implicitly in Goncalves and White (2004) when N is fixed and we do not impose variation freeness. Dropping these two conditions and proving it for our case would be interesting further work.

6.3 Beta CL

All statistical models are misspecified. If the goal is to estimate market betas, that is the dependence between the market and individual assets, it may make sense to define the "beta CL" based on the pairs

$$\begin{aligned} Y_{1t} &= (r_{1t}, r_{2t})', \\ Y_{2t} &= (r_{1t}, r_{3t})', \\ &\vdots \\ Y_{(K-1)t} &= (r_{1t}, r_{Kt})', \end{aligned}$$

where $N = K - 1$ and $\{r_{1t}\}$ is the return on the market. Statistically, if the model was correctly specified, this is likely to be less efficient than using K randomly chosen pairs, as the corresponding submodel quasi-likelihoods $\log L_{jt}(\theta, \lambda_j)$ will be tightly dependent across j. However, as the models will be incorrect then having this highly tuned to estimating betas may be beneficial — in effect allowing one to pool information on the estimation of betas across assets.

6.4 CL and λ

CL estimation of θ does not necessarily deliver estimates of all λ_j, for some CL estimators do not use all available pairs. Of course once θ is estimated all the missing elements in λ can be filled in rapidly. In the scalar BEKK and DCC cases this will costs $O(K^2)$.

6.5 Engle's method

Before we wrote our paper, Engle (2008b) proposed a method for estimating large dimensional models. He called it the MacGyver strategy, basing it on pairs of returns. Instead of averaging the log-likelihoods of pairs of observations, the log-likelihoods were separately maximised and then the resulting estimators were robustly averaged using medians. This overcomes the difficulty of inverting H, but has the difficulty that (i) it is not clear that the pooled estimators should have equal weight, (ii) it involves $K(K-1)/2$ maximisations, (iii) no properties of this estimator were derived, (iv) the resulting estimator may not be in the permissible parameter space[15]. Engle's MacGyver method has some similarities, but is distinct, with the Ledoit, Santa-Clara, and Wolf (2003) flexible multivariate GARCH estimation procedure which also fits models to many pairs of observations. It is distinctive as it is focused entirely on estimating a small number of common parameters.

It is not difficult to study the asymptotic properties of this estimator in the case where we replace the median by an average. This linear version of the MacGyver estimation method of Engle (2008b) would average the submodels maximum quasi-likelihood estimators, which asymptotically behave like

$$\frac{1}{N_T}\sum_{j=1}^{N_T}\widehat{\theta}_j = \frac{1}{N_T}\sum_{j=1}^{N_T}\theta_j - \frac{1}{TN_T}\sum_{j=1}^{N_T} D_{j,\theta\theta,T}^{-1}\sum_{t=1}^{T}\left(\frac{\partial l_{jt}}{\partial\theta'} - F_{j,T}g_{jt}\right),$$

using the notation defined in Section 3.6. Hence its asymptotic variance can be estimated by applying a HAC estimator to

$$Z_{t,T}^{M} = \sum_{j=1}^{N_T} D_{j,\theta\theta,T}^{-1}\left(\frac{\partial l_{jt}}{\partial\theta'} - F_{j,T}g_{jt}\right).$$

In the linear MacGyver case the estimator is dominated by the submodel estimators with largest variances — i.e. components which are least informative. We do not know how to extend this analysis to when we replace the mean by the median.

6.6 Imposing factor structure on Σ

In some stationary multivariate models it might make sense to impose structure on Σ, particularly when K is very large. There is a long history of using factor models in financial economics, see for example, Chamberlain and Rothschild (1983), King, Sentana, and Wadhwani (1994) and Diebold and Nerlove (1989). A leading candidate would be that Σ obeys a strict factor structure

$$\Sigma = ff' + \Omega,$$

[15] An example is the scalar BEKK model where $\alpha, \beta \in [0,1)$ and $\alpha+\beta<1$. The median of pairs based estimators of α and β, each constrained to satisfy the above conditions, will be in $[0,1)$ but there is no reason why the resulting estimated $\alpha+\beta<1$.

where f is a $K \times M$ matrix of factor loadings and Ω is a K by K diagonal matrix containing the residual variances. This means that in the long run the covariances in the model obey a factor structure but in the short run there can be departures from it. This is simple to carry out in the m-profile case, using a two step procedure of estimating the constrained Σ and then plugging this into a composite likelihood to estimate α and β.

