Fitting vast dimensional time-varying covariance models[*]

Cavit Pakel
Department of Economics, Bilkent University
06800 Ankara, Turkey
cavit.pakel@bilkent.edu.tr

Neil Shephard
Department of Economics, Harvard University, Littauer Center
and Department of Statistics, Harvard University, Science Center
Cambridge, MA 02138, USA
shephard@fas.harvard.edu

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

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

September 16, 2014

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 existing 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.

1 Introduction

The estimation of time-varying covariances between the returns on hundreds of assets is a key input in modern risk management and asset allocation. Typically this is carried out by calculating the

[*] An earlier version of this paper was circulated in 2008 under the same title but with authors Engle, Shephard and Sheppard.

sample covariance matrix based on the last 100 or 250 days of data or through an 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), where the volatilities of each asset were allowed to vary through time but the correlations were time invariant. Surveys of more sophisticated models are given by Bauwens, Laurent, and Rombouts (2006), Silvennoinen and Teräsvirta (2009) and Engle (2009a).

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 (2012) and the MacGyver estimation method of Engle (2009b). Engle and Kelly (2012) assume that the time-changing correlation amongst assets is common across the cross-section of L assets, allowing the log-likelihood to be computed in $O(L)$ calculations. However, that 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. Formally this is a special case of the scalar integrated BEKK process discussed by Engle and Kroner (1995). It has been widely used in industry and was until recently the only viable proposed method.

An alternative method was suggested by Engle (2009b) where he fit many pairs of bivariate estimators, governed by simple dynamics, and then took a median of these estimators. This method requires $O(L^2)$ calculations, is not invariant to reparameterisation and formalising this method in order to conduct inference is difficult. Our method has some similarities to the Engle (2009b) strategy but is more efficient and is invariant.

A further set of papers advocate methods which can be used on moderately high dimensional problems, such as 50 assets. The first was the covariance targeting and scalar dynamics BEKK model of Engle and Kroner (1995), the second was the DCC model introduced by Engle (2002) and studied in detailed by Engle and Sheppard (2001) — recent developments in this area include Aielli (2013) and Engle (2009a). 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 dimensions of many hundreds, 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. We provide a proof of the consistency of the composite likelihood estimator in the presence of incidental parameters and show that the estimator is asymptotically normal. Multivariate volatility modelling is typically applied to datasets that have long time-series dimensions but a much smaller number of assets. An important implication of our study then is that the scope of multivariate volatility modelling now encompasses both fixed-N large-T and large-N large-T panels.

The theoretical analysis is general in the sense that it focuses on a generic likelihood estimation problem for large-N large-T nonlinear dynamic panels with incidental parameters. In this part, we build upon Pakel (2014) who analyses the first-order bias of the integrated composite likelihood estimator under different types of dependence. We extend their results for the strong dependence setting by allowing for a vector-valued nuisance parameter.

Our approach 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, e.g. King, Sentana, and Wadhwani (1994), Fiorentini, Sentana, and Shephard (2004), Engle, Ng, and Rothschild (1990) and Rangel and Engle (2012).

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 targeting, 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 models in order to carry out inference on a large model. We show this method has a hidden incidental parameter problem and that the use of composite likelihoods largely overcomes this problem. The formal theoretical analysis and the main results are presented in Section 4. Section 6 provides a Monte Carlo investigation comparing the finite sample properties of our estimator with the quasi-maximum likelihood. Section 7 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 8 we discuss some important additional topics. Section 9 concludes, while the Appendix contains proofs.

2 The framework

Our primary objective is large scale modelling in a statistically parsimonious and computationally efficient fashion. Let r_{lt}, be the (log) return on individual asset l at time t, where $l = 1, ..., L$ and $t = 1, ..., T$. Data are assumed to exhibit both time and cross-section dependence. Typically, the interest would be on estimating the time-varying conditional covariance matrix,

$$Cov(r_t|\mathcal{F}_{t-1}) = H_t,$$

where $r_t = (r_{1t}, ..., r_{Lt})'$, $\mathcal{F}_t$ is the information set at time t and $\mathbb{E}[r_t|\mathcal{F}_{t-1}] = 0$. H_t is modelled parametrically, indexed by a parameter vector ψ. Two examples are below.

Example 2.1 *The scalar BEKK (Baba, Engle, Kraft and Kroner) model,*

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

which is a special case of Engle and Kroner (1995). Typically, this model is completed by setting $H_0 = \Sigma$. *Hence,* $\psi = (\lambda', \theta')'$, *where* $\lambda = vech(\Sigma)$ *and* $\theta = (\alpha, \beta)'$.

Example 2.2 *Nonstationary covariances with scalar dynamics,* $H_t = \alpha r_{t-1}r_{t-1}' + (1 - \alpha)H_{t-1}$, $\alpha \in [0, 1)$. *A simple case of this EWMA dynamics is RiskMetrics, which puts* $\alpha = 0.06$ *for daily returns and* 0.03 *for monthly returns. Inference is typically made conditional on* $\lambda = vech(H_0)$, *which has to be estimated. Therefore,* $\psi = (\lambda', \theta)'$, *where* $\lambda = vech(H_0)$ *and* $\theta = \alpha$.

Section 5 discusses the popular Dynamic Conditional Correlation (DCC) model of Engle (2002). Estimation in these multivariate GARCH style models gets increasingly complicated as L increases. To illustrate, consider the popular setting of Gaussian normality where $r_t|\mathcal{F}_{t-1} \sim N(\mathbf{0}, H_t)$, where $\mathbf{0}$ is a $(L \times 1)$ vector of zeroes. Then, the quasi-likelihood is given by

$$\log L(\psi; r) = \sum_{t=1}^{T} \ell(\psi; r_t), \quad \text{where} \quad \ell(\psi; r_t) = -\frac{1}{2}\log|H_t| - \frac{1}{2}r_t' H_t^{-1} r_t. \tag{1}$$

with $r = (r_1', ..., r_T')'$. Estimating ψ by the maximum likelihood method using (1) is prone to several issues in high-dimensional models. The first problem is the modelling of H_t which has $O(L^2)$ parameters. Although this curse of dimensionality issue could be solved by parameter reduction methods, such as factor modelling, there still remains a computational issue: the solution to the optimisation problem will not exist in closed form, requiring numerical optimisation. This in turn implies that H_t^{-1} will have to be calculated for each t, and many times until the numerical optimiser converges; a daunting task even for moderate L as the computational load is $O\left(L^3\right)$. This paper's objective is to side-step these issues.

In these examples, the parameter vector ψ was divided into two parts: λ and θ. Typically parameters that govern the volatility dynamics (θ in our examples) would be considered as parameters of interest while the remaining parameters, λ, would be considered nuisances. The dimension of the nuisance parameter is an important issue. In Example 2.1, Σ has $L(L+1)/2$ free parameters, which will be vast if L is large. Similar issues arise in many multivariate models.

2.1 Empirical illustration

Here we estimate the models given in Examples 2.1 and 2.2 (and the DCC model discussed in Section 5) 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 $L \leq 96$. To allow L to increase, which allows us to assess the sensitivity to L, we set the first asset as the market and the other assets are arranged alphabetically by ticker (for stocks that changed tickers during the sample, the ticker on the first day of the sample was used). The estimated θ parameters from an expanding cross-section of assets are contained in Table 1. Throughout θ is estimated using the conventional multistep procedures for each model, which we refer to as 2MLE here and spelt out later in the paper.

	S&P 100 Components					S&P 500 Components				
	Scalar BEKK		EWMA	DCC			Scalar BEKK		DCC	
L	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$	L	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
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 (2MLE). 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.

The empirical results suggest the increasing L destroys the 2MLE as $\widetilde{\alpha}$ falls dramatically as L increases. These results will be confirmed by detailed simulation studies in Section 6 which produce the same results by simulating BEKK or DCC models and then estimating them using 2MLE techniques. In addition Section 7 suggests the 2MLE parameter values when $L = 96$ are poor when judged using a simple economic criteria.

These results are reinforced by an empirical study using the same type of database, but now based on the corresponding components of the S&P 500. Including the index this produces a dataset with $L = 480$. The results in Table 1 show dramatic distortions — where the estimated β also crash towards zero as L increases.

We now turn to our preferred estimator which allows L to have any relationship to T.

3 The composite likelihood method

3.1 Main ideas

Our strategy in dealing with these statistical and computational issues will be to use a composite likelihood. The main idea behind this is that if the joint density is difficult to specify or to work with, then one can instead use an approximation based on lower dimensional marginal densities. This idea can be traced back to Lindsay (1988).[1] The simplest composite log-likelihood is given by

$$\sum_{l=1}^{L}\sum_{t=1}^{T}\ell(\psi; r_{kt}), \tag{2}$$

where $\sum_{t=1}^{T}\ell(\psi; r_{kt})$ is the log-likelihood for the l^{th} individual. Obviously, if data are independent across L, then (2) coincides with the true joint likelihood function.[2] In this study we will focus on composite likelihoods constructed from bivariate marginal densities. The objective is to obtain a consistent and computationally fast valid estimator of the parameter of interest by reducing the dimensionality of the problem from L to 2. Beyond computational gains, estimation based on lower dimensional marginal densities is also useful in a robustness sense: one is more likely to misspecify the more complicated joint density compared to the simpler univariate or bivariate marginal densities (Xu and Reid (2011)). Hence, composite likelihood estimation is potentially more robust to misspecification. There will, of course, be some efficiency loss. However, this would very much depend on the particular model at hand.

We now introduce the notation. Let the pairs of observations be given by $Y_{jt} = (r_{j_1 t}, r_{j_2 t})$ where $(j_1, j_2) \in \{1, ..., L\}^2$ and $j_1 \neq j_2$ for all $j = 1, ..., N$. Obviously, all cases where $(r_{i_1 t}, r_{i_2 t}) = (r_{j_1 t}, r_{j_2 t})$ and $i \neq j$ are ruled out, in order to exclude redundant pairs from analysis. N depends on the particular sampling approach. For example, if all unique bivariate pairs are considered, then $N = L(L-1)/2$. Another possibility is to take all (or some) contiguous pairs:

$$Y_{1t} = (r_{1t}, r_{2t}), \quad Y_{2t} = (r_{2t}, r_{3t}), \ldots \quad Y_{Nt} = (r_{L-1,t}, r_{L,t}),$$

where $N = O(L)$. These samples are studied here, though other sampling strategies are interesting.

[1] See also Cox and Reid (2004), Varin and Vidoni (2005), Varin (2008) and Varin, Reid, and Firth (2011).

[2] This type of marginal analysis has appeared before outside the time-series statistics literature. Examples include Besag (1974) in his analysis of spatial processes, Fearnhead (2003) in bioinformatics, deLeon (2005) on grouped data, Kuk and Nott (2000) and LeCessie and van Houwelingen (1994) for correlated binary data. Cox and Reid (2004) discuss this problem in the non-time-series case.

Implicit here is that for each pair j the lower dimensional likelihood $\ell(\theta, \lambda_j; Y_{jt})$ are such that

$$(\hat{\theta}, \hat{\lambda}_j) = \arg\max_{\theta, \lambda_j} \sum_{t=1}^{T} \ell(\theta, \lambda_j; Y_{jt})$$

is a consistent estimator of the (pseudo) true parameter vector (θ_0, λ_{j0}) as $T \to \infty$. In other words, in our particular case, the bivariate likelihood functions carry enough information to consistently estimate (θ_0,λ_{j0}). However, more information can be obtained by using the composite likelihood function, formed by averaging across all pairs in the sample:

$$CL_{NT}(\psi) = \frac{1}{NT} \sum_{j=1}^{N} \sum_{t=1}^{T} \ell_{jt}(\theta, \lambda_j), \tag{3}$$

where $\ell_{jt}(\theta, \lambda_j) = \ell(\theta, \lambda_j; Y_{jt})$. What is essentially being done here is pooling the information in the dataset in order to estimate (θ_0, λ_{j0}), without having to use the complicated full likelihood function. Hence, using the whole information pool for estimation is now much cheaper in terms of computational costs. Another advantage is that in short panels, where time-series variation is not sufficient, the individual specific parameter λ_{j0} might be poorly identified, due to an almost flat score function. Averaging across pairs would improve the curvature of the score and, thus, improve identification of the pair-specific parameter.

Depending on the particular sampling strategy, computational gains can be substantial. Evaluation of $CL_{NT}(\psi)$ costs $O(N)$ calculations. In the case where all distinct pairs is used, this means that the CL costs $O(L^2)$ calculations - which is distinctively better than the $O(L^3)$ implied by the full likelihood, based on the joint density. One can also use the subset of contiguous pairs or an economically motivated selection like the so called "beta CL" discussed in Section 8.1 which is based on all pairs involving the market index returns. Both of these would cost $O(L)$ calculations. When it comes to constructing the composite likelihood function, another alternative is to choose only $O(1)$ pairs, which is computationally faster. It is also tempting to randomly select N pairs and make inference conditional on the selected pairs as the selection is strongly exogenous. We will see in a moment that the efficiency loss of using large, but ultimately only $O(1)$ numbers of subsets compared to computing all possible pairs can be extremely small.

3.2 Parameter Space for λ_j

A peculiar feature of λ_j is that, there is no guarantee that there will be no links between λ_i and λ_j for different values of i and j. For instance, in the scalar BEKK model of Example 2.1, for $Y_{1t} = (r_{1t}, r_{2t})'$ and $Y_{2t} = (r_{2t}, r_{3t})'$, one has

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

Although it is possible to make gains in estimation by using these links, extension towards this direction would not be trivial. Therefore, we leave this interesting avenue for future research. Then, the explicit assumption is that λ_{j0} are variation free (Engle, Hendry, and Richard (1983)), in the sense that $(\lambda_1, ..., \lambda_N) \in \Lambda_1 \times ... \times \Lambda_N$, where Λ_j is the parameter space for λ_j. In the BEKK example this is achieved by having no pairs that contain the same r_{lt}.

Remark 1 *Ignoring possible links means that estimation for λ_i should be carried out based solely on the time variation for the j^{th} pair, using $Y_{j1}, ..., Y_{jT}$. Thus the 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 imposed on Σ however, some care needs to be taken that* $\dim(Y_{jt})$ *is larger than the dimension of the factor.*

Of course, this risks efficiency loss. However, our experiments which use cross-submodel constraints, not reported here, indicate that the efficiency loss in practice is tiny when N is large.

3.3 Theoretical framework: nonlinear and dynamic panels with incidental parameters

Although the primary focus of this study is volatility modelling, the theoretical setting considered here has a general scope as the analysis is based on a generic likelihood estimation problem. In Section 4, we will make specific assumptions on the dependence structure of data, smoothness properties of the likelihood function and the existence of moments. However, the main theoretical results will not be with respect to a particular model, such as BEKK or DCC.

The estimation problem of this paper belongs to the more general class of estimation in the presence of incidental parameters in a nonlinear and dynamic panel data model with large-N large-T asymptotics. To illustrate, let the full joint likelihood function be given by $\ell(\theta, \lambda_1, ..., \lambda_N; Y)$ where Y is the $(N \times T)$ data matrix. Then, the concentrated (or profile) likelihood estimator, based on the joint density, is given by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta, \hat{\lambda}_1(\theta), ..., \hat{\lambda}_N(\theta); Y) \quad \text{and} \quad \hat{\lambda}_j(\theta) = \arg\max_{\lambda_j \in \Lambda_j} \sum_{t=1}^{T} \ell(\theta, \lambda_j; Y_{jt}), \tag{4}$$

In the case of the composite likelihood method, we accordingly have

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{j=1}^{N} \sum_{t=1}^{T} \ell_{jt}(\theta, \hat{\lambda}_j(\theta)) \quad \text{and} \quad \hat{\lambda}_j(\theta) = \arg\max_{\lambda_j \in \Lambda_j} \sum_{t=1}^{T} \ell(\theta, \lambda_j; Y_{jt}), \tag{5}$$

Here, $\lambda_1, ..., \lambda_N$ are called the incidental or nuisance parameters. It is well known that estimation in this setting is prone to the incidental parameter issue, first analysed by Neyman and Scott (1948). There is a rich literature dealing with this issue in statistics (e.g. Barndorff-Nielsen (1983), Cox and Reid (1987), McCullagh and Tibshirani (1990) and Sartori (2003)). A classic reference in econometrics is Nickell (1981). Recently, there has been a particular interest on this problem in the microeconometrics literature, within the framework of estimation in the presence of unobserved heterogeneity (e.g. Hahn and Kuersteiner (2011), Hahn and Newey (2004), Carro (2007), Arellano and Bonhomme (2009), Bester and Hansen (2009), Fernández-Val (2009), Dhaene and Jochmans (2011) and Pakel (2014). Lancaster (2000) and Arellano and Hahn (2007) provide detailed surveys). We will adopt a similar theoretical setting as in and build upon Pakel (2014) who allows for both serial and cross-section dependence, which is appropriate for financial data.

The classical incidental parameter bias in large-N large-T panel data models under cross-section independence is asymptotically non-vanishing in the following sense: the estimation error associated with each $\hat{\lambda}_j$ gets accumulated in $\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}(\theta,\hat{\lambda}_j(\theta))$ (the joint likelihood function under cross-section independence) as $N\to\infty$. Then, as $N,T\to\infty$, despite the increasing information from the time-series dimension, the accumulated estimation error remains sufficiently large, leading to $\sqrt{NT}(\hat{\theta}-\theta_0)\overset{d}{\to}N\left(\mathcal{B},\Omega\right),\mathcal{B}=O(\sqrt{N/T})$ as $N,T\to\infty$, where Ω is some asymptotic covariance matrix. This necessitates the use of bias correction methods. An important distinction of our paper is that, under the dependence setting considered here, this bias turns out to be an asymptotically vanishing time-series small-sample bias. Of course, some bias correction might still be desirable in small samples, but, as the simulation results later reveal, this is not required for the sample sizes we consider.

In volatility modelling, often we can side step the optimising over λ by concentrating at some moment based estimator $\tilde{\lambda}$, by using the *covarince-targeting* idea of Engle and Mezrich (1996). In the particular case of Example 2.1, they suggest using $\widehat{\Sigma}=T^{-1}\sum_{t=1}^{T}r_tr_t'$ in which case $\widehat{\lambda}=vech(\widehat{\Sigma})$. By the same idea, in Example 2.2 one can put $\widehat{H}_0=T^{-1}\sum_{t=1}^{T}r_tr_t'$ and $\widehat{\lambda}=vech(\widehat{H}_0)$.[3] Appropriate two-step estimator versions of (4) and (5) are then given by

$$\hat{\theta}=\arg\max_{\theta\in\Theta}\ell(\theta,\tilde{\lambda}_1,...,\tilde{\lambda}_N;Y)\quad\text{ and }\quad\hat{\theta}=\arg\max_{\theta\in\Theta}\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}(\theta,\tilde{\lambda}_j),$$

respectively. These estimators are not concentrated likelihood estimators anymore, since $\tilde{\lambda}_j$ are not likelihood based. Instead, they can be considered as two-step GMM estimators (see Newey and McFadden (1994)). We will provide the theory for the concentrated likelihood estimator, but will use the moments based approach in applications. Note that in large samples the two approaches are expected to deliver similar results. In what follows, we refer to (4) as 2MLE. In the case of (5),

[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.

L	2MLE	2MCLE	2MSCLE
5	24s	0.1s	0.2s
25	46s	2.1s	0.2s
50	2m 10s	10s	0.5s
100	1h 50m	39s	0.8s
250	15h 11m	4m 7s	1.6s
480	85h 33m	18m 6s	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.*

2MCLE corresponds to the case where all unique bivariate pairs are used, while 2MSCLE is the CL estimator based on contiguous pairs. We finish this part by presenting some computational times for a problem based on modelling up to 480 assets (Table 2). Clearly, the concentrated likelihood approach does deliver in terms of computational efficiency. A detailed discussion of this will be given in Section 7.4.

4 Dependence structure and large sample theory

4.1 Dependence setting

The main idea underlying the dependence setup is mixing-type dependence across time and a strong-type of cross-sectional dependence. The essence of mixing is that although a mixing random variable is not independently distributed, dependence between observations at two points in time vanishes as the time distance increases.

Mixing is a property related to the sigma-algebras generated by random variables. Therefore, we start by defining some relevant sigma-fields:

$$\mathcal{F}^t_{j,-\infty} = \sigma(..., Y_{j,t-1}, Y_{jt}), \quad \mathcal{F}^\infty_{j,t+m} = \sigma(Y_{j,t+m}, Y_{j,t+m+1}, ...),$$
$$\mathcal{F}^t_{ij,-\infty} = \sigma(..., Y_{i,t-1}, Y_{j,t-1}, Y_{it}, Y_{jt}), \quad \mathcal{F}^\infty_{ij,t+m} = \sigma(Y_{i,t+m}, Y_{j,t+m}, Y_{i,t+m+1}, Y_{j,t+m+1}, ...).$$

Note that the sigma-field generated by some sequence of Y_{jt} will essentially be a sigma-field generated by a corresponding sequence of r_{lt}. For example, $\mathcal{F}^t_{j,-\infty} = \sigma(..., r_{j_1,t-1}, r_{j_2,t-1}, r_{j_1t}, r_{j_2t})$ and $\mathcal{F}^\infty_{j,t+m} = \sigma(r_{j_1,t+m+1}, r_{j_2,t+m+1}, r_{j_1,t+m+2}, r_{j_2,t+m+2}, ...)$.

Definition 4.1 (α-mixing) *For two events $G \in \mathcal{G}$ and $H \in \mathcal{H}$, where $\mathcal{G}$ and $\mathcal{H}$ are some sigma-fields, the α-mixing coefficient is defined as $\alpha(\mathcal{G}, \mathcal{H}) = \sup\{|P(G \cap H) - P(G)P(H)| : G \in \mathcal{G}, H \in \mathcal{H}\}$. A random sequence $\{Y_{jt}\}_{t=1}^T$ is called α-mixing if $\lim_{m\to\infty} \sup_t \alpha(\mathcal{F}^t_{j,-\infty}, \mathcal{F}^\infty_{j,t+m}) = 0$.*

As we focus on a panel of random variables, we have to consider some slightly more complicated dependence structures. For example, what is the magnitude of dependence between Y_{it} and

$Y_{j,t+m}$ (as opposed to the magnitude of dependence between observations on the *same* pair at different points in time, e.g. Y_{it} and $Y_{i,t+m}$)? Therefore, it is necessary to introduce the following mixing coefficients: $\alpha_{i,j}(m) = \sup_t \alpha(\mathcal{F}^t_{i,-\infty}, \mathcal{F}^\infty_{j,t+m})$, $\alpha_{ij,k}(m) = \sup_t \alpha(\mathcal{F}^t_{ij,-\infty}, \mathcal{F}^\infty_{k,t+m})$ and $\alpha_{i,jk}(m) = \sup_t \alpha(\mathcal{F}^t_{i,-\infty}, \mathcal{F}^\infty_{jk,t+m})$, where $i, j, k = 1, ..., N$. Intuitively, $\alpha_{i,j}(m)$ measures the dependence between m-period apart observatons on pairs i and j (note that in what follows, we use $\alpha_i(m) = \alpha_j(m) = \alpha_{i,j}(m)$ whenever $i = j$). On the other hand, $\alpha_{ij,k}(m)$ measures the dependence between m-period apart observations that belong to (i) the sigma-field generated by pairs i and j together and (ii) the sigma-field generated by pair k. Such coefficients will be useful for understanding the dependence properties of non-standard covariances, such as $Cov(r_{i_1t}r_{j_2t}, r_{k_1t})$.

These mixing coefficients are slightly different from the usual ones, such as $\alpha_i(m)$ which controls the dependence structure for a single time-series only. Our modification is due to the necessity that in financial panels one has to allow for dependence between observations on two different pairs that are m units apart in time. If $\lim_{m\to\infty} \alpha_{i,j}(m) = 0$, then this dependence is restricted to be of α-mixing type. Therefore, for instance, the daily return on the IBM equity at time t is allowed to have a mixing-type dependence with the return on the JP Morgan equity at time $t + m$. The essential idea is that dependence between any two observations is determined entirely by their time distance, independent of the cross-sectional indices. Implicitly, the only assumption made on contemporaneous cross-section dependence is that it is finite uniformly across all pairs and time periods, which allows for strong dependence across cross-section. Dependence can be weakened by assuming some form of weak dependence across cross-section but we do not pursue this route here.

4.2 Assumptions

We consider a multidimensional setting where $\dim(\lambda_j) = P$ and $\dim(\theta) = R$. The (pseudo) true parameter values are given by $(\lambda_{10}, ..., \lambda_{N0})$ and θ_0. In what follows, we use the following short hand notation:

$$\ell_{jT}(\theta, \lambda_j) = \frac{1}{T}\sum_{t=1}^{T} \ell_{jt}(\theta, \lambda_j), \qquad \ell_{NT}(\theta, \lambda) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} \ell_{jt}(\theta, \lambda_j),$$

$$\ell^{\lambda}_{jT}(\theta, \hat{\lambda}(\theta)) = \frac{\partial}{\partial\lambda}\ell_{jT}(\theta, \hat{\lambda}(\theta)), \qquad \ell^{\theta\lambda}_{jT}(\theta, \hat{\lambda}(\theta)) = \frac{\partial^2}{\partial\theta\partial\lambda'}\ell_{jT}(\theta, \hat{\lambda}(\theta)),$$

$$\ell^{\theta\theta}_{jT}(\theta, \hat{\lambda}(\theta)) = \frac{\partial^2}{\partial\theta\partial\theta'}\ell_{jT}(\theta, \hat{\lambda}(\theta)), \qquad \ell^{\lambda\theta}_{jT}(\theta, \hat{\lambda}(\theta)) = [\ell^{\theta\lambda}_{jT}(\theta, \hat{\lambda}(\theta))]',$$

$$\nabla_\theta \ell^{\lambda}_{jT}(\theta, \hat{\lambda}(\theta)) = \frac{d}{d\theta}\frac{\partial}{\partial\lambda'}\ell_{jT}(\theta, \hat{\lambda}(\theta)) \qquad \text{etc.}$$

Therefore, partial differentiation is denoted by superscripts, while total derivatives are denoted by the gradient operator ∇. Whenever a term is evaluated at (θ_0, λ_{j0}), the arguments are dropped for conciseness. Hence, for instance, $\ell_{jT} = \ell_{jT}(\theta_0, \lambda_{j0})$. Also, all moments are evaluated with respect to the true underlying density.

A further likelihood concept, which we use as the theoretical benchmark in the asymptotic expansions, is the *target likelihood.* The target likelihood estimator is given by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{j=1}^{N}\sum_{t=1}^{T} \ell_{jt}(\theta, \bar{\lambda}_{jT}(\theta)) \quad \text{and} \quad \bar{\lambda}_{jT}(\theta) = \arg\max_{\lambda_j\in\Lambda_j} \sum_{t=1}^{T} \mathbb{E}[\ell(\theta, \lambda_j; Y_{jt})]. \tag{6}$$

This looks similar to the concentrated likelihood estimator. However, a major difference is that $\bar{\lambda}_{jT}(\theta)$ is an infeasible quantity as it depends on θ_0 and λ_{j0}, as well as θ. In essence, $\hat{\lambda}_j(\theta)$ is an estimator of $\bar{\lambda}_{jT}(\theta)$. Consequently, $\ell_{jt}(\theta, \hat{\lambda}_j(\theta))$ can be shown to be asymptotically equivalent to $\ell_{jt}(\theta, \bar{\lambda}_{jT}(\theta))$. Importantly, $\bar{\lambda}_{jT}(\theta_0) = \lambda_{j0}$ and the target likelihood is maximised at θ_0. See Severini and Wong (1992) for a more detailed discussion.

We use the following notation to denote likelihood derivatives of any order: Let $r = \{r_1, ..., r_R\}$ and $p = \{p_1, ..., p_P\}$ be sets of R and P non-negative integers where $|r| = \sum_{i=1}^{R} r_i$ and $|p| = \sum_{i=1}^{P} p_i$. Then, define $f_{jt}^{(r,p)}(\theta, \lambda) = \frac{d^{(|r|+|p|)}\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}...d\theta_{r_R}d\lambda_{p_1},...,d\lambda_{p_P}}$. Hence, for example, $f_{jt}^{(r,p)}(\theta, \lambda)$ where $|r| + |p| = 1$ yields the set of all first order derivatives. Similarly, the same term evaluated across all (r, p) such that $|r| + |p| \leq 2$ yields all derivatives up to and including the second order. As a final example, consider $f_{jt}^{(r,p)}(\theta, \lambda)$ where $|r| = 0$ and $|p| \leq 3$: this encompasses all the derivatives of the likelihood function with respect to λ up to and including the third order. Notice that $f_{jt}^{(r,p)}(\theta, \lambda)$ where $|r| + |p| = 0$ gives back the likelihood function. Finally, for any (r, p) let $\bar{f}_{jt}^{(r,p)}(\theta, \lambda) = f_{jt}^{(r,p)}(\theta, \lambda) - \mathbb{E}[f_{jt}^{(r,p)}(\theta, \lambda)]$.

Assumption 4.1 *(i) $\lambda \in \Lambda$ and $\theta \in \Theta$ where Λ and Θ are compact convex subsets of $\mathbb{R}^P$ and $\mathbb{R}^R$, respectively; (ii) N and T tend to ∞ such that $N/T \to c$ where c is some positive finite constant; (iii) $\ell_{jt}(\theta, \lambda) \in \mathcal{C}^5$ for all j and t, where $\mathcal{C}^c$ is the class of functions whose derivatives up to and including order c are continuous.*

Assumption 4.2 *(i) For all $(\theta', \lambda'), (\theta'', \lambda'') \in \Psi = \Theta \times \Lambda$ we have*

$$\left|\ell_{jt}(\theta', \lambda') - \ell_{jt}(\theta'', \lambda'')\right| \leq c(Y_{jt}) \left|\left|(\theta', \lambda') - (\theta'', \lambda'')\right|\right|,$$

where $c(\cdot)$ is a measurable function of Y_{jt} and $\sup_{j,t}\mathbb{E}[|c(Y_{jt})|] < \infty$; (ii) the same holds for all $f_{jt}^{(r,p)}(\theta, \lambda)$, where (r, p) is such that $(|r|, |p|) \in \{(0, 4), (1, 4), (2, 3), (5, 0), (3, 1), (4, 1)\}$. Note that the function $c(\cdot)$ is not necessarily the same across its all appearances.

Assumption 4.3 *(i) For any $\eta > 0$,*

$$\varepsilon = \inf_{1\leq j\leq N} \left[\mathbb{E}[\ell_{jT}(\theta_0, \lambda_{j0})] - \sup_{\{(\theta,\lambda):||(\theta,\lambda)-(\theta_0,\lambda_{j0})||>\eta\}} \mathbb{E}[\ell_{jT}(\theta, \lambda)]\right] > 0;$$

(ii) for any $\tilde{\eta} > 0$,

$$\tilde{\varepsilon} = \inf_{\theta\in\Theta} \inf_{1\leq j\leq N} \left[\mathbb{E}[\ell_{jT}(\theta, \bar{\lambda}_{jT}(\theta))] - \sup_{\{\lambda:||\lambda-\bar{\lambda}_{jT}(\theta)||>\tilde{\eta}\}} \mathbb{E}[\ell_{jT}(\theta, \lambda)] \right] > 0.$$

Assumption 4.4 *(i) The underlying random variable r_{lt}, $(l = 1, ..., L;\ t = 1, ..., T)$ is such that for any $j = 1, ..., N$, $\{Y_{jt}\}_{t=1}^{T}$ is an α-mixing process with mixing coefficients of size $-(2+\varepsilon)/\varepsilon$ for some $\varepsilon > 0$; (ii) $\lim_{m\to\infty} \alpha_{i,j}(m) = 0$, $\lim_{m\to\infty} \alpha_{ij,k}(m) = 0$, $\lim_{m\to\infty} \alpha_{i,jk}(m) = 0$ while $\sum_{m=1}^{\infty} m\alpha_{ij,k}(m)^{\varepsilon/3+\varepsilon} < \infty$ and $\sum_{m=1}^{\infty} m\alpha_{i,jk}(m)^{\varepsilon/3+\varepsilon} < \infty$ uniformly for all $i, j, k = 1, ..., N$; (iii) $Var[T^{-1/2}\sum_{t=1}^{T} \bar{f}_{jt}^{(r,p)}(\theta, \bar{\lambda}_{jT}(\theta))] > 0$ uniformly for all j, T, θ such that $|r| \leq 4$ and $|p| = 1$ or $|r| \leq 1$ and $|p| = 2$.*

Assumption 4.5 *Let $\left(\frac{\partial^2 \ell_{jT}(\theta,\lambda)}{\partial\lambda_j \partial\lambda_j'}\right)_{p_1,p_2}$ be the row p_1 and column p_2 entry of the matrix $\frac{\partial^2 \ell_{jT}(\theta,\lambda)}{\partial\lambda_j \partial\lambda_j'}$ and $\left(\frac{d^2 \ell_{NT}(\theta,\lambda)}{d\theta d\theta'}\right)_{r_1,r_2}$ be the row p_1 and column p_2 entry of the matrix $\left(\frac{d^2 \ell_{NT}(\theta,\lambda)}{d\theta d\theta'}\right)_{r_1,r_2}$. Then, (i) $\frac{\partial^2 \ell_{jT}(\theta,\lambda)}{\partial\lambda_j \partial\lambda_j'}$ and $\frac{d^2 \ell_{NT}(\theta,\lambda)}{d\theta d\theta'}$ are invertible and (ii) $\left(\frac{\partial^2 \ell_{jT}(\theta,\lambda)}{\partial\lambda_j \partial\lambda_j'}\right)_{p_1,p_2}$ and $\left(\frac{d^2 \ell_{NT}(\theta,\lambda)}{d\theta d\theta'}\right)_{r_1,r_2}$ are bounded, uniformly for all j, N, T, θ and λ_j.*

Assumption 4.6 *For $1 \leq |r| \leq 3$ and $|p| = 0$, $\sigma_{r,NT}^2 = Var\left[\sqrt{T}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} \bar{f}_{jt}^{(r,p)}(\theta_0, \lambda_{j0})\right]$ is positive for all N, T and*

$$\sqrt{T}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} \bar{f}_{jt}^{(r,p)}(\theta_0, \lambda_{j0}) \xrightarrow{d} N(0, \sigma_r^2) \quad \text{as } N, T \to \infty, \quad \text{where} \quad \sigma_r^2 = p\lim_{N,T\to\infty} \sigma_{r,NT}^2.$$

In addition $Var(\sqrt{T}d\ell_{NT}/d\theta)$ is positive definite.

Assumption 4.1 contains the standard conditions on the parameter space (e.g. Hahn and Kuersteiner (2011)) and formalises the double asymptotic setup where both dimension sizes tend to infinity at the same rate. This is a sensible setting for financial datasets where both the time and cross-section dimensions are of non-negligible sizes. The continuity assumption serves two purposes. First, the likelihood function is smooth enough to guarantee the existence of asymptotic expansions to sufficiently high orders. Second, it also implies that the objective function and its derivatives are measurable functions. This is useful since mixing properties of a sequence of random variables are directly inherited by measurable functions of their finite sub-sequences (see Doukhan (1994) and Bradley (2005)). Thus, for instance, existence of a law of large numbers (LLN) or central limit theorem (CLT) for a mixing sequence will imply the existence of the same large sample results for continuous transformations of the same mixing sequence, under standard regularity conditions. Assumption 4.2 is a Lipschitz continuity type assumption, which imposes further smoothness conditions on some particular derivatives of the likelihood function. This is necessary for proving consistency and for bounding the remainder terms in mean value expansions. The first part of Assumption 4.3 is standard which imposes unique identifiability of (θ_0, λ_{j0}) for all j. The second part, which is not as standard, ensures that $\bar{\lambda}_{jT}(\theta)$ is the unique maximiser of $\mathbb{E}[\ell_{jT}(\theta, \lambda)]$ for all j, T. Note that the two parts of this assumption do not contradict since $\left(\theta_0, \bar{\lambda}_{jT}(\theta_0)\right) = (\theta_0, \lambda_{j0})$.