We will take this model to the data. We estimate the factor model using the method of Jöreskog (1967) which assumes the returns, factors and innovations are i.i.d. Gaussian. This method means that the estimate Σ has the same diagonal elements of $T^{-1}\sum_{t=1}^{T} r_t r_t'$ and so only the correlations estimates differ.

The parameters controlling the dynamics were estimated for $M = 1, 2, 3$ using a composite likelihood. The estimates are presented in Table 9. The estimated parameters vary substantially as the cross-sectional dimension increases. The m-profile estimates that use a factor intercept are very close to $\alpha+\beta = 1$, although the sum moves marginally away from this boundary as the cross section increases. This is the classic sign of misspecification (Monte Carlo experiments, not reported here, indicate the above estimation method does not yield biased estimators when the factor structure is used as the data generator process), where the data wants to ignore the log-run Σ matrix and it does this by imposing a near unit root on the parameters.

	$M=1$		$M=2$		$M=3$	
K	α	β	α	β	α	β
5	0.0261	0.9715	0.0261	0.9715	0.0261	0.9715
25	0.0082	0.9909	0.0081	0.9909	0.0080	0.9908
50	0.0057	0.9935	0.0057	0.9934	0.0057	0.9933
100	0.0041	0.9949	0.0040	0.9947	0.0039	0.9946
250	0.0026	0.9955	0.0025	0.9953	0.0024	0.9950
480	0.0017	0.9964	0.0016	0.9963	0.0016	0.9961

Table 9: Parameter estimates from fitting a scalar BEKK to the S&P 500 components continuously available between 1998 and 2007 using a factor-model-based estimate of the intercept and a composite likelihood function for α and β. The dimension of the factor model is M. K denotes the number of assets analysed.

6.7 Insights from panel data literature

Consider the diagonal BEKK model

$$H_t = (1-\alpha)\Sigma + \alpha r_{t-1} r_{t-1}' + \beta H_{t-1}$$

then

$$\gamma_t = H_t - H_{t-1} = \alpha\left(r_{t-1}r_{t-1}' - r_{t-2}r_{t-2}'\right) + \beta\left(H_{t-1} - H_{t-2}\right),$$

so

$$\gamma_t - \beta\gamma_{t-1} = \alpha\left(r_{t-1}r'_{t-1} - r_{t-2}r'_{t-2}\right),$$

which is free of the incidental parameter Σ. This is similar in spirit, but somewhat more sophisticated due to the lagged H_t, to the influential approach to autoregressive panel data model of Arellano and Bond (1991) who estimate the parameters of interest based upon differences of data, differencing out their incidental individual effects.

7 Conclusions

This paper has introduced a new way of estimating large dimensional time-varying covariance models, based upon the sum of quasi-likelihoods generated by time series of pairs of asset returns. This CL procedure leads to a loss in efficiency compared to a full quasi-likelihood approach, but it is easy to implement, is not effected by the incidental parameter problem and scales well with the dimension of the problem. These new methods can be used to estimate models in many hundreds of dimensions, indeed the dimension could be larger than the time series dimension.

References

Aielli, G. P. (2006). Consistent estimation of large scale dynamic conditional correlations. Unpublished paper: Department of Statistics, University of Florence.

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

Arellano, M. and S. R. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies 58*, 277–297.

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.

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH approach. *Review of Economics and Statistics 72*, 498–505.

Bourgoin, F. (2002). Fast calculation of GARCH correlation. Presented at the 2002 Forecasting Financial Markets Conference.

Campbell, J. Y., M. Lettau, B. G. Malkeil, and Y. Xu (2001). Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *Journal of Finance 56*, 1–43.

Chamberlain, G. and M. Rothschild (1983). Arbitrage and mean-variance analysis of large asset markets. *Econometrica 51*, 1281–1301.