Dependence is formalised in Assumption 4.4. The first part establishes the standard α-mixing property for each individual series. The size assumption is necessary to invoke mixing LLNs and CLTs. In the second part, the α-mixing property across different pairs is established in the sense of the α-mixing coefficients introduced in Section 4.1. The summability conditions are required to bound the expectations of the terms appearing in expansions. Note that these assumptions are strong enough to generate other summability conditions such as $\sum_{m=1}^{\infty} m\alpha_{i,j}(m)^{\varepsilon/2+\varepsilon} < \infty$, $\sum_{m=1}^{\infty} \alpha_i(m)^{\varepsilon/3+\varepsilon} < \infty$ etc. The variance assumption of the final part is essential for obtaining CLTs for centred likelihood terms. When coupled with the size assumption and existence of $2+\varepsilon$ moments, this assumption implies a mixing CLT (see, e.g. White (2001)). Assumption 4.5 is required for the existence of the asymptotic expansions.

Assumption 4.6 is key to the existence of a large-N large-T CLT for the composite likelihood estimator $\hat{\theta}$. As usual, asymptotic normality of the estimator is achieved by the asymptotic normality of the score, $\nabla_\theta \ell_{NT}(\theta_0, \lambda_0)$. Let $Z_{t,T,N} = N^{-1} \sum_{j=1}^{N} (\ell^{\theta}_{jt} - \mathbb{E}[\ell^{\theta\lambda}_{jT}]\{\mathbb{E}[\ell^{\lambda\lambda}_{jT}]\}^{-1} \ell^{\lambda}_{jt})$. Then, this assumption implies that

$$\sqrt{T} \nabla_\theta \ell_{NT}(\theta_0, \lambda_0) = \sqrt{T} \frac{1}{T} \sum_{t=1}^{T} Z_{t,T,N} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}),$$

where $\mathcal{I} = \lim_{N,T\to\infty} Var\left[\sqrt{T} \nabla_\theta \ell_{NT}(\theta, \lambda_0)\right]$. Asymptotic normality for the second and third order derivatives with respect to θ are required for the higher order terms in the expansion. An important implication of this assumption is that due to strong dependence, cross-section information does not contribute to asymptotic convergence. Hence $\sqrt{T}$-convergence ensues. A similar assumption has also appeared in the recent paper by Gonçalves (2011). This is a high-level, yet intuitively reasonable assumption. Given that N is linearly related to T as implied by Assumption 4.1, one can also consider $T^{-1/2} \sum_{t=1}^{T} Z_{t,T,N_T} = T^{-1/2} \sum_{t=1}^{T} Z_{t,T}$ and use a triangular array CLT. Two examples are Bosq, Merlevède, and Peligrad (1999) and Utev (1991). The former study is made with kernel estimation in mind, and so is based on assumptions that would not be required in our case. However, diluting their assumptions is beyond the scope of our study. We leave the proof of the existence of a CLT valid for our case for future research. Finally, the moment conditions used in the paper are listed in Section B.1 of the Mathematical Appendix. These moment assumptions are sufficient enough to ensure that all contemporaneous covariance terms considered in the proofs are bounded uniformly across all i, j, k and t.

4.3 Main results

Our first main result is the consistency of the composite likelihood estimator $\hat{\theta}$ in the presence of incidental parameters.

Theorem 4.1 *Let Assumptions 4.1(i)-(ii), 4.2(i), 4.3, 4.4(i)-(ii) and B.1 hold.*[4] *Then, for all* $\theta \in \Theta$ *and* $\eta, \tilde{\eta} > 0$,

$$P\left[\max_{1\le j\le N}\left|\left|\hat{\lambda}_j(\theta) - \bar{\lambda}_{jT}(\theta)\right|\right| < \eta\right] = 1 - o(1) \quad \textit{and} \quad P\left[\left|\left|\hat{\theta} - \theta_0\right|\right| < \tilde{\eta}\right] = 1 - o(1).$$

Notice that, by the definition of the target likelihood, $\bar{\lambda}_{jT}(\theta_0) = \lambda_{j0}$ for all j and T. Therefore, Theorem 4.1 in particular implies that the composite likelihood estimator $(\hat{\theta}, \hat{\lambda}_j(\hat{\theta}))$ converges in probability to (θ_0, λ_{j0}). We need the more general version of convergence to $\bar{\lambda}_{jT}(\theta)$ because we use the target likelihood estimator as the theoretical benchmark in the asymptotic analysis of $\hat{\theta}$.

Theorem 4.2 *Let Assumptions 4.1-4.6 and B.1 hold. Then,*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathcal{D}^{-1}\mathcal{I}\mathcal{D}^{-1}) \quad \textit{as} \quad N, T \to \infty, \quad \textit{where}$$

$$\mathcal{D} = \lim_{N,T\to\infty} \mathbb{E}[\ell^{\theta\theta}_{NT}] - \left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\ell^{\theta\lambda}_{jT}]\right\}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\ell^{\lambda\lambda}_{jT}]\right\}^{-1}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\ell^{\lambda\theta}_{jT}]\right\}.$$

This result formalises the earlier claim that in our particular case, characterised by strong cross-section dependence and $\sqrt{T}$-convergence, the asymptotic distribution is not asymptotically biased anymore. Intuitively, with $\sqrt{T}$-consistency, the incidental parameter bias turns into a pure small-sample time-series bias. This $O(T^{-1})$ small sample bias is precisely characterised in Theorem 4.3 below. The crucial implication of this discussion is that our estimator is valid independent of how N is related to T as both tend to infinity. To be explicit, in small samples one might of course still have to do some bias correction, independent of whether the asymptotic distribution is biased or not. However, as will be demonstrated later, for the data dimensions we consider there will be no need to employ bias correction in our context.

Another important message of Theorem 4.2 is that with $\sqrt{T}$-convergence, the incidental parameter bias turns into purely a small-sample time-series bias. This is precisely characterised in Theorem 4.3 below. To be explicit, there is no need to employ bias correction in our context.

[4] Note that some of these assumptions are stronger than necessary. For example, we do not require all of the moment conditions listed in Assumption B.1. In fact, $\mathbb{E}\left[|\ell_{jt}(\theta,\lambda)|^{2+\varepsilon}\right] < \infty$ is sufficient for proving consistency. Similarly, the summability conditions of Assumption 4.4(ii) are stronger than the condition that $\sum_{m=1}^{\infty} \alpha_i(m)^{1/q-1/r} < \infty$ for some $r > q > 1$,which is a key condition in proving consistency. However, the stronger portions of these assumptions will be used elsewhere to prove Theorems 4.2 and 4.3. Especially when it comes to moment conditions, it is impossible to construct a separate non-overlapping list of conditions for each Theorem. Therefore, in order to prevent a large list of overlapping assumptions, we choose to group similar assumptions together.

Theorem 4.3 *Let Assumptions 4.1-4.6 and B.1 hold. Then* $\mathbb{E}[\hat{\theta}-\theta_0]=\mathbb{E}[\mathcal{A}_{NT}(\theta_0,\lambda_0)]+O\left(T^{-2}\right)$ *as* N,$T\to\infty$, *where* $\mathcal{A}_{NT}(\theta,\lambda)$ *depends on* N,T *and* $\mathbb{E}[\mathcal{A}_{NT}(\theta_0,\lambda_0)]=O(T^{-1})$ *as* $N,T\to\infty$.

The term $\mathcal{A}_{NT}(\theta,\lambda)$ has a complicated expression based on higher order likelihood derivatives. This is given in (28) in the Appendix.

5 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 2.1 and 2.2. Write the submodel based on a pair as

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

where we construct a model for the conditional variance $h_{ljt}=\mathrm{Var}(r_{ljt}|\mathcal{F}_{t-1},\eta_{lj})$, where η_{lj} are parameters.[5] This has a log-likelihood for the $\{r_{ljt}\}$ return sequence of

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

The devolatilised series are defined as

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

We build a model for R_{jt} using the cDCC dynamic introduced by Aielli (2013). It is defined as

$$\begin{aligned} R_{jt} &= P_{jt}^{-1/2}Q_{jt}P_{jt}^{-1/2},\quad P_{jt}=\left(\begin{array}{cc} Q_{11jt} & 0\\ 0 & Q_{22jt}\end{array}\right),\quad \text{where}\\ Q_{jt} &= \Psi_j\left(1-\alpha-\beta\right)+\alpha P_{jt-1}^{1/2}\left(S_{jt-1}S_{jt-1}'-R_{jt-1}\right)P_{jt-1}^{1/2}+\left(\alpha+\beta\right)Q_{jt-1},\quad \Psi_j=\left(\begin{array}{cc} 1 & \varphi_j\\ \varphi_j & 1\end{array}\right).\end{aligned}$$

It has the virtue that if we let $S_{jt}^*=P_{jt}^{1/2}S_{jt}$, then $\mathrm{Cov}\left(S_{jt}^*|\mathcal{F}_{t-1}\right)=P_{jt}^{1/2}R_{jt}P_{jt}^{1/2}=Q_{jt}$, 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

[5] The first step of fitting the cDCC models is to model $h_{lt}=\mathrm{Var}(r_{lt}|\mathcal{F}_{t-1})$. It is important to note that although it is common to fit standard GARCH models for this purpose, allowing the h_{lt} to depend the lagged squared returns on the l-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 $\mathrm{Var}(r_{lt}|r_{lt-1},r_{lt-2},...)$ 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, $\overline{r}_t=\frac{1}{L}\sum_{l=1}^L r_{l,t}$. The volatility was modeled using an EGARCH specification Nelson (1991),

$$\ln h_{\bullet,t}=\omega_\bullet+\alpha_\bullet|\epsilon_{\bullet,t-1}-\sqrt{2/\pi}|+\kappa_\bullet\epsilon_{\bullet,t-1}+\beta_\bullet\ln h_{\bullet,t-1},\quad \epsilon_{\bullet,t}=\overline{r}_t h_{\bullet,t}^{-1/2}. \tag{7}$$

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

$$\ln\tilde{h}_{l,t}=\omega_l+\alpha_l|\epsilon_{l,t-1}-\sqrt{2/\pi}|+\kappa_l\epsilon_{l,t-1}+\beta_l\ln h_{l,t-1},\quad h_{l,t}=h_{\bullet,t}\tilde{h}_{l,t},\quad \epsilon_{l,t}=r_{l,t}h_{l,t}^{-1/2}. \tag{8}$$

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

	Bias						RMSE					
	2MLE		2MCLE		2MSCLE		2MLE		2MCLE		2MSCLE	
L	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 (2MSCLE), CL (2MCLE), and likelihood (2MLE) estimators. Based on* $2,500$ *replications.*

corresponding ingredients into the estimation of θ from this model is the common structure

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

6 Monte Carlo experiments

6.1 Relative performance of estimators

Here we explore the effectiveness of three estimators of the parameters in the DCC model outlined in Section 5 above. These are the 2MLE, 2MCLE and 2MSCLE methods discussed in Section 3.3. The Appendix A 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 $h_{jt} = 1$. Throughout we used $T = 2,000$, L is one of $\{3, 10, 50, 100\}$ and the returns were simulated according to a cDCC model given in Section 5. Three choices spanning the range of empirically relevant values of the temporal dependence in the Q process were used $(\alpha, \beta) = (0.02, 0.97)$, $(0.05, 0.93)$ or $(0.10, 0.80)$. The parameters were estimated using a constraint that $0 \leq \alpha < 1$, $0 \leq \beta < 1$, $\alpha + \beta < 1$. None of the estimators hit the parameter space boundary.

The intercept Ψ was chosen to match the properties of the S&P 100 returns studied in the previous Section. The unconditional correlations were generated from a single-factor model, so that $\Psi_{l_1 l_2} = \pi_{l_1}\pi_{l_2}$ for $l_1 \neq l_2$ and 1 if $l_1 = l_2$. Here the π_l are distributed according to a truncated normal with mean 0.5, standard deviation 0.1 where the truncation occurs at ±4 standard deviations. This means $\pi \in (0.1, 0.9)$ and the average correlation in the cross section is 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 L.[6]

	Bias						RMSE					
	2MLE		2MCLE		2SCLE		2MLE		2MCLE		2SCLE	
T	α	β	α	β	α	β	α	β	α	β	α	β
						$L = 10$						
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
						$L = 50$						
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
						$L = 100$						
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
						$L = 200$						
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 (2MSCLE), CL (2MCLE), and likelihood (2MLE) estimators. Based on* $2,500$ *replications.*

Table 3 contains the bias and root mean square error of the estimates. The two-step maximum likelihood (2MLE) method develops a significant bias in estimating α as L increases. This is consistent with the findings of Engle and Sheppard (2001) and our earlier theoretical discussion on the nuisance parameter issue.

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

[6] 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.

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

Overall the Monte Carlo provides evidence of the 2MCLE has better RMSE for all cross-section sizes and parameter configurations. There seems little difference between the 2MCLE and 2MSCLE. In simulations not reported here, both estimators substantially outperform the Engle (2009b) McGyver estimator. The evidence presented here suggests 2MSCLE is attractive from statistical and computational viewpoints for large dimensional problems.

6.2 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 2MCLE and 2MSCLE. Both standard deviations rapidly decline as the cross-section dimension grows and the standard deviation of the 2MCLE is always slightly smaller than the 2MSCLE for a fixed cross-section size. Recall that the 2MCLE uses many more submodels than the 2MSCLE when the cross-section size is large, and so when $L = 50$ the 2MCLE is based on $1,225$ submodels while the 2MSCLE 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 L is around 50.

Another interesting feature of the Figure is the expected result that as L increases the standard error of the 2MCLE and 2MSCLE estimators become very close. In the limit they asymptote to a value above zero — it looks like this asymptote is close to being realised by the time $L = 100$.

6.3 Performance of asymptotic standard errors

The Monte Carlo study was extended to assess the accuracy of the asymptotic based covariance estimator in Section 4.3. Data was simulated according to a cDCC model using the above configuration for $\alpha = .05$, $\beta = .93$ with $T = 2,000$. The 2MCL estimator and the 2MSCL estimator 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, $\bar{\sigma}^2_\alpha = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}^2_{i,\alpha}$ and the corresponding Monte Carlo's estimated parameters, $\hat{\sigma}^2_\alpha = \frac{1}{1000}\sum_{i=1}^{1000}\left(\tilde{\alpha}_i - \bar{\tilde{\alpha}}\right)^2$ with $\bar{\tilde{\alpha}} = \frac{1}{1000}\sum_{i=1}^{1000}\tilde{\alpha}_i$, for both α and β.

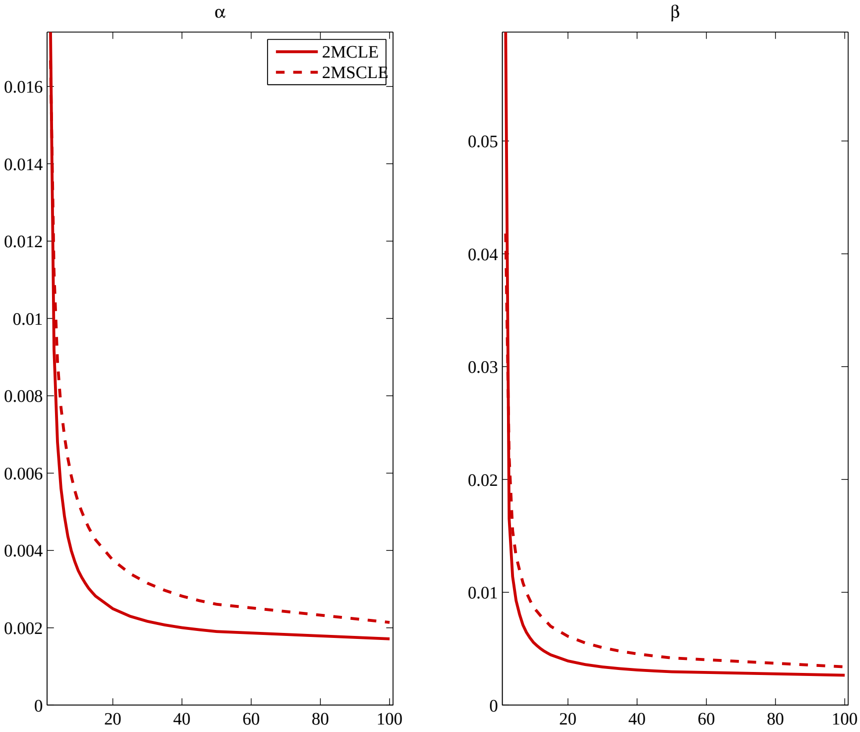

Figure 1: *Standard deviation of the CL estimators drawn against L calculated from a Monte Carlo based upon* $\alpha = .05$, $\beta = .93$ *using* $T = 2,000$. *L varies from* 2 *up to* 100. *Graphed are the results for the maximum CL estimator (2MCLE) and the subset version (2MSCLE) based on only contiguous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

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

7 Empirical comparison

7.1 Database

The data used in this empirical illustration is the same as used in Section 2.1. 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 2MCLE estimators for a variety of models. One is based on all distinct pairs, which has $N = L(L-1)/2$. The other just looks at contiguous pairs $Y_{lt} = (r_{lt}, r_{l+1,t})'$ so $N = L - 1$. The results, given in Table 6, are directly comparable with Table 1. The figures in brackets are asymptotic standard errors.

	2MCLE				2MSCLE			
L	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$	$\bar{\sigma}_\alpha$	$\hat{\sigma}_\alpha$	$\bar{\sigma}_\beta$	$\hat{\sigma}_\beta$
α=.02, β=.97								
3	.010	.008	.261	.152	.008	.007	.052	.028
10	.002	.002	.004	.004	.003	.003	.008	.007
50	.001	.001	.002	.002	.002	.002	.003	.003
100	.001	.001	.002	.001	.001	.001	.002	.002
α=.05, β=.93								
3	.009	.009	.016	.015	.011	.010	.021	.019
10	.003	.003	.006	.006	.005	.005	.009	.009
50	.002	.002	.003	.003	.003	.003	.004	.004
100	.002	.002	.003	.003	.002	.002	.003	.003
α=.10, β=.80								
3	.017	.016	.041	.040	.020	.019	.052	.049
10	.007	.006	.015	.014	.009	.010	.022	.022
50	.004	.004	.008	.008	.005	.005	.011	.011
100	.003	.003	.007	.007	.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 $\hat{\sigma}_\beta$.*

The results for the two-step CL are reasonably stable with respect to L 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 L.

Estimates from the 2MLE are markedly different from those of any of the CL based estimators, which largely agree with each other. The parameter estimates of the 2MLE and other estimators also produced meaningfully different fits.

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 $L-1$ pairs. We see the estimate is hardly effected.

To examine the fit of the models, the conditional correlations of the 95 individual stocks with the S&P 500 from the 2MCLE and 2MLE 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 2MCLE 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 2MLE produces conditional correlations which are nearly constant and exhibiting little variation even at the height of the technology bubble in 2001.

	Scalar BEKK		EWMA	DCC	
L	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$
2MCLE					
5	.0287 (.0081)	.9692 (.0092)	.0205 (.0037)	.0143 (.0487)	.9829 (.0846)
10	.0281 (.0055)	.9699 (.0063)	.0211 (.0027)	.0107 (.0012)	.9881 (.0016)
25	.0308 (.0047)	.9667 (.0055)	.0234 (.0023)	.0100 (.0009)	.9871 (.0017)
50	.0319 (.0046)	.9645 (.0056)	.0225 (.0026)	.0101 (.0008)	.9856 (.0018)
96	.0334 (.0041)	.9636 (.0049)	.0249 (.0019)	.0103 (.0009)	.9846 (.0019)
2MSCLE					
5	.0284 (.0083)	.9696 (.0094)	.0189 (.0037)	.0099 (.0033)	.9885 (.0045)
10	.0272 (.0054)	.9709 (.0062)	.0201 (.0027)	.0093 (.0016)	.9886 (.0018)
25	.0307 (.0049)	.9668 (.0056)	.0227 (.0024)	.0089 (.0011)	.9889 (.0012)
50	.0316 (.0047)	.9647 (.0057)	.0220 (.0029)	.0092 (.0010)	.9869 (.0019)
96	.0335 (.0043)	.9634 (.0051)	.0247 (.0020)	.0094 (.0009)	.9860 (.0014)

Table 6: *Based on the maximum m-profile and maximum CL estimator (2MCLE) using real and simulated data. Top part uses $L(L-1)/2$ pairs based subsets, the bottom part uses $L-1$ 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.*

7.2 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 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 $\widehat{\beta}_{l,t} = \widehat{h}_{l,t}^{1/2}\widehat{\rho}_{lm,t}/\widehat{h}_{m,t}^{1/2}$ where $h_{l,t} = \text{Var}(r_{l,t}|\mathcal{F}_{t-1})$, $\rho_{lm,t} = \text{Cor}(r_{l,t}, r_{m,t}|\mathcal{F}_{t-1})$ and $l = 1, 2, ..., L$. The corresponding hedging errors were computed as $\widehat{\nu}_{l,t} = r_{l,t} - \widehat{\beta}_{l,t}r_{m,t}$. Here $r_{l,t}$ is the return on the l-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 2MCLE to the 2MLE. 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.

Defining the difference in the squared hedging error $\widehat{\delta}_{l,t} = \left\{\widehat{\nu}_{l,t}\left(\widehat{\rho}_{l,t}^{2MCLE}\right)\right\}^2 - \left\{\widehat{\nu}_{l,t}\left(\widehat{\rho}_{l,t}^{2MLE}\right)\right\}^2$

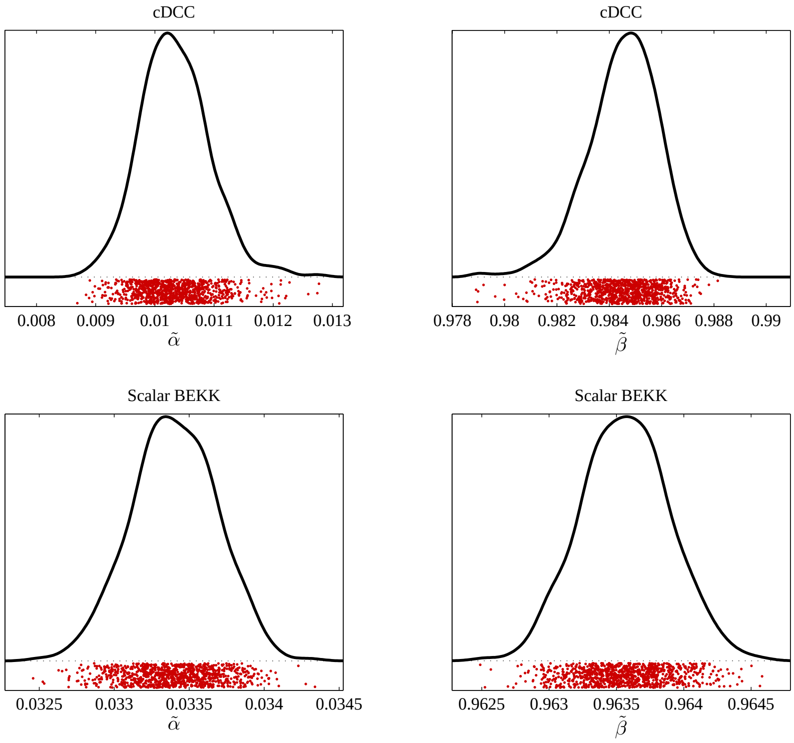

Figure 2: *Sensitivity to random selection of pairs. Density of the maximum m-profile CL estimator based on $L-1$ distinct but randomly choosen pairs. Top row are the estimators of the cDCC model and the bottom row are the corresponding estimators for the scalar BEKK.*

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 2MCLE would be the preferred model while significant positive results would indicate favor for the 2MLE. The null of $H_0 : \mathrm{E}\left(\widehat{\delta}_{l,t}\right) = 0$ was tested using a t-test, $GW = \bar{\delta}_l/\sqrt{avar\left(\sqrt{T}\bar{\delta}_l\right)}$, $\bar{\delta}_l = P^{-1}\sum_{t=R}^{P}\delta_{l,t}$. Here $\bar{\delta}_l$ is the average loss differential. Under mild regularity conditions GW is asymptotically normal. See Giacomini and White (2006) for further details (note we also tried a heteroskedastically adjusted version of the GW test, in order to increase its power, but this had no impact.).

The test statistic was computed for each asset excluding the market, resulting in 95 test statistics. Table 7 holds the results, where 37 series favour the 2MCLE estimator compared to 2 which prefer the 2MLE based estimated model. 56 are inconclusive. The results for the maximum two-step CL estimator are 24 in favour of that estimator, 8 preferring 2MLE and 63 inconclusive.

7.3 Out of sample comparison with other models

7.3.1 Scalar BEKK

We can use the CL methods to estimate the scalar BEKK model using this database. The results

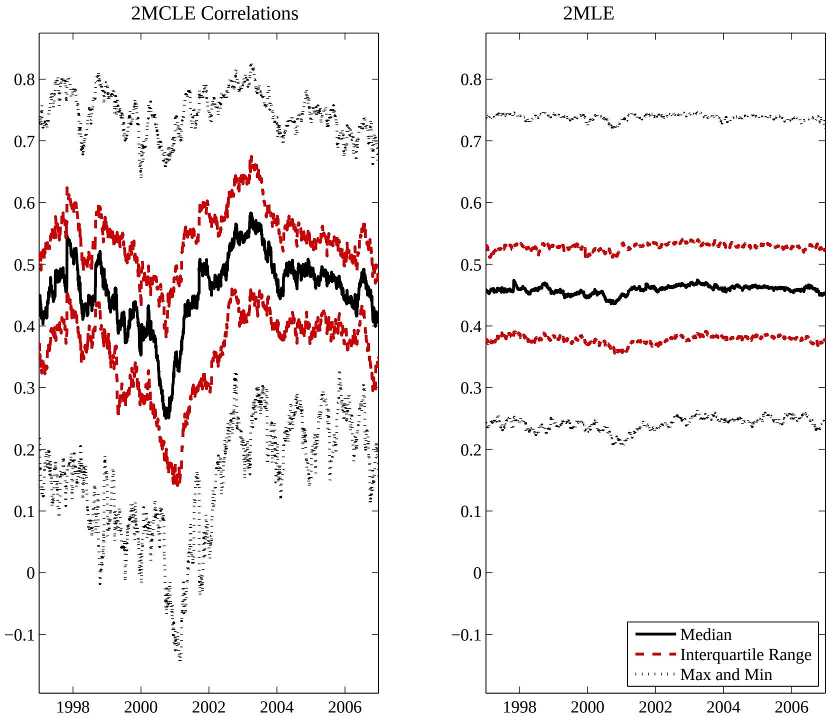

Figure 3: *How do the correlations with the market temporally change? Plot of the median, interquartile range and min & max of the correlations of the 95 included S&P 100 components with the index return using the estimates produced by the maximum CL estimator (2MCLE) & maximum m-profile likelihood estimator. Each day the 95 correlations were sorted to produce the quantiles.*

are in Table 1 and 6. They follow the same theme 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 2MLE and 2MCLE. They show the CL method delivering estimators which produce smaller hedging errors than the conventional 2MLE technique.

7.3.2 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 l. 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_l and β_l for each of the l-th submodels — it demonstrates a very significant scatter. It has 22 of the estimated $\alpha_l + \beta_l$ on their unit boundary. We will see in a moment such

M-profile

Model A	Favours A	No Decision	Favours B	Model B
DCC 2MCLE	24	63	8	DCC 2MLE
DCC 2MCLE	92	3	0	DECO
DCC 2MCLE	18	68	9	Bivariate DCC
DCC 2MCLE	9	82	4	EWMA
BEKK 2MCLE	29	65	1	BEKK 2MLE
BEKK 2MCLE	50	44	1	Bivariate BEKK

Table 7: *Which model and estimation strategy leads to smallest hedging errors? GW t-statistics for 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. 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.*

unit root models, which are often called EWMA models, perform very poorly indeed in terms of hedging. Once in a while the estimates of $\alpha_l + \beta_l$ 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 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.

7.3.3 Equicorrelation model

The Engle and Kelly (2012) 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 equicorrelation $R_t = \rho_t \iota\iota' + (1-\rho_t) I$, with $\rho_t = \omega + \gamma u_{t-1} + \beta \rho_{t-1}$, 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.

7.3.4 RiskMetrics

The 2MCLE fit of the cDCC model can be compared to the RiskMetrics method in Example 2.2 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.

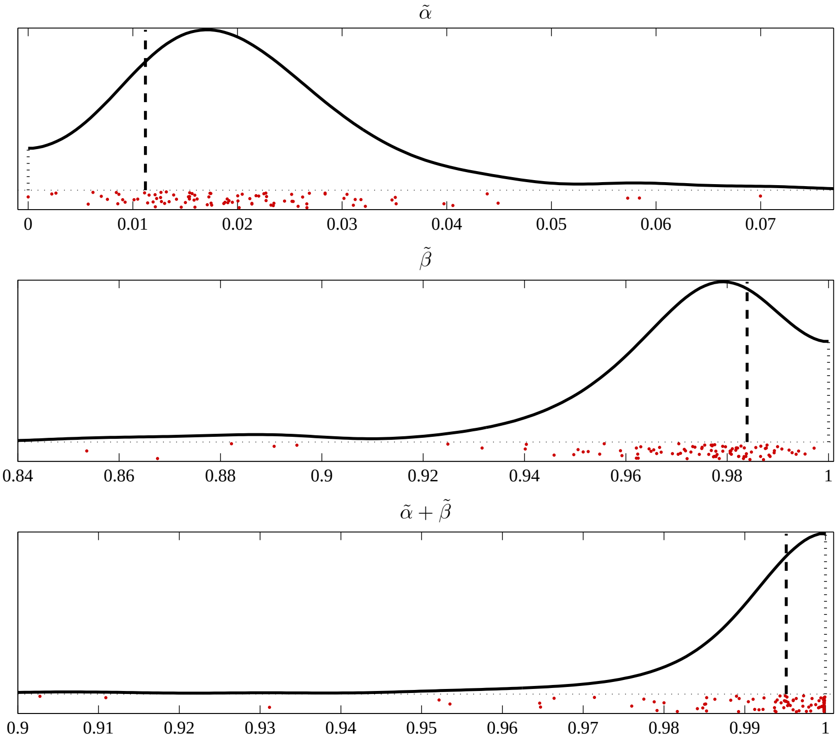

Figure 4: *Should the data be pooled across pairs? Seperately estimated α_j and β_j for each bivariate submodel for the beta-pair of the market and an individual asset. Dotted line is the CL estimator — which acts as a pooling device.*

7.4 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 2MLE, 2MCLE and 2MSCLE (contiguous pairs). The model was estimated across $L = \{5, 25, 50, 100, 250, 480\}$. Results are presented in Table 8.

The 2MLE shows 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 CL estimates are very similar, all with $\alpha \approx .03$, $\beta \approx .96$, and the standard errors decline quickly and then modestly as L increases. For large L the difference between the contiguous and all pairs estimators is small.

In the analysis of the cDCC model, for this wider set of data the best performing volatility model was the GJR-GARCH(1,1) $h_{l,t} = \omega_l + \delta_l r^2_{l,t-1} + \gamma_l r^2_{l,t-1} I_{[r_{l,t-1}<0]} + \kappa_l h_{l,t-1}$ for each margin.

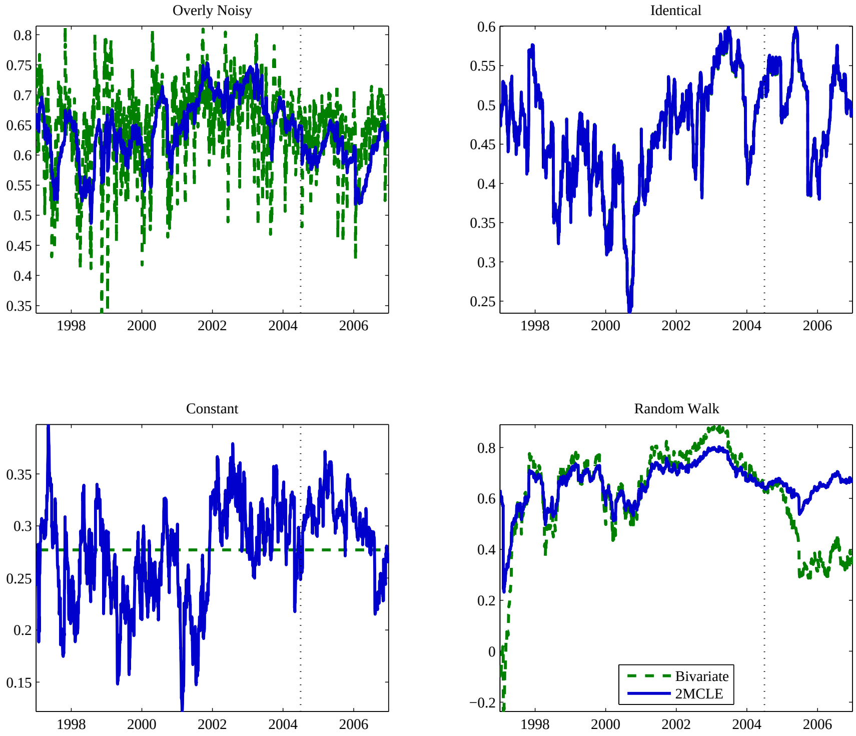

Figure 5: *Comparison of estimated conditional correlations for j-th model, including out of sampling projections, using the high dimensional model and the bivariate model. Top left looks like the bivariate model is overly noisy. Top right give results which are basically the same. Bottom left gives 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 results for the cDCC model are presented in Table 8. The 2MLE of α for the cDCC model exhibits a strong bias as the sample size increases and for $L > 250$ the β estimate is also badly affected. This contrast with the estimates from the maximum composite and maximum m-profile composite likelihood where $\alpha \approx .008$ and $\alpha + \beta \approx .995$ (The maximized CL 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).

Table 2 contains the times for each of the methods for estimating the scalar BEKK model — the simpler of the two models. The 2MLE method takes around 3.5 days on the $L = 480$ problem, while for $L = 25$ the time is quite modest being under a minute. This shows the impact of the $O(L^3)$ computational load. The composite methods are much more rapid than 2MLE, with the all pairs method still being quite fast for $L = 100$ and being around 200 times faster than 2MLE in that case. The contiguous pair method is fast even when $L = 480$, just taking a small handful of seconds. This means it is around $68,000$ times faster than 2MLE in this vast dimensional case.