Chen, X., D. T. Jacho-Chavez, and O. Linton (2007). An alternative way of computing efficient instrumental variable estimators. Unpublished paper, Department of Economics, London School of Economics.

Chib, S., F. Nardari, and N. Shephard (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics 134*, 341–371.

Cox, D. R. (1961). Tests of seperate families of hypotheses. *Proceedings of the Berkeley Symposium 4*, 105–123.

Cox, D. R. and N. Reid (2003). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*, 729–737.

deLeon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics and Probability Letters 75*, 49–57.

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

Diebold, F. X. and M. Nerlove (1989). The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics 4*, 1–21.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 59–82. Berkeley: University of California.

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. (2008a). *Anticipating Correlations*. Princeton University Press. Forthcoming.

Engle, R. F. (2008b). 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*. Oxford University Press. Forthcoming.

Engle, R. F., D. F. Hendry, and J. F. Richard (1983). Exogeneity. *Econometrica 51*, 277–304.

Engle, R. F. and B. Kelly (2007). Dynamic equicorrelation. Unpublished paper, Stern Business School, NYU.

Engle, R. F. and K. F. Kroner (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory 11*, 122–150.

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

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

Fan, J., J. Zhang, and K. Yu (2008). Asset allocation with gross exposure constraints for vast portfolios. Unpublished paper: Bendheim Center for Finance, Princeton University.

Fearnhead, P. (2003). Consistency of estimators of the population-scaled recombination rate. *Theoretical Population Biology 64*, 67–79.

Fiorentini, G., E. Sentana, and N. Shephard (2004). Likelihood-based estimation of latent generalised ARCH structures. *Econometrica 72*, 1481–1517.

Gallant, A. R. and H. White (1988). A unified theory of estimation and inference for nonlinear dynamic models.

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

Goncalves, S. and H. White (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics 119*, 199–219.

Harvey, A. C., E. Ruiz, and E. Sentana (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics 52*, 129–158.

Jöreskog, K. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika 32*, 443–482.

King, M., E. Sentana, and S. Wadhwani (1994). Volatility and links between national stock markets. *Econometrica 62*, 901–933.

Kuk, A. Y. C. and D. J. Nott (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistical and Probability Letters 47*, 329–335.

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

LeCessie, S. and J. C. van Houwelingen (1994). Logistic regression for correlated binary data. *Applied Statistics 43*, 95–108.

Ledoit, O., P. Santa-Clara, and M. Wolf (2003). Flexible multivariate GARCH modeling with an application to international stock markets. *The Review of Economics and Statistics 85*, 735–747.

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

Nelson, D. B. (1991). Conditional heteroskedasticity in asset pricing: a new approach. *Econometrica 59*, 347–370.

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.

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

Nickell, S. J. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*, 1417–1426.

Pesaran, M. H. and B. Pesaran (2007). Modelling volatilities and conditional correlations in futures markets with a multivariate t distribution. Unpublished paper: Department of Economics, University of Cambridge.

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

Silvennoinen, A. and T. Terasvirta (2008). Multivariate GARCH models. In T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Springer-Verlag. Forthcoming.

Strassen, V. (1969). Gaussian elimination is not optimal. *Numerische Mathematik 13*, 354–356.

Tse, Y. (2000). A test for constant correlations in a multivariate GARCH model. *Journal of Econometrics 98*, 107–127.

Tsui, A. K. and Q. Yu (1999). Constant conditional correlation in a bivariate GARCH model: evidence from the stock market in China. *Mathematics and Computers in Simulation 48*, 503–509.

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

Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika 92*, 519–528.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