	Scalar BEKK						DCC					
	2MLE		2MCLE		2MSCLE		2MLE		2MCLE		2MSCLE	
L	α	β	α	β	α	β	α	β	α	β	α	β
5	.0261	.9715	.0369 (.0057)	.9603 (.0065)	.0312 (.0053)	.9664 (.0061)	.0101	.9823	.0133 (.0041)	.9794 (.0081)	.0070 (.0033)	.9912 (.0038)
25	.0080	.9909	.0300 (.0062)	.9670 (.0075)	.0289 (.0055)	.9682 (.0067)	.0030	.9908	.0083 (.0015)	.9885 (.0031)	.0071 (.0011)	.9911 (.0016)
50	.0055	.9932	.0282 (.0051)	.9692 (.0062)	.0277 (.0049)	.9698 (.0059)	.0018	.9882	.0078 (.0010)	.9887 (.0021)	.0073 (.0010)	.9901 (.0019)
100	.0034	.9934	.0296 (.0046)	.9670 (.0057)	.0292 (.0045)	.9674 (.0056)	.0015	.9524	.0073 (.0007)	.9881 (.0015)	.0076 (.0010)	.9866 (.0028)
250	.0015	.9842	.0322 (.0049)	.9633 (.0064)	.0322 (.0048)	.9633 (.0063)	.0020	.5561	.0076 (.0007)	.9872 (.0015)	.0080 (.0016)	.9858 (.0039)
480	.0032	.5630	.0290 (.0041)	.9672 (.0054)	.0290 (.0040)	.9672 (.0053)	.0013	.2556	.0073 (.0007)	.9874 (.0016)	.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. L is the dimension of problem fitted.*

8 Additional remarks

8.1 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 $Y_{1t} = (r_{1t}, r_{2t})'$, $Y_{2t} = (r_{1t}, r_{3t})'$,..., $Y_{(L-1)t} = (r_{1t}, r_{Lt})'$, where $N = L - 1$ 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 L randomly chosen pairs, as the corresponding submodel quasi-likelihoods $\log L_{lt}(\theta, \lambda_l)$ will be tightly dependent across l. 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.

8.2 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 $O(L^2)$.

8.3 Engle's method

Engle (2009b) 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 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 $L(L-1)/2$ maximisations, (iii) no properties of this estimator were derived, (iv) the resulting estimator may not be in the permissible parameter space. Engle's method has some similarities, but is distinct, to the Ledoit, Santa-Clara, and Wolf (2003) procedure which also fits models to many pairs of observations. It is distinctively focused 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 method would average the submodels maximum quasi-likelihood estimators, which asymptotically behave like

$$\frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_j = \frac{1}{N}\sum_{j=1}^{N}\theta_j - \frac{1}{NT}\sum_{j=1}^{N}\left(\ell_{jT}^{\theta\theta} - \ell_{jT}^{\theta\lambda}(\ell_{jT}^{\lambda\lambda})^{-1}\ell_{jT}^{\lambda\theta}\right)^{-1}\sum_{t=1}^{T}\left(\ell_{jt}^{\theta} - \ell_{jT}^{\theta\lambda}(\ell_{jT}^{\lambda\lambda})^{-1}\ell_{jt}^{\lambda}\right).$$

Hence its asymptotic variance can be estimated by applying a HAC estimator to

$$\sum_{j=1}^{N}\left(\ell_{jT}^{\theta\theta} - \ell_{jT}^{\theta\lambda}(\ell_{jT}^{\lambda\lambda})^{-1}\ell_{jT}^{\lambda\theta}\right)^{-1}\left(\ell_{jt}^{\theta} - \ell_{jT}^{\theta\lambda}(\ell_{jT}^{\lambda\lambda})^{-1}\ell_{jt}^{\lambda}\right).$$

In the linear case the estimator is dominated by the submodel estimators with largest variances — i.e. components which are least informative.

8.4 Imposing factor structure on Σ

In some stationary multivariate models it might make sense to impose a factor structure on Σ, particularly when L is very large (e.g. 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 $\Sigma = ff' + \Omega$, where f is a $L \times M$ matrix of factor loadings and Ω is an L by L diagonal matrix containing the residual variances. This implies the long run the covariances in the model obey a factor structure but in the short run there can be departures from it. This can be carried out using a two step procedure: estimating the constrained Σ and then plugging this into a composite likelihood to estimate α and β.

Taking this model to the data, we estimate the factor model using the Jöreskog (1967) method which assumes the returns, factors and innovations are i.i.d. Gaussian. This implies the estimated 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 first step 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.

	$M = 1$		$M = 2$		$M = 3$	
L	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 M dimensional factor based estimate of the intercept and a composite likelihood function for α and β. L denotes the number of assets analysed.*

9 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 dimensions of many hundreds, indeed the dimension could be larger than the time series dimension.

References

Aielli, G. P. (2013). Consistent estimation of large scale dynamic conditional correlations. *Journal of Business and Economic Statistics 31*, 282–299.

Arellano, M. and S. Bonhomme (2009). Robust priors in nonlinear panel data models. *Econometrica 77*, 489–536.

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: some recent developments. In R. Blundell, W. Newey, and T. Persson (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress - Volume III*, pp. 381–409. CUP.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*, 343–65.

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. *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Bester, C. A. and C. B. Hansen (2009). A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *Journal of Business and Economic Statistics 27*, 131–148.

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.

Bosq, D., F. Merlevède, and M. Peligrad (1999). Asymptotic normality for density kernel estimators in discrete and continuous time. *Journal of Multivariate Analysis 68*, 78–95.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys 2*, 107–144.

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.

Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics 140*, 503–528.

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

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B 49*, 1–39.

Cox, D. R. and N. Reid (2004). 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.

Dhaene, G. and K. Jochmans (2011). An adjusted profile likelihood for non-stationary panel data models with fixed effects. working paper.

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.

Doukhan, P. (1994). *Mixing, Properties and Examples*. Springer.

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. (2009a). *Anticipating Correlations*. Princeton University Press.

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

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

Engle, R. F. and B. Kelly (2012). Dynamic equicorrelation. *Journal of Business and Economic Statistics 30*, 212–228.

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 9*, 36–40.

Engle, R. F., V. K. Ng, and M. Rothschild (1990). Asset pricing with a factor ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics 45*, 213–238.

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

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

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics 150*, 71–85.

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

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

Gonçalves, S. (2011). The moving blocks bootstrap for panel linear regression models with individual fixed effects. *Econometric Theory 27*, 1048–1082.

Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory 27*, 1152–1191.

Hahn, J. and W. K. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*(4), 1295–1319.

Hansen, B. E. (1991). Strong laws for dependent heterogenous processes. *Econometric Theory 7*, 213–221.

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

McCullagh, P. (1987). *Tensor Methods in Statistics.* London: Chapman & Hall.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological) 52*, 325–344.

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*, 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.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* Singapore: World Scientific.

Pakel, C. (2014). Bias reduction in nonlinear and dynamic panels in the presence of cross-section dependence. Working paper.

Rangel, J. G. and R. F. Engle (2012). The factor-spline-GARCH model for high and low-frequency correlations. *Journal of Business and Economic Statistics 30*, 109–124.

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

Severini, T. A. and W. H. Wong (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics 20*, 1768–1802.

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

Utev, S. (1991). On the central limit theorem for φ-mixing arrays of random variables. *Theory of Probability & Its Applications 35*, 131–139.

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

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

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

Xu, X. and N. Reid (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference 141*, 3047–3054.

A Scalar BEKK simulation

Here we report the results from repeating the experiments discussed in Section 6 but on the scalar BEKK model given in Example 2.1. 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 6.

B Mathematical Appendix

B.1 Moment conditions

The moment conditions used in the rest of the proofs are provided below.

	Bias						RMSE					
	2MLE		2MCLE		2MSCLE		2MLE		2MCLE		2MSCLE	
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 are the subset CL (2MSCLE), the CL (2MCLE), and the likelihood (2MLE) estimator. All results based on $2,500$ replications.*

Assumption B.1 *For some $\varepsilon > 0$ we have*

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{d^4\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\lambda_{p_1}d\lambda_{p_2}d\lambda_{p_3}}\right|^{1+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^4\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\lambda_{p_1}d\lambda_{p_2}}\right|^{1+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^3\ell_{jt}(\theta,\lambda)}{d\lambda_{p_1}d\lambda_{p_2}d\lambda_{p_3}}\right|^{1+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^4\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\theta_{r_3}d\theta_{r_4}}\right|^{1+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^4\ell_{jt}(\theta,\lambda)}{d\lambda_{p_1}d\lambda_{p_2}d\lambda_{p_3}d\lambda_{p_4}}\right|^{2+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^5\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\lambda_{p_1}d\lambda_{p_2}d\lambda_{p_3}d\lambda_{p_4}}\right|^{2+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^5\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\lambda_{p_1}d\lambda_{p_2}d\lambda_{p_3}}\right|^{2+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^5\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\theta_{r_3}d\theta_{r_4}d\theta_{r_5}}\right|^{2+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^4\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\theta_{r_3}d\lambda_{p_1}}\right|^{2+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^5\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\theta_{r_3}d\theta_{r_4}d\lambda_{p_1}}\right|^{2+\varepsilon}\right] < \infty,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[|\ell_{jt}(\theta,\lambda)|^{2+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d\ell_{jt}(\theta,\lambda)}{d\lambda_{p_1}}\right|^{3+\varepsilon}\right] < \infty, \quad \mathbb{E}\left[\left|\frac{d\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}}\right|^{3+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^2\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\lambda_{p_1}}\right|^{3+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^3\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\lambda_{p_1}}\right|^{3+\varepsilon}\right] < \infty, \quad \mathbb{E}\left[\left|\frac{d^2\ell_{jt}(\theta,\lambda)}{d\lambda_{p_1}d\lambda_{p_2}}\right|^{3+\varepsilon}\right] < \infty,\\
\mathbb{E}\left[\left|\frac{d^2\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}}\right|^{3+\varepsilon}\right] &< \infty, \quad \mathbb{E}\left[\left|\frac{d^3\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\lambda_{p_1}d\lambda_{p_2}}\right|^{3+\varepsilon}\right] < \infty, \quad \mathbb{E}\left[\left|\frac{d^3\ell_{jt}(\theta,\lambda)}{d\theta_{r_1}d\theta_{r_2}d\theta_{r_3}}\right|^{3+\varepsilon}\right] < \infty,
\end{aligned}$$

uniformly for all θ, λ, j, t and $r_1, ..., r_5 \in \{1, ..., R\}$ and $p_1, ..., p_4 \in \{1, ..., P\}$.

B.2 Consistency

The analysis in this section is to a large extent based on Lemmas 1 and 4, and Theorems 3 and 4 of Hahn and Kuersteiner (2011), and their proofs. In what follows, let $\hat{Q}_{(j)}(\theta,\lambda)=\ell_{jT}(\theta,\lambda)$ and $Q_{(j)}(\theta,\lambda)=E[\ell_{jT}(\theta,\lambda)]$. Throughout this section we maintain the Assumptions of Theorem 4.1. As mentioned before, the particular moment and summability conditions required in this section are (i) $\mathbb{E}[|\ell_{jt}(\theta,\lambda)|^{2+\varepsilon}]<\infty$ and (ii) $\sum_{m=1}^{\infty}\alpha_j(m)^{1/q-1/r}<\infty$ for some $r>q>1$.

We start with two preliminary results, which will be followed by the proof of consistency.

Lemma B.1 *Let $\{Y_{jt}\}_{t=1}^{T}$ be a zero-mean α-mixing sequence where, for some r and q such that $r>q>1$, $\sum_{m=1}^{\infty}\alpha_j(m)^{1/q-1/r}<\infty$ for all $j=1,...,N$. Moreover, $\sup_{j,t}\mathbb{E}[||Y_{jt}||^{2+\delta}]<\infty$ for some $\delta>0$. Then, $P[||\frac{1}{T}\sum_{t=1}^{T}Y_{jt}||\geq\eta]=o(T^{-1})$ for every $\eta>0$ as $T\to\infty$.*

Proof of Lemma B.1. We start with

$$P\left[\left|\left|\frac{1}{T}\sum_{t=1}^{T}Y_{jt}\right|\right|\geq\eta\right]=P\left[\left|\left|\sum_{t=1}^{T}Y_{jt}\right|\right|^{2+\delta/2}\geq\eta^{2+\delta/2}T^{2+\delta/2}\right]\leq\frac{1}{(T\eta)^{2+\delta/2}}\mathbb{E}\left[\left|\left|\sum_{t=1}^{T}Y_{jt}\right|\right|^{2+\delta/2}\right],$$

which follows from Markov's inequality. Then, by Corollary 3 of Hansen (1991), for some $K<\infty$

$$\left\{\mathbb{E}\left[\left(\max_{\bar{T}\leq T}\left|\left|\sum_{t=1}^{\bar{T}}Y_{jt}\right|\right|\right)^{q}\right]\right\}^{1/q}\leq K\left(\sum_{t=1}^{T}\{\mathbb{E}\left[||Y_{jt}||^{r}\right]\}^{2/r}\right)^{1/2}.$$

Choosing $q=2+\delta/2$ and $r=2+\delta$, this yields, after some rearranging,[7]

$$\mathbb{E}\left[\left(\left|\left|\sum_{t=1}^{T}Y_{jt}\right|\right|\right)^{2+\delta/2}\right]\leq K^{2+\delta/2}\left\{\sup_{j,t}\mathbb{E}\left[||Y_{jt}||^{2+\delta}\right]\right\}^{\frac{2+\delta/2}{2+\delta}}T^{1+\delta/4}=O(T^{1+\delta/4}).$$

Hence,

$$P\left[\left|\left|\frac{1}{T}\sum_{t=1}^{T}Y_{jt}\right|\right|\geq\eta\right]\leq\frac{1}{(T\eta)^{2+\delta/2}}O(T^{1+\delta/4})=O(T^{-1-\delta/4}),$$

which is $o(T^{-1})$, as desired. ■

Lemma B.2 *For all $\eta>0$,*

$$P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}|\ell_{jT}(\theta,\lambda)-\mathbb{E}[\ell_{jT}(\theta,\lambda)]|>\eta\right]=o(1).$$

Proof of Lemma B.2. For some $\eta>0$,

$$P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right]\leq\sum_{j=1}^{N}P\left[\sup_{(\theta,\lambda)\in\Psi}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right].$$

Let $\mathcal{S}_{\delta}(\psi_k)=\{(\theta,\lambda):||(\theta,\lambda)-(\theta_k,\lambda_k)||<\delta\}$ and $\mathcal{S}_{\delta}^{k}=\mathcal{S}_{\delta}(\psi_k)$. Since Θ and Λ are compact,

[7] Note that, $\mathbb{E}\left[\left(\left|\left|\sum_{t=1}^{T}Y_{jt}\right|\right|\right)^{q}\right]\leq\mathbb{E}\left[\left(\max_{\bar{T}\leq T}\left|\left|\sum_{t=1}^{\bar{T}}Y_{jt}\right|\right|\right)^{q}\right]$, since $\left|\left|\sum_{t=1}^{T}Y_{jt}\right|\right|\leq\max_{\bar{T}\leq T}\left|\left|\sum_{t=1}^{\bar{T}}Y_{jt}\right|\right|$.

there exists a finite collection of subsets $\mathcal{S}_\delta^1, ..., \mathcal{S}_\delta^{K(\delta)}$ such that $\Psi \subset \cup_{k=1}^{K(\delta)} \mathcal{S}_\delta^k$. Hence,

$$P\left[\sup_{(\theta,\lambda)\in\Psi}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right]\leq\sum_{k=1}^{K(\delta)}P\left[\sup_{(\theta,\lambda)\in\mathcal{S}_\delta^k}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right].$$

Now, consider some particular value of k. Then, for some $(\theta,\lambda)\in\mathcal{S}_\delta^k$,

$$\begin{aligned}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| &\leq \left|\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}(\theta,\lambda)-\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}(\theta_k,\lambda_k)\right|+\left|\ell_{jT}(\theta_k,\lambda_k)-Q_{(j)}(\theta_k,\lambda_k)\right|\\ &\quad+\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\ell_{jt}(\theta_k,\lambda_k)\right]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\ell_{jt}(\theta,\lambda)\right]\right|,\end{aligned}$$

which follows from the Triangle Inequality. Then,

$$\begin{aligned}\sup_{(\theta,\lambda)\in\mathcal{S}_\delta^k}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| &\leq \left|\ell_{jT}(\theta_k,\lambda_k)-Q_{(j)}(\theta_k,\lambda_k)\right|+\frac{\delta}{T}\sum_{t=1}^{T}c(Y_{jt})+\frac{\delta}{T}\sum_{t=1}^{T}\mathbb{E}[c(Y_{jt})]\\ &\leq \left|\ell_{jT}(\theta_k,\lambda_k)-Q_{(j)}(\theta_k,\lambda_k)\right|+\frac{\delta}{T}\left|\sum_{t=1}^{T}c(Y_{jt})-\sum_{t=1}^{T}\mathbb{E}[c(Y_{jt})]\right|\\ &\quad+2\frac{\delta}{T}\sum_{t=1}^{T}\mathbb{E}[c(Y_{jt})],\end{aligned}$$

where the first inequality follows from Assumption 4.2(i) and the fact that for $(\theta,\lambda)\in\mathcal{S}_\delta^k$, $||(\theta,\lambda)-(\theta_k,\lambda_k)||<\delta$. One can pick δ such that $2\delta\max_{j,t}\mathbb{E}[c(Y_{jt})]<\eta/3$. Then,

$$\begin{aligned}P\left[\sup_{(\theta,\lambda)\in\mathcal{S}_\delta^k}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|>\eta\right] &\leq P\left[\left|\ell_{jT}(\theta_k,\lambda_k)-Q_{(j)}(\theta_k,\lambda_k)\right|>\frac{\eta}{3}\right]\\ &\quad+P\left[\frac{1}{T}\left|\sum_{t=1}^{T}c(Y_{jt})-\sum_{t=1}^{T}\mathbb{E}[c(Y_{jt})]\right|>\frac{\eta}{3\delta}\right]\\ &\quad+P\left[2\frac{\delta}{T}\sum_{t=1}^{T}\mathbb{E}[c(Y_{jt})]>\frac{\eta}{3}\right],\end{aligned}$$

where the final probability on the right hand side is equal to zero for the particular choice of δ here. Hence, by Lemma B.1, $P\left[\sup_{(\theta,\lambda)\in\mathcal{S}_\delta^k}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right]=o(T^{-1})$ and, finally,

$$P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|\geq\eta\right]=o(1),$$

since $N=O(T)$. The mixing conditions of Assumption 4.4(ii) are stronger than necessary for Lemma B.1. However, they will be needed when proving Theorems 4.2 and 4.3. ■

The main proof follows next.