A Appendix

A.1 Proof of Theorem 1

Due to assumption 2, we solely have to show that the following goes to zero:

$$\begin{aligned} V &= \sup_{\theta\in\Theta}\left|\frac{1}{TN_T}\sum_{t=1}^{T}\sum_{j=1}^{N_T} l_{jt}(\theta,\lambda_{j\theta}^*) - \frac{1}{TN_T}\sum_{t=1}^{T}\sum_{j=1}^{N_T} l_{jt}(\theta,\widehat{\lambda}_j)\right| \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_T}\sum_{j=1}^{N_T}\sup_{\theta\in\Theta}\left|l_{jt}(\theta,\lambda_{j\theta}^*) - l_{jt}(\theta,\widehat{\lambda}_j)\right| \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_T}\sum_{j=1}^{N_T}\sup_{\theta\in\Theta}\left|\frac{\partial l_{jt}(\theta,\overline{\lambda}_j)}{\partial\lambda_j}\left(\widehat{\lambda}_j - \lambda_{j\theta}^*\right)\right|, \end{aligned}$$

where $\overline{\lambda}_j \in \left[\min(\widehat{\lambda}_j,\lambda_j), \max(\widehat{\lambda}_j,\lambda_j)\right]$ by the mean value theorem using Assumption 3. Then

$$V \leq \max_{j\in\{1,2,\ldots,N_T\}}\left|\widehat{\lambda}_j - \lambda_j\right| \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_T}\sum_{j=1}^{N_T}\sup_{\theta\in\Theta,\lambda_j\in\Lambda_j}\left|\frac{\partial l_{jt}(\theta,\lambda_j)}{\partial\lambda_j}\right|$$

$$= \max_{j\in\{1,2,...,N_T\}} \left|\widehat{\lambda}_j - \lambda_j\right| \frac{1}{T}\sum_{t=1}^{T} c_{t,T},$$

where

$$c_{t,T}(r_t) = \frac{1}{N}\sum_{j=1}^{N_T} \sup_{\theta\in\Theta,\lambda_j\in\Lambda_j} \left|\frac{\partial l_{jt}(\theta,\lambda_j)}{\partial\lambda_j}\right|.$$

We assume that the fixed dimensional

$$\frac{1}{T}\sum_{t=1}^{T} c_{t,T}(r_t)$$

exhibits a weak law of large numbers. Then we have the result from Assumption 5.

A.2 Proof of Theorem 2

The asymptotic properties of these types of CL estimators were derived in Cox and Reid (2003) in the non-time series context when there are no nuisance parameters. Our analysis relaxes these conditions, although imposes another one which will appear in equation (B.1) which controls the rate of convergence. We will assume consistency.

To study the properties of the CL estimator it is helpful to stack the moment constraints and estimators. Write

$$\frac{1}{T}\sum_{t=1}^{T} k_{t,T}(\widehat{\phi}_T) = 0,$$

where

$$\phi_T = \left(\lambda_1', ..., \lambda_{N_T}', \theta'\right)'$$

and

$$k_{t,T}(\phi_T) = \left(g_{1t}', ..., g_{N_T t}', \frac{1}{N_T}\sum_{j=1}^{N_T}\frac{\partial l_{jt}}{\partial\theta'}\right)'.$$

We assume for all t and T that $k_{t,T}$ is continuously differentiable in all elements of $\phi_T \in \Psi$, where Ψ is a compact subset of $\mathbb{R}^{\dim(\phi_T)}$.

Then by the mean value theorem

$$0 = \frac{1}{T}\sum_{t=1}^{T} k_{t,T}(\phi_T) + D_T(\overline{\phi}_T)\left(\widehat{\phi}_T - \phi_T\right),$$

where for each element of ϕ_T, $\overline{\phi}_{j,T} \in \min\left(\widehat{\phi}_{jT}, \phi_{jT}\right), \max\left(\widehat{\phi}_{jT}, \phi_{jT}\right)$.

D_T has the important block structure

$$D_T(\phi_T) = \begin{pmatrix} A_T(\phi_T) & C_T(\phi_T) \\ B_T(\phi_T) & J_T(\phi_T) \end{pmatrix},$$

where

$$\begin{aligned}
A_T(\phi_T) &= \frac{1}{T}\sum_{t=1}^{T} diag\left(\frac{\partial g_{1t}}{\partial \lambda_{1t}'},...,\frac{\partial g_{N_T t}}{\partial \lambda_{N_T t}'}\right),\\
B_T(\phi_T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial^2 l_{1t}}{\partial\theta\lambda_{1t}'},...,\frac{\partial^2 l_{N_t t}}{\partial\theta\partial\lambda_{N_T t}'}\right),\\
C_T(\phi_T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial g_{1t}}{\partial\theta'},...,\frac{\partial g_{N_T t}}{\partial\theta'}\right)',\\
J_T(\phi_T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_T}\sum_{j=1}^{N_T}\frac{\partial^2 l_{jt}}{\partial\theta\partial\theta'}\right).
\end{aligned}$$

The block diagonality of A_T is vital here as it is an extremely large dimensional matrix — which if unstructured would be difficult to deal with computationally and statistically.