Proof of Theorem 4.1. Let $\eta>0$. By Lemma B.2, for any $\tau>0$

$$P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}\left|\ell_{jT}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right|>\tau\right]=o(1).$$

Let $\tau = \varepsilon/2$, where $\varepsilon = \inf_{1\le j\le N}\left[Q_{(j)}(\theta_0,\lambda_{j0}) - \sup_{\{(\theta,\lambda):||(\theta,\lambda)-(\theta_0,\lambda_{j0})||>\eta\}} Q_{(j)}(\theta,\lambda)\right] > 0$ by Assumption 4.3(i). Then, with probability $1-o(1)$,

$$\begin{aligned}
&\max_{\{(\theta,\lambda_1,\ldots,\lambda_N):||\theta-\theta_0||\ge\eta;\lambda_1,\ldots,\lambda_N\}} \frac{1}{N}\sum_{j=1}^{N}\ell_{jT}(\theta,\lambda_j)\\
\le\ & \max_{\{(\theta,\lambda_1,\ldots,\lambda_N):||(\theta,\lambda_j)-(\theta_0,\lambda_{j0})||\ge\eta\ \forall j\}} \frac{1}{N}\sum_{j=1}^{N}\ell_{jT}(\theta,\lambda_j) \le \frac{1}{N}\sum_{j=1}^{N}\max_{\{(\theta,\lambda_j):||(\theta,\lambda_j)-(\theta_0,\lambda_{j0})||\ge\eta\}}\ell_{jT}(\theta,\lambda_j)\\
<\ & \frac{1}{N}\sum_{j=1}^{N}\max_{\{(\theta,\lambda_j):||(\theta,\lambda_j)-(\theta_0,\lambda_{j0})||\ge\eta\}} Q_{(j)}(\theta,\lambda_j) + \frac{\varepsilon}{2} < \frac{1}{N}\sum_{j=1}^{N}Q_{(j)}(\theta_0,\lambda_{j0}) - \frac{\varepsilon}{2} < \frac{1}{N}\sum_{j=1}^{N}\ell_{jT}(\theta_0,\lambda_{j0})\\
\le\ & \max_{(\theta,\lambda)\in\Psi}\frac{1}{N}\sum_{j=1}^{N}\ell_{jT}(\theta,\lambda_j),
\end{aligned}$$

The 3rd and 5th inequalities follow from Lemma B.2, while the 4th is due to Assumption 4.3(i). Hence, with probability $1-o(1)$, $||\hat\theta-\theta_0||<\eta$, giving the desired result. Proving consistency of the nuisance parameter estimator follows using the same ideas. Let $\eta>0$. Again, by Lemma B.2, $P[\max_{1\le j\le N}\sup_{(\theta,\lambda)\in\Psi}|\hat Q_{(j)}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)|>\tau]=o(1)$, for some $\tau>0$. We choose

$$2\tau=\varepsilon=\inf_{\theta\in\Theta}\inf_{1\le j\le N}\left[Q_{(j)}(\theta,\bar\lambda_{jT}(\theta)) - \sup_{\{\lambda:\left|\left|\lambda-\bar\lambda_{jT}(\theta)\right|\right|>\eta\}} Q_{(j)}(\theta,\lambda)\right]>0,$$

where $\varepsilon>0$ by Assumption 4.3(ii). Now, with probability $1-o(1)$,

$$\begin{aligned}
\max_{1\le j\le N}\max_{\{\lambda:\left|\left|\lambda-\bar\lambda_{jT}(\theta)\right|\right|>\eta\}}\hat Q_{(j)}(\theta,\lambda) &< \max_{1\le j\le N}\max_{\{\lambda:\left|\left|\lambda-\bar\lambda_{jT}(\theta)\right|\right|>\eta\}} Q_{(j)}(\theta,\lambda)+\varepsilon/2\\
&< \max_{1\le j\le N}Q_{(j)}(\theta,\bar\lambda_{jT}(\theta))-\varepsilon/2 < \max_{1\le j\le N}\hat Q_{(j)}(\theta,\bar\lambda_{jT}(\theta))\\
&\le \max_{1\le j\le N}\max_{\lambda\in\Lambda}\hat Q_{(j)}(\theta,\lambda).
\end{aligned}$$

Hence, for each j and any η, $P[\max_{1\le j\le N}||\hat\lambda_j(\theta)-\bar\lambda_{jT}(\theta)||<\eta]=1-o(1)$. ■

We finish this section by providing one more result, which will be used below in finding the orders of the remainder terms in mean value expansions.

Theorem B.1 *Let $\tilde\theta\in\Theta$ be such that $\tilde\theta\xrightarrow{p}\theta_0$ as $N,T\to\infty$. Then,*

$$P\left[\max_{1\le j\le N}\left|\left|\hat\lambda_j(\tilde\theta)-\bar\lambda_{jT}(\theta_0)\right|\right|<\eta\right]=1-o(1),\qquad \textit{for all } \eta>0.$$

Proof of Theorem B.1. For conciseness, define $\mathcal{E}=\max_{1\le j\le N}\sup_{\lambda\in\Lambda}|\hat Q_{(j)}(\tilde\theta,\lambda)-Q_{(j)}(\theta_0,\lambda)|$. Then,

$$\begin{aligned}
\mathcal{E} &\le \max_{1\le j\le N}\sup_{\lambda\in\Lambda}\left|\hat Q_{(j)}(\tilde\theta,\lambda)-Q_{(j)}(\tilde\theta,\lambda)\right| + \max_{1\le j\le N}\sup_{\lambda\in\Lambda}\left|Q_{(j)}(\tilde\theta,\lambda)-Q_{(j)}(\theta_0,\lambda)\right|\\
&\le \max_{1\le j\le N}\sup_{(\theta,\lambda)\in\Psi}\left|\hat Q_{(j)}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| + \max_{1\le j\le N}\sup_{\lambda\in\Lambda}\mathbb{E}[c(Y_{jt})]\left|\left|(\tilde\theta,\lambda)-(\theta_0,\lambda)\right|\right|\\
&= \max_{1\le j\le N}\sup_{(\theta,\lambda)\in\Psi}\left|\hat Q_{(j)}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| + \max_{1\le j\le N}\mathbb{E}[c(Y_{jt})]\left|\left|\tilde\theta-\theta_0\right|\right|.
\end{aligned}$$

Therefore, for some $\tau > 0$,

$$\begin{aligned} P\left[\mathcal{E} \geq \tau\right] &\leq P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}\left|\hat{Q}_{(j)}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| \geq \tau/2\right] + P\left[\max_{1\leq j\leq N}\mathbb{E}[c(Y_{jt})]\left|\left|\tilde{\theta}-\theta_0\right|\right| \geq \tau/2\right] \\ &= P\left[\max_{1\leq j\leq N}\sup_{(\theta,\lambda)\in\Psi}\left|\hat{Q}_{(j)}(\theta,\lambda)-Q_{(j)}(\theta,\lambda)\right| \geq \tau/2\right] + P\left[\left|\left|\tilde{\theta}-\theta_0\right|\right| \geq \frac{\tau}{2\max_{1\leq j\leq N}\mathbb{E}[c(Y_{jt})]}\right], \end{aligned}$$

which, by the assumption that $||\tilde{\theta}-\theta_0|| \xrightarrow{p} 0$ and by Lemma B.2, implies that $P[\mathcal{E} < \tau] = 1 - o(1)$. Now, let τ be such that $2\tau = \varepsilon = \inf_{1\leq j\leq N}\left[Q_{(j)}(\theta_0,\bar{\lambda}_{jT}(\theta_0)) - \sup_{\{\lambda:||\lambda-\bar{\lambda}_{jT}(\theta_0)||>\eta\}} Q_{(j)}(\theta_0,\lambda)\right]$, which is positive by Assumption 4.3(ii). Now, for any $\hat{\Lambda} \subseteq \Lambda$, $\left|\sup_{\lambda\in\hat{\Lambda}}\hat{Q}_{(j)}(\tilde{\theta},\lambda) - \sup_{\lambda\in\hat{\Lambda}} Q_{(j)}(\theta_0,\lambda)\right| \leq \sup_{\lambda\in\hat{\Lambda}}\left|\hat{Q}_{(j)}(\tilde{\theta},\lambda) - Q_{(j)}(\theta_0,\lambda)\right|$. Then, since $P[\mathcal{E} < \tau] = 1 - o(1)$, this implies that

$$P\left[\max_{1\leq i\leq j}\left|\sup_{\lambda\in\hat{\Lambda}}\hat{Q}_{(j)}(\tilde{\theta},\lambda) - \sup_{\lambda\in\hat{\Lambda}} Q_{(j)}(\theta_0,\lambda)\right| < \varepsilon/2\right] = 1 - o(1), \tag{9}$$

as well. Therefore, with probability $1 - o(1)$, we have for all j

$$\begin{aligned} \sup_{\{\lambda:||\lambda-\bar{\lambda}_{jT}(\theta_0)||>\eta\}} \hat{Q}_{(j)}(\tilde{\theta},\lambda) &< \sup_{\{\lambda:||\lambda-\bar{\lambda}_{jT}(\theta_0)||>\eta\}} Q_{(j)}(\theta_0,\lambda) + \varepsilon/2 \\ &\leq Q_{(j)}(\theta_0,\bar{\lambda}_{jT}(\theta_0)) - \varepsilon/2 < \hat{Q}_{(j)}(\tilde{\theta},\hat{\lambda}_j(\tilde{\theta})), \end{aligned}$$

where the first and last inequalities follow from (9) while the second inequality is due to Assumption 4.3(ii). Since, by definition, $\hat{\lambda}_j(\tilde{\theta}) = \arg\max_{\lambda\in\Lambda}\hat{Q}_{(j)}(\tilde{\theta},\lambda)$, it must be the case that $||\hat{\lambda}_j(\tilde{\theta}) - \bar{\lambda}_{jT}(\theta_0)|| \leq \eta$, for all j with probability $1 - o(1)$, which proves the theorem. ■

B.3 Proofs of theorems 4.2 and 4.3

Proofs in this part are inspired by the strong dependence case of Pakel (2014), who considers the *integrated composite likelihood method*. Our main contribution is extending these results to multivariate incidental parameters for the standard *composite likelihood method*.

B.3.1 A short overview of the index notation

Due to θ and λ being vector-valued parameters, we have to use multivariate asymptotic expansions in the remainder. Unfortunately, these become quite tedious, even with low order expansions. To simplify the algebra, we will use *index notation* and the *Einstein summation convention*. Below, we provide a short overview; for a more detailed treatment of these notational techniques see McCullagh (1987) and Pace and Salvan (1997).

Here to make the likelihood notation more concise, we use the following short hand notation:

$$\begin{gathered} \bar{\ell}_{jT} = \ell_{jT}(\theta,\bar{\lambda}_{jT}(\theta)), \quad \bar{\ell}_{NT} = \ell_{NT}(\theta,\bar{\lambda}_{1,T}(\theta),...,\bar{\lambda}_{N,T}(\theta)), \quad \hat{\ell}_{NT} = \ell_{NT}(\theta,\hat{\lambda}_1(\theta),...,\hat{\lambda}_N(\theta)), \\ \hat{\ell}_{jT} = \ell_{jT}(\theta,\hat{\lambda}_j(\theta)), \quad \ell_{jT} = \ell_{jT}(\theta_0,\bar{\lambda}_{jT}(\theta_0)), \quad \ell_{NT} = \ell_{NT}(\theta_0,\bar{\lambda}_{1,T}(\theta_0),...,\bar{\lambda}_{N,T}(\theta_0)). \end{gathered}$$

In index notation an array of any dimension can be written as a scalar where the array structure is made explicit by the use of indices. For example, let $[x_d]$ denote a D-dimensional vector where $d = 1,...,D$. Using index notation, this can be written more concisely as x_d. Similarly, a $D_1 \times D_2$ matrix $[x_{d_1,d_2}]$, where $d_1 = 1,...,D_1$ is the row index and $d_2 = 1,...,D_2$ is the column index, would be written as x_{d_1,d_2}. We adopt this convention for all likelihood derivatives. For example,

$$\ell_{j;p} = \frac{d\ell_{jT}}{d\lambda_{j;p}}, \quad \hat{\ell}_{j;r} = \frac{d\hat{\ell}_{jT}}{d\theta_r}, \quad \ell_{j;p_1,p_2,r_1} = \frac{d^3\ell_{jT}}{d\lambda_{j;p_1} d\lambda_{j;p_2} d\theta_{r_1}} \ ,$$
$$\bar{\ell}_p = \frac{d\bar{\ell}_{NT}}{d\lambda_{j;p}}, \quad \ell_r = \frac{d\ell_{NT}}{d\theta_r}, \quad \ell_{p_1,p_2,r_1} = \frac{d^3\ell_{NT}}{d\lambda_{j;p_1} d\lambda_{j;p_2} d\theta_{r_1}} \quad \text{etc.}$$

where $p_i \in \{1, ..., P\}$ and $r_i \in \{1, ..., R\}$. Note that, we use the indices q and r for derivatives with respect to θ only, while $a, b, c, d, e, f, l, m, n, o$ and p are reserved for derivatives with respect to λ. Hence, for example, $[\ell_p]$ is the P-dimensional score vector $d\ell_{NT}/d\lambda_j$ while $[\ell_{j;r_i,r_j}]$ is the Hessian matrix $d^2\ell_{jT}/d\theta d\theta'$ where $r_1, r_2 = \{1, ..., R\}$ etc. In the following, we will use the index notation both to denote a particular entry and the whole array itself, e.g. depending on context $\ell_{j;r_i,r_j}$ can stand for $d^2\ell_{jT}/d\theta d\theta'$ or the row r_i and column r_j entry of this matrix. We also use the following notation for expected values of likelihood terms and their centred versions:

$$\bar{v}_{j;p_1,p_2} = \mathbb{E}[\bar{\ell}_{j;p_1,p_2}], \quad v_{p_1,p_2} = \mathbb{E}[\ell_{p_1,p_2}], \quad \mathcal{H}_{j;a,b} = \ell_{j;a,b} - v_{j;a,b} \quad \bar{\mathcal{H}}_{q,m} = \bar{\ell}_{q,m} - \bar{v}_{q,m} \quad \text{etc.}$$

To denote the inverse of a matrix, we use upperscript indices: for example, $\bar{v}_j^{p_1,p_2}$ denotes row p_1 column p_2 entry of the matrix $\left(\frac{d^2\ell_{jT}(\theta,\bar{\lambda}_j(\theta))}{d\lambda_j d\lambda_j'}\right)^{-1}$. The Kronecker Delta is given by κ_p^q where $\kappa_p^q = 1$ if $p = q$ and $\kappa_p^q = 0$ if $p \neq q$. In what follows, we will use $\epsilon_{j;p}(\theta) = \hat{\lambda}_{j;p}(\theta) - \bar{\lambda}_{jT;p}(\theta)$ and $\delta_r = \hat{\theta}_r - \theta_{0;r}$, where $\lambda_{j;p}$ is entry p of λ_j and θ_r is entry r of θ.

Another technique employed here is the Einstein summation convention, which is used to represent multiple summations concisely. The essence of this notation is that, whenever an index appears twice in a given expression, that expression is to be summed across that index. For example, consider the following term written in the Einstein summation notation: $x_{r,q}x_{q,p}x^{r,p}y_z$. The indices p, q and r appear twice while z appears only once. Therefore, in the standard notation this term is equal to $\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} x_{r,q}x_{q,p}x^{r,p}y_z$. The number of indices that appear once (the so called free indices) then determine the dimension of the expression. For example, the previous expression is a vector while $x_{r,q}y_{q,z}$ is a (2×2) matrix since r and z appear only once. Note that one can freely change the letters used for the indices, as long as the relationship between the indices remains the same. For example $x_{a,b}x^{b,c}$ is the same as $x_{d,f}x^{f,e}$ whereas $x_{b,d}x^{b,a}$ is a different expression.

Remark 2 *Notice that, since by definition* $\frac{\partial}{\partial\lambda_j}\sum_{t=1}^{T}\mathbb{E}[\ell_{jt}(\theta,\lambda_j)]\Big|_{\lambda_j=\bar{\lambda}_{jT}(\theta)} = 0$ *for all* θ*, we have*

$$\nabla_\theta \left\{\frac{\partial}{\partial\lambda_j}\sum_{t=1}^{T}\mathbb{E}[\ell_{jt}(\theta,\lambda_j)]\right\}\Big|_{\lambda_j=\bar{\lambda}_{jT}(\theta)} = 0, \quad \nabla_{\theta\theta} \left\{\frac{\partial}{\partial\lambda_j}\sum_{t=1}^{T}\mathbb{E}[\ell_{jt}(\theta,\lambda_j)]\right\}\Big|_{\lambda_j=\bar{\lambda}_{jT}(\theta)} = 0,$$

etc. $\forall\theta$.

B.3.2 Some preliminary lemmas

First recall a result from Pakel (2014, Lemma A.3).