Now our sole interest is in $\widehat{\theta}-\theta$ so we focus only on the lower block of the inverse of $D_T(\phi_T)$. It is of the form of the low dimensional variables

$$\widehat{\theta}-\theta = -D_{\theta\theta,T}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_{t,T}\right)$$

where

$$D_{\theta\theta,T} = J_T - B_T A_T^{-1} C_T, \quad F_{j,T} = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\lambda_{jt}'}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\lambda_{jt}'}\right)^{-1}$$

and the average projected score

$$Z_{t,T} = \frac{1}{N_T}\sum_{j=1}^{N_T}\left(\frac{\partial l_{jt}}{\partial\theta'} - F_{j,T} g_{jt}\right).$$

Notice that

$$D_{\theta\theta,T} = \frac{1}{N_T}\sum_{j=1}^{N_T} D_{j,\theta\theta,T}$$

where

$$D_{j,\theta\theta,T} = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\partial\theta'}\right) - \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\lambda_{jt}'}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\lambda_{jt}'}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\theta'}\right).$$

Assumption 4 means

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} Z_{t,T} \xrightarrow{d} N(0,\mathcal{I}_{\theta\theta}), \tag{B.1}$$

and $\mathcal{I}_{\theta\theta} > 0$.

We will assume as $T\to\infty$ that

$$D_{\theta\theta,T} \xrightarrow{p} \mathcal{D}_{\theta\theta} > 0.$$

Taken together this delivers the result using Slutsky's theorem.

A.3 Scalar BEKK simulation

Here we report the results from repeating the experiments discussed in Section 4 but on the scalar BEKK model given in Example 1. In this experiment the same values of α and β are used but with Ψ being replaced by Σ.

The results are presented in Table 10, their structure exactly follows that discussed for the cDCC model given in Section 4.

	Bias						RMSE					
	MMLE		MCLE		MSCLE		MMLE		MCLE		MSCLE	
N	α	β	α	β	α	β	α	β	α	β	α	β
$\alpha = .02, \beta = .97$												
3	.000	-.005	.000	-.005	.000	-.006	.005	.009	.005	.010	.006	.012
10	-.001	-.003	.000	-.004	.000	-.004	.002	.004	.003	.006	.003	.007
50	-.005	-.000	.000	-.004	.000	-.004	.005	.001	.002	.005	.002	.005
100	-.009	-.001	.000	-.004	.000	-.004	.009	.001	.002	.005	.002	.005
$\alpha = .05, \beta = .93$												
3	-.000	-.008	-.000	-.009	.000	-.010	.008	.023	.009	.025	.010	.029
10	-.001	-.005	-.000	-.007	-.000	-.007	.003	.009	.005	.014	.006	.015
50	-.006	-.003	-.000	-.006	-.000	-.006	.006	.004	.003	.009	.003	.009
100	-.012	-.004	-.000	-.006	-.000	-.006	.012	.004	.003	.009	.003	.009
$\alpha = .10, \beta = .80$												
3	-.001	-.005	-.001	-.006	-.001	-.006	.013	.028	.014	.030	.015	.033
10	-.003	-.003	-.001	-.005	-.001	-.005	.006	.011	.009	.019	.009	.019
50	-.014	.001	-.001	-.005	-.001	-.005	.015	.004	.006	.012	.006	.012
100	-.026	.001	-.001	-.005	-.001	-.005	.026	.003	.006	.012	.006	.012

Table 10: *Bias and RMSE results from a simulation study for the covariance estimators of the covariance targeting scalar BEKK model. We only report the estimates of α and β and their sum. The estimators include the subset CL (MSCLE), the full CL (MCLE), and the m-profile likelihood (MMLE) estimator. All results based on $2,500$ replications.*