Lemma B.3 *For any given* j*, let* Y_{jt}*,* $t = 1, ..., T$ *be an* α*-mixing random sequence such that* $\lim_{m\to\infty}\alpha_{ij,k}(m) = 0$*,* $\lim_{m\to\infty}\alpha_{i,jk}(m) = 0$*, and for some* $\delta > 0$ $\sum_{m=1}^{\infty} m\alpha_{ij,k}(m)^{\delta/3+\delta} < \infty$ *and* $\sum_{m=1}^{\infty} m\alpha_{i,jk}(m)^{\delta/3+\delta} < \infty$*, uniformly for all* $i, j, k = 1, ..., N$*. Let* $f(\cdot)$*,* $g(\cdot)$ *and* $h(\cdot)$ *be some measurable functions of* Y_{jt} *where* $\mathbb{E}[f(\cdot)] = \mathbb{E}[g(\cdot)] = \mathbb{E}[h(\cdot)] = 0$*. In addition assume* *that* $\mathbb{E}[|f(Y_{jt})|^{3+\delta}] < \infty$, $\mathbb{E}[|g(Y_{jt})|^{3+\delta}] < \infty$ *and* $\mathbb{E}[|h(Y_{jt})|^{3+\delta}] < \infty$ *for all* j,t. *Then,*

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T} f(Y_{is})g(Y_{jt})\right] &= O\left(\frac{1}{T}\right), & (10)\\
\mathbb{E}\left[\frac{1}{T^3}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{q=1}^{T} f(Y_{is})g(Y_{jt})h(Y_{kq})\right] &= O\left(\frac{1}{T^2}\right), & (11)\\
\mathbb{E}\left[\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T} f(Y_{is})g(Y_{jt})\right] &= O\left(\frac{1}{T}\right), & (12)\\
\mathbb{E}\left[\frac{1}{N^3T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{q=1}^{T} f(Y_{is})g(Y_{jt})h(Y_{kq})\right] &= O\left(\frac{1}{T^2}\right), & (13)
\end{aligned}$$

where (10) and (11) hold for all $i, j, k = 1, ..., N$.

Lemma B.4 *Let* $\tilde{\lambda}_j(\theta)$ *be the mean value between* $\hat{\lambda}_j(\theta)$ *and* $\bar{\lambda}_{jT}(\theta)$ *and* $\tilde{\theta}$ *be the mean value between* $\hat{\theta}$ *and* θ_0. *Note that, in what follows,* $\tilde{\lambda}_j(\theta)$ *and* $\tilde{\theta}$ *do not necessarily take on the same value for all the terms considered. Then, using the index notation as defined before,*

$$\begin{aligned}
\ell_{j;p_1,p_2,p_3,p_4}(\theta, \tilde{\lambda}_j(\theta))\epsilon_{j;p_2}(\theta)\epsilon_{j;p_3}(\theta)\epsilon_{j;p_4}(\theta) &= O_p(T^{-3/2}), & (14)\\
\ell_{j;r_1,r_2,m,n,l}(\theta_0, \tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0) &= O_p(T^{-3/2}), & (15)\\
\ell_{j;r_1,m,n,l,o}(\theta_0, \tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0)\epsilon_{j;o}(\theta_0) &= O_p(T^{-2}), & (16)\\
\ell_{r_1,r_2,r_3,r_4,r_5}(\tilde{\theta}, \hat{\lambda}_j(\tilde{\theta}))\delta_{r_2}\delta_{r_3}\delta_{r_4}\delta_{r_5} &= O_p(T^{-2}), & (17)\\
\ell_{j;r_1,r_2,r_3,m}(\theta_0, \tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0) &= o_p(T^{-1/2}), & (18)\\
\ell_{j;r_1,r_2,r_3,r_4,m}(\theta_0, \tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0) &= o_p(T^{-1/2}). & (19)
\end{aligned}$$

Proof of Lemma B.4. Consider (14). By the same arguments as in the proof of Lemma B.2, a uniform convergence result for $\ell_{j;p_1,p_2,p_3,p_4}(\theta, \lambda)$ (and any of the other likelihood derivatives given in Lemma B.4) can be obtained. Then, by uniform convergence of $\ell_{j;p_1,p_2,p_3,p_4}(\theta, \lambda)$ and since $\tilde{\lambda}_j(\theta) \xrightarrow{p} \bar{\lambda}_{jT}(\theta)$, using standard arguments one obtains

$$\ell_{j;p_1,p_2,p_3,p_4}(\theta, \tilde{\lambda}_j(\theta)) = \mathbb{E}[\ell_{j;p_1,p_2,p_3,p_4}(\theta, \bar{\lambda}_{jT}(\theta))] + o_p(1).$$

Since, $\hat{\lambda}_j(\theta) - \bar{\lambda}_{jT}(\theta) = O_p(T^{-1/2})$ for all j and θ, we have, $\ell_{j;p_1,p_2,p_3,p_4}(\theta, \tilde{\lambda}_j(\theta))\epsilon_j^{p_2}(\theta)\epsilon_j^{p_3}(\theta)\epsilon_j^{p_4}(\theta) = O_p(T^{-3/2})$, as desired. By using exactly the same line of arguments as above, (15) and (16) can be shown to be $O_p(T^{-3/2})$ and $O_p(T^{-2})$, respectively.

Next, take (17). We know that $\tilde{\theta} - \theta_0 \xrightarrow{p} 0$. In addition, by Theorem B.1, $\hat{\lambda}_j(\tilde{\theta}) - \bar{\lambda}_{jT}(\theta_0) \xrightarrow{p} 0$ for all j. Hence, $(\tilde{\theta}, \hat{\lambda}_j(\tilde{\theta})) - (\theta_0, \bar{\lambda}_{jT}(\theta_0)) \xrightarrow{p} 0$. By the above arguments,

$$\ell_{r_1,r_2,r_3,r_4,r_5}(\tilde{\theta}, \hat{\lambda}_j(\tilde{\theta})) = \mathbb{E}[\ell_{r_1,r_2,r_3,r_4,r_5}(\theta_0, \bar{\lambda}_{jT}(\theta_0))] + o_p(1),$$

which, together with $||\hat{\theta}-\theta_0|| = O_p(T^{-1/2})$, implies that $\ell_{r_1,r_2,r_3,r_4,r_5}(\tilde{\theta}, \hat{\lambda}_j(\tilde{\theta}))\delta_{r_2}\delta_{r_3}\delta_{r_4}\delta_{r_5} = O_p(T^{-2})$.

Two terms are left. By definition both $\mathbb{E}[\ell_{j;r_1,r_2,r_3,m}(\theta_0, \bar{\lambda}_{jT}(\theta_0))]$ and $\mathbb{E}[\ell_{j;r_1,r_2,r_3,r_4,m}(\theta_0, \bar{\lambda}_{jT}(\theta_0))]$ are equal to zero (see Remark 2). Then, by using similar arguments as before,

$$\begin{aligned}
\ell_{j;r_1,r_2,r_3,m}(\theta_0, \tilde{\lambda}_j(\theta_0)) &= \mathbb{E}[\ell_{j;r_1,r_2,r_3,m}(\theta_0, \bar{\lambda}_{jT}(\theta_0))] + o_p(1) = o_p(1),\\
\ell_{j;r_1,r_2,r_3,r_4,m}(\theta_0, \tilde{\lambda}_j(\theta_0)) &= \mathbb{E}[\ell_{j;r_1,r_2,r_3,r_4,m}(\theta_0, \bar{\lambda}_{jT}(\theta_0))] + o_p(1) = o_p(1).
\end{aligned}$$

Since $\hat{\lambda}_j(\theta) - \bar{\lambda}_{jT}(\theta) = O_p(T^{-1/2})$ for all j and θ, we have the desired results. ■

B.3.3 First step: multivariate expansion for $\hat{\lambda}_j(\theta) - \bar{\lambda}_{jT}(\theta)$

We start with an expansion of the score with respect to λ_j around $\hat{\lambda}_j(\theta) = \bar{\lambda}_{jT}(\theta)$:

$$\begin{aligned}
\hat{\ell}_{j;p_1} &= \bar{\ell}_{j;p_1} + \bar{\ell}_{j;p_1,p_2}\epsilon_{j;p_2}(\theta) + \frac{1}{2}\bar{\ell}_{j;p_1,p_2,p_3}\epsilon_{j;p_2}(\theta)\epsilon_{j;p_3}(\theta) \\
&\quad + \frac{1}{6}\bar{\ell}_{j;p_1,p_2,p_3,p_4}(\theta, \tilde{\lambda}_j(\theta))\epsilon_{j;p_2}(\theta)\epsilon_{j;p_3}(\theta)\epsilon_{j;p_4}(\theta), \\
&= \bar{\ell}_{j;p_1} + \bar{v}_{j;p_1,p_2}\epsilon_{j;p_2}(\theta) + \bar{\mathcal{H}}_{j;p_1,p_2}\epsilon_{j;p_2}(\theta) + \frac{1}{2}\bar{v}_{j;p_1,p_2,p_3}\epsilon_{j;p_2}(\theta)\epsilon_{j;p_3}(\theta) + O_p\left(T^{-3/2}\right)
\end{aligned}$$

where the remainder term is $O_p(T^{-3/2})$ by Lemma B.4. Since $\hat{\ell}_{j;p_1} = 0$, we have

$$\bar{v}_{j;p_1,p_2}\epsilon_{j;p_2}(\theta) = -\left(\bar{\ell}_{j;p_1} + \bar{\mathcal{H}}_{j;p_1,p_2}\epsilon_{j;p_2}(\theta) + \frac{1}{2}\bar{v}_{j;p_1,p_2,p_3}\epsilon_{j;p_2}(\theta)\epsilon_{j;p_3}(\theta)\right) + O_p\left(T^{-3/2}\right). \quad (20)$$

The next step is to invert (20) in order to obtain an expression for $\epsilon_{j;m}(\theta) = \hat{\lambda}_{j;m}(\theta) - \bar{\lambda}_{jT;m}(\theta)$ in terms of likelihood terms only. First, notice that, $\bar{v}_{j;p_1,p_2}\bar{v}_j^{p_1,m}\epsilon_{j;p_2}(\theta) = \kappa^m_{p_2}\epsilon_{j;p_2}(\theta) = \epsilon_{j;m}(\theta)$; see e.g. Pace and Salvan (1997). Then, multiplying both sides of (20) by $\bar{v}_j^{p_1,m}$ yields

$$\begin{aligned}
\epsilon_{j;m}(\theta) &= -\bar{\ell}_{j;p_1}\bar{v}_j^{p_1,m} - [\bar{\mathcal{H}}_{j;p_1,p_2}\bar{v}_j^{p_1,m} + \frac{1}{2}\bar{v}_{j;p_1,p_2,p_3}\bar{v}_j^{p_1,m}\epsilon_{j;p_3}(\theta)]\epsilon_{j;p_2}(\theta) + O_p(T^{-3/2}), \quad (21) \\
\epsilon_{j;p_2}(\theta) &= -\left(\bar{\ell}_{j;a}\bar{v}_j^{a,p_2} + \bar{\mathcal{H}}_{j;a,b}\bar{v}_j^{a,p_2}\epsilon_{j;b}(\theta) + \frac{1}{2}\bar{v}_{j;a,b,c}\bar{v}_j^{a,p_2}\epsilon_{j;b}(\theta)\epsilon_{j;c}(\theta)\right) + O_p\left(T^{-3/2}\right), \quad (22) \\
\epsilon_{j;p_3}(\theta) &= -\left(\bar{\ell}_{j;d}\bar{v}_j^{d,p_3} + \bar{\mathcal{H}}_{j;d,e}\bar{v}_j^{d,p_3}\epsilon_{j;e}(\theta) + \frac{1}{2}\bar{v}_{j;d,e,f}\bar{v}_j^{d,p_3}\epsilon_{j;e}(\theta)\epsilon_{j;f}(\theta)\right) + O_p\left(T^{-3/2}\right). (23)
\end{aligned}$$

Note that (22) and (23) are copies of (21), although with a different set of indices to prevent confusion. Substituting (22) and (23) into (21) finally yields an expansion for $\hat{\lambda}_j(\theta) - \bar{\lambda}_{jT}(\theta)$:

$$\epsilon_{j;m}(\theta) = -\left(\bar{\ell}_{j;p_1}\bar{v}_j^{p_1,m} + \bar{\mathcal{H}}_{j;p_1,p_2}\bar{v}_j^{p_1,m}\bar{\ell}_{j;a}\bar{v}_j^{a,p_2} + \frac{1}{2}\bar{v}_{j;p_1,p_2,p_3}\bar{v}_j^{p_1,m}\bar{\ell}_{j;a}\bar{v}_j^{a,p_2}\bar{\ell}_{j;d}\bar{v}_j^{d,p_3}\right) + O_p(T^{-3/2}).$$

B.3.4 Second step: multivariate expansion for $\hat{\theta} - \theta_0$

Now, by a mean value expansion of $\ell_{r_1}(\hat{\theta}, \hat{\lambda}_j(\hat{\theta}))$ around $\hat{\theta} = \theta_0$, we have

$$\begin{aligned}
\ell_{r_1}(\hat{\theta}, \hat{\lambda}_j(\hat{\theta})) &= \ell_{r_1}(\theta_0, \hat{\lambda}_j(\theta_0)) + \ell_{r_1,r_2}(\theta_0, \hat{\lambda}_j(\theta_0))\delta_{r_2} + \frac{1}{2}\ell_{r_1,r_2,r_3}(\theta_0, \hat{\lambda}_j(\theta_0))\delta_{r_2}\delta_{r_3} \\
&\quad + \frac{1}{6}\ell_{r_1,r_2,r_3,r_4}(\theta_0, \hat{\lambda}_j(\theta_0))\delta_{r_2}\delta_{r_3}\delta_{r_4} + \frac{1}{24}\ell_{r_1,r_2,r_3,r_4,r_5}(\tilde{\theta}, \hat{\lambda}_j(\tilde{\theta}))\delta_{r_2}\delta_{r_3}\delta_{r_4}\delta_{r_5}. (24)
\end{aligned}$$

By Lemma B.4, the remainder term is $O_p(T^{-2})$. Unfortunately, the asymptotic behaviour of the likelihood derivatives evaluated at $(\theta_0, \hat{\lambda}_j(\theta_0))$ is not clear. However, by expanding these terms around $(\theta_0, \bar{\lambda}_{jT}(\theta_0)) = (\theta_0, \lambda_{j0})$ we can obtain an asymptotically equivalent expression, which can be analysed conveniently. Doing this for each term on the right hand side of (24) yields

$$\begin{aligned}
\ell_{j;r_1}(\theta_0, \hat{\lambda}_j(\theta_0)) &= \ell_{j;r_1} + \ell_{j;r_1,m}\epsilon_{j;m}(\theta_0) + \frac{1}{2}\ell_{j;r_1,m,n}\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0) + \frac{1}{6}\ell_{j;r_1,m,n,l}\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0) \\
&\quad + \frac{1}{24}\ell_{j;r_1,m,n,l,o}(\theta_0, \tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0)\epsilon_{j;o}(\theta_0) \\
&= \ell_{j;r_1} + \ell_{j;r_1,m}\epsilon_{j;m}(\theta_0) + \frac{1}{2}\left(v_{j;r_1,m,n} + \mathcal{H}_{j;r_1,m,n}\right)\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0) \\
&\quad + \frac{1}{6}v_{j;r_1,m,n,l}\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0) + O_p(T^{-2})
\end{aligned}$$

$$\begin{aligned} &= \ell_{j;r_1} - \ell_{j;r_1,m}\left(\ell_{j;p_1}v_j^{p_1,m} + \mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} + \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right) \\ &\quad +\frac{1}{2}v_{j;r_1,m,n}\left[\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}\right. \\ &\quad +\ell_{j;p_1}v_j^{p_1,m}\left(\mathcal{H}_{j;p_1',p_2'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'} + \frac{1}{2}v_{j;p_1',p_2',p_3'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}\ell_{j;d'}v_j^{d',p_3'}\right) \\ &\quad \left.+\ell_{j;p_1'}v_j^{p_1',n}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} + \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right)\right] \\ &\quad +\frac{1}{2}\mathcal{H}_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n} \\ &\quad -\frac{1}{6}v_{j;r_1,m,n,l}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}\ell_{j;p_1''}v_j^{p_1'',l} + O_p(T^{-2}), \end{aligned}$$

while

$$\begin{aligned} \ell_{j;r_1,r_2}(\theta_0,\hat{\lambda}_j(\theta_0)) &= v_{j;r_1,r_2} + \mathcal{H}_{j;r_1,r_2} + \ell_{j;r_1,r_2,m}\epsilon_{j;m}(\theta_0) + \frac{1}{2}\ell_{j;r_1,r_2,m,n}\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0) \\ &\quad +\frac{1}{6}\ell_{j;r_1,r_2,m,n,l}(\theta_0,\tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0)\epsilon_{j;n}(\theta_0)\epsilon_{j;l}(\theta_0), \\ &= v_{j;r_1,r_2} + \mathcal{H}_{j;r_1,r_2} - \ell_{j;r_1,r_2,m}(\ell_{j;p_1}v_j^{p_1,m}) + \frac{1}{2}v_{j;r_1,r_2,m,n}[\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}] \\ &\quad +O_p(T^{-3/2}), \\ \ell_{j;r_1,r_2,r_3}(\theta_0,\hat{\lambda}_j(\theta_0)) &= v_{j;r_1,r_2,r_3} + \mathcal{H}_{j;r_1,r_2,r_3} + \ell_{j;r_1,r_2,r_3,m}(\theta_0,\tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0) \\ &= v_{j;r_1,r_2,r_3} + \mathcal{H}_{j;r_1,r_2,r_3} + O_p(T^{-1}), \\ \ell_{j;r_1,r_2,r_3,r_4}(\theta_0,\hat{\lambda}_j(\theta_0)) &= v_{j;r_1,r_2,r_3,r_4} + \mathcal{H}_{j;r_1,r_2,r_3,r_4} + \ell_{j;r_1,r_2,r_3,r_4,m}(\theta_0,\tilde{\lambda}_j(\theta_0))\epsilon_{j;m}(\theta_0) \\ &= v_{j;r_1,r_2,r_3,r_4} + O_p(T^{-1/2}). \end{aligned}$$

The orders of the remainder terms in these four mean value expansions are given by Lemma B.4 (see (15), (16), (18) and (19)). In the above calculations we have used

$$\begin{aligned} \epsilon_{j;m}(\theta_0) &= -\ell_{j;p_1}v_j^{p_1,m} - \mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} - \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3} + O_p(T^{-3/2}), \\ \epsilon_{j;n}(\theta_0) &= -\ell_{j;p_1'}v_j^{p_1',n} - \mathcal{H}_{j;p_1',p_2'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'} - \frac{1}{2}v_{j;p_1',p_2',p_3'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}\ell_{j;d'}v_j^{d',p_3'} + O_p(T^{-3/2}), \\ \epsilon_{j;l}(\theta_0) &= -\ell_{j;p_1''}v_j^{p_1'',l} - \mathcal{H}_{j;p_1'',p_2''}v_j^{p_1'',l}\ell_{j;a''}v_j^{a'',p_2''} - \frac{1}{2}v_{j;p_1'',p_2'',p_3''}v_j^{p_1'',l}\ell_{j;a''}v_j^{a'',p_2''}\ell_{j;d''}v_j^{d'',p_3''} + O_p(T^{-3/2}). \end{aligned}$$

These expressions are identical, except that in each case different indices have been used, in order to keep track of different entries of the arrays appearing in the expansion. Now, substituting the above derived asymptotically equivalent expressions into (24) and summing across i gives

$$\begin{aligned} \ell_{r_1}(\hat{\theta},\hat{\lambda}_j(\hat{\theta})) &= \ell_{r_1} + \delta_{r_2}v_{r_1,r_2} + \delta_{r_2}\mathcal{H}_{r_1,r_2} + \frac{1}{2}\delta_{r_2}\delta_{r_3}v_{r_1,r_2,r_3} + \frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n} \\ &\quad -\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,m}\ell_{j;p_1}v_j^{p_1,m} + \frac{1}{6}\delta_{r_2}\delta_{r_3}\delta_{r_4}v_{r_1,r_2,r_3,r_4} + \frac{1}{2}\delta_{r_2}\delta_{r_3}\mathcal{H}_{r_1,r_2,r_3} \\ &\quad -\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,m}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} + \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right) \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2N}\sum_{j=1}^{N} v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\left(\mathcal{H}_{j;p_1',p_2'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}+\frac{1}{2}v_{j;p_1',p_2',p_3'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}\ell_{j;d'}v_j^{d',p_3'}\right)\\
&+\frac{1}{2N}\sum_{j=1}^{N} v_{j;r_1,m,n}\ell_{j;p_1'}v_j^{p_1',n}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}+\frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right)\\
&+\frac{1}{2N}\sum_{j=1}^{N}\mathcal{H}_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}-\frac{1}{6N}\sum_{j=1}^{N}v_{j;r_1,m,n,l}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}\ell_{j;p_1''}v_j^{p_1'',l}\\
&-\delta_{r_2}\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,r_2,m}\ell_{j;p_1}v_j^{p_1,m}+\delta_{r_2}\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,r_2,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}+O_p(T^{-2}).
\end{aligned}
$$

The next step is to formally invert this expression in order to obtain an asymptotic expansion for δ in terms of likelihood terms only. Observe that (i) $\ell_{r_1}(\hat{\theta},\hat{\lambda}_j(\hat{\theta}))=0$ for all possible values of r_1, (ii) $\ell_{r_1,r_2}=v_{r_1,r_2}+\mathcal{H}_{r_1,r_2}$ and (iii) by the definition of Kronecker's Delta, $v_{r_1,r_2}v^{r_1,q}\delta_{r_2}=\kappa^q_{r_2}\delta_{r_2}=\delta_q$. Then, multiplying both sides by $v^{r_1,q}$ and rearranging gives

$$
\begin{aligned}
\delta_q \quad = \quad & -\ell_{r_1}v^{r_1,q}+\frac{1}{N}\sum_{j=1}^{N}v^{r_1,q}\ell_{j;r_1,m}\ell_{j;p_1}v_j^{p_1,m}-\delta_{r_2}\mathcal{H}_{r_1,r_2}v^{r_1,q}-\frac{1}{2}\delta_{r_2}\delta_{r_3}v_{r_1,r_2,r_3}v^{r_1,q}\\
&-\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}\\
&+\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,m}v^{r_1,q}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}+\frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right)\\
&-\frac{1}{6}\delta_{r_2}\delta_{r_3}\delta_{r_4}v_{r_1,r_2,r_3,r_4}v^{r_1,q}-\frac{1}{2}\delta_{r_2}\delta_{r_3}v^{r_1,q}\mathcal{H}_{r_1,r_2,r_3}\\
&-\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}v^{r_1,q}\left(\mathcal{H}_{j;p_1',p_2'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}+\frac{1}{2}v_{j;p_1',p_2',p_3'}v_j^{p_1',n}\ell_{j;a'}v_j^{a',p_2'}\ell_{j;d'}v_j^{d',p_3'}\right)\\
&-\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,m,n}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}+\frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right)\\
&-\frac{1}{2N}\sum_{j=1}^{N}\mathcal{H}_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}+\frac{1}{6N}\sum_{j=1}^{N}v_{j;r_1,m,n,l}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}\ell_{j;p_1''}v_j^{p_1'',l}v^{r_1,q}\\
&+\delta_{r_2}\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,r_2,m}\ell_{j;p_1}v_j^{p_1,m}v^{r_1,q}-\delta_{r_2}\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,r_2,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}+O_p(T^{-2})
\end{aligned}
\tag{25}
$$

For convenience, we list below the copies of δ_q which we will use in the remainder of the inversion process. These are

$$
\begin{aligned}
\delta_{r_2} \quad = \quad & -\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}+\frac{1}{N}\sum_{j=1}^{N}v^{\bar{r}_1,r_2}\ell_{j;\bar{r}_1,\bar{m}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}}-\delta_{\bar{r}_2}\mathcal{H}_{\bar{r}_1,\bar{r}_2}v^{\bar{r}_1,r_2}-\frac{1}{2}\delta_{\bar{r}_2}\delta_{\bar{r}_3}v_{\bar{r}_1,\bar{r}_2,\bar{r}_3}v^{\bar{r}_1,r_2}\\
&-\frac{1}{2N}\sum_{j=1}^{N}v_{j;\bar{r}_1,\bar{m},\bar{n}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}}\ell_{j;\bar{p}_1'}v_j^{\bar{p}_1',\bar{n}}v^{\bar{r}_1,r_2}+O_p(T^{-3/2}),\\
\delta_{r_3} \quad = \quad & -\ell_{\tilde{r}_1}v^{\tilde{r}_1,r_3}+\frac{1}{N}\sum_{j=1}^{N}v^{\tilde{r}_1,r_3}\ell_{j;\tilde{r}_1,\tilde{m}}\ell_{j;\tilde{p}_1}v_j^{\tilde{p}_1,\tilde{m}}-\delta_{\tilde{r}_2}\mathcal{H}_{\tilde{r}_1,\tilde{r}_2}v^{\tilde{r}_1,r_3}-\frac{1}{2}\delta_{\tilde{r}_2}\delta_{\tilde{r}_3}v_{\tilde{r}_1,\tilde{r}_2,\tilde{r}_3}v^{\tilde{r}_1,r_3}
\end{aligned}
$$

$$\begin{aligned}
& -\frac{1}{2N}\sum_{j=1}^{N} v_{j;\tilde{r}_1,\tilde{m},\tilde{n}}\ell_{j;\tilde{p}_1}v_j^{\tilde{p}_1,\tilde{m}}\ell_{j;\tilde{p}'_1}v_j^{\tilde{p}'_1,\tilde{n}}v^{\tilde{r}_1,r_3} + O_p(T^{-3/2}), \\
\delta_{r_4} \quad = \quad & -\ell_{\dot{r}_1}v^{\dot{r}_1,r_4} + \frac{1}{N}\sum_{j=1}^{N} v^{\dot{r}_1,r_4}\ell_{j;\dot{r}_1,\dot{m}}\ell_{j;\dot{p}_1}v_j^{\dot{p}_1,\dot{m}} - \delta_{\dot{r}_2}\mathcal{H}_{\dot{r}_1,\dot{r}_2}v^{\dot{r}_1,r_4} - \frac{1}{2}\delta_{\dot{r}_2}\delta_{\dot{r}_3}v_{\dot{r}_1,\dot{r}_2,\dot{r}_3}v^{\dot{r}_1,r_4} \\
& -\frac{1}{2N}\sum_{j=1}^{N} v_{j;\dot{r}_1,\dot{m},\dot{n}}\ell_{j;\dot{p}_1}v_j^{\dot{p}_1,\dot{m}}\ell_{j;\dot{p}'_1}v_j^{\dot{p}'_1,\dot{n}}v^{\dot{r}_1,r_4} + O_p(T^{-3/2}), \\
\delta_{\bar{r}_2} \quad = \quad & -\ell_r v^{r,\bar{r}_2} + O_p(T^{-1}), \quad \delta_{\bar{r}_3} = -\ell_r v^{r,\bar{r}_3} + O_p(T^{-1}), \quad \delta_{\tilde{r}_2} = -\ell_r v^{r,\tilde{r}_2} + O_p(T^{-1}), \\
\delta_{\tilde{r}_3} \quad = \quad & -\ell_r v^{r,\tilde{r}_3} + O_p(T^{-1}), \quad \delta_{\dot{r}_2} = -\ell_r v^{r,\dot{r}_2} + O_p(T^{-1}), \quad \delta_{\dot{r}_3} = -\ell_r v^{r,\dot{r}_3} + O_p(T^{-1}).
\end{aligned}$$

Substituting these terms repeatedly into (25) until the right hand side of this equation is free of any δ terms and rearranging according to order gives

$$\begin{aligned}
\delta_q \quad = \quad & -\ell_{r_1}v^{r_1,q} + \frac{1}{N}\sum_{j=1}^{N} v^{r_1,q}\ell_{j;r_1,m}\ell_{j;p_1}v_j^{p_1,m} + \mathcal{H}_{r_1,r_2}v^{r_1,q}\ell_{\bar{r}_1}v^{\bar{r}_1,r_2} \\
& -\frac{1}{2}\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}\ell_{\tilde{r}_1}v^{\tilde{r}_1,r_3}v_{r_1,r_2,r_3}v^{r_1,q} - \frac{1}{2N}\sum_{j=1}^{N} v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p'_1}v_j^{p'_1,n}v^{r_1,q} \\
& -\mathcal{H}_{r_1,r_2}v^{r_1,q}\left(\frac{1}{N}\sum_{j=1}^{N} v^{\bar{r}_1,r_2}\ell_{j;\bar{r}_1,\bar{m}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}} + \ell_r v^{r,\bar{r}_2}\mathcal{H}_{\bar{r}_1,\bar{r}_2}v^{\bar{r}_1,r_2}\right. \\
& \left. -\frac{1}{2}\ell_r v^{r,\bar{r}_2}\ell_r v^{r,\bar{r}_3}v_{\bar{r}_1,\bar{r}_2,\bar{r}_3}v^{\bar{r}_1,r_2} - \frac{1}{2N}\sum_{j=1}^{N} v_{j;\bar{r}_1,\bar{m},\bar{n}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}}\ell_{j;\bar{p}'_1}v_j^{\bar{p}'_1,\bar{n}}v^{\bar{r}_1,r_2}\right) \\
& +\frac{1}{2}\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}v_{r_1,r_2,r_3}v^{r_1,q}\left(\frac{1}{N}\sum_{j=1}^{N} v^{\tilde{r}_1,r_3}\ell_{j;\tilde{r}_1,\tilde{m}}\ell_{j;\tilde{p}_1}v_j^{\tilde{p}_1,\tilde{m}} + \ell_r v^{r,\tilde{r}_2}\mathcal{H}_{\tilde{r}_1,\tilde{r}_2}v^{\tilde{r}_1,r_3}\right. \\
& \left. -\frac{1}{2}\ell_r v^{r,\tilde{r}_2}\ell_r v^{r,\tilde{r}_3}v_{\tilde{r}_1,\tilde{r}_2,\tilde{r}_3}v^{\tilde{r}_1,r_3} - \frac{1}{2N}\sum_{j=1}^{N} v_{j;\tilde{r}_1,\tilde{m},\tilde{n}}\ell_{j;\tilde{p}_1}v_j^{\tilde{p}_1,\tilde{m}}\ell_{j;\tilde{p}'_1}v_j^{\tilde{p}'_1,\tilde{n}}v^{\tilde{r}_1,r_3}\right) \\
& +\frac{1}{2}\ell_{\tilde{r}_1}v^{\tilde{r}_1,r_3}v_{r_1,r_2,r_3}v^{r_1,q}\left(\frac{1}{N}\sum_{j=1}^{N} v^{\bar{r}_1,r_2}\ell_{j;\bar{r}_1,\bar{m}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}} + \ell_r v^{r,\bar{r}_2}\mathcal{H}_{\bar{r}_1,\bar{r}_2}v^{\bar{r}_1,r_2}\right. \\
& \left. -\frac{1}{2}\ell_r v^{r,\bar{r}_2}\ell_r v^{r,\bar{r}_3}v_{\bar{r}_1,\bar{r}_2,\bar{r}_3}v^{\bar{r}_1,r_2} - \frac{1}{2N}\sum_{j=1}^{N} v_{j;\bar{r}_1,\bar{m},\bar{n}}\ell_{j;\bar{p}_1}v_j^{\bar{p}_1,\bar{m}}\ell_{j;\bar{p}'_1}v_j^{\bar{p}'_1,\bar{n}}v^{\bar{r}_1,r_2})\right) \\
& +\frac{1}{N}\sum_{j=1}^{N} \ell_{j;r_1,m}v^{r_1,q}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} + \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right) \\
& +\frac{1}{6}\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}\ell_{\tilde{r}_1}v^{\tilde{r}_1,r_3}\ell_{\dot{r}_1}v^{\dot{r}_1,r_4}v_{r_1,r_2,r_3,r_4}v^{r_1,q} - \frac{1}{2}\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}\ell_{\tilde{r}_1}v^{\tilde{r}_1,r_3}v^{r_1,q}\mathcal{H}_{r_1,r_2,r_3} \\
& -\frac{1}{2N}\sum_{j=1}^{N} v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}v^{r_1,q}\left(\mathcal{H}_{j;p'_1,p'_2}v_j^{p'_1,n}\ell_{j;a'}v_j^{a',p'_2} + \frac{1}{2}v_{j;p'_1,p'_2,p'_3}v_j^{p'_1,n}\ell_{j;a'}v_j^{a',p'_2}\ell_{j;d'}v_j^{d',p'_3}\right) \\
& -\frac{1}{2N}\sum_{j=1}^{N} v_{j;r_1,m,n}\ell_{j;p'_1}v_j^{p'_1,n}v^{r_1,q}\left(\mathcal{H}_{j;p_1,p_2}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2} + \frac{1}{2}v_{j;p_1,p_2,p_3}v_j^{p_1,m}\ell_{j;a}v_j^{a,p_2}\ell_{j;d}v_j^{d,p_3}\right)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2N}\sum_{j=1}^{N}\mathcal{H}_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}+\frac{1}{6N}\sum_{j=1}^{N}v_{j;r_1,m,n,l}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}\ell_{j;p_1''}v_j^{p_1'',l}v^{r_1,q}\\
&-\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,r_2,m}\ell_{j;p_1}v_j^{p_1,m}v^{r_1,q}+\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}\frac{1}{2N}\sum_{j=1}^{N}v_{j;r_1,r_2,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_1'}v_j^{p_1',n}v^{r_1,q}\\
&+O_p(T^{-2}). \qquad (26)
\end{aligned}
$$

Our objective is to derive an analytical expression for $\mathbb{E}[\hat{\theta}-\theta_0]$ up to a $O(T^{-2})$ remainder. Although (26) looks quite complicated, determining the orders of the terms comprising $\mathbb{E}[\delta^q]$ is straightforward using Lemma B.3. The main idea is as follows: All the terms comprising (26) are products of some expectations (which are all $O(1)$) and some zero-mean likelihood derivatives. Therefore, by using Lemma B.3, if a term contains two zero-mean likelihood derivatives, then it will be $O(T^{-1})$ in expectation. If, on the other hand, a term contains three zero-mean likelihood derivatives, then it will be $O(T^{-2})$ in expectation (although the product itself is $O_p(T^{-3/2})$). This reveals that all the terms except for the first five are $O(T^{-2})$ in expectation. Below we illustrate these points formally by considering some specific terms.

Example B.1 *Start with $\mathcal{T}_0=\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}v^{r_1,q}\mathcal{H}_{r_1,r_2}$. Here, $v^{\bar{r}_1,r_2}$ is the row $\bar{r}_1$ and column r_2 entry and $v^{r_1,q}$ is the row r_1 and column q entry of $\left\{\mathbb{E}\left[\frac{d^2}{d\theta d\theta'}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}\right]\right\}^{-1}$. The vector $\frac{d}{d\theta}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}$ is given by the array $[\ell_{\bar{r}_1}]$ whereas $\mathcal{H}_{r_1,r_2}$ is the row r_1 and column r_2 entry of $\frac{d^2}{d\theta d\theta'}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}-\mathbb{E}\left[\frac{d^2}{d\theta d\theta'}\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\ell_{jt}\right]$. Notice that $v^{\bar{r}_1,r_2}$ and $v^{r_1,q}$ are $O(1)$ while $\ell_{\bar{r}_1}$ and $\mathcal{H}_{r_1,r_2}$ are zero-mean. Let $f(Y_{is})=\ell_{is;\bar{r}_1}=\frac{d\ell(\theta_0,\bar{\lambda}_{iT}(\theta_0);Y_{is})}{d\theta_{\bar{r}_1}}$ and $g(Y_{jt})=\mathcal{H}_{jt;r_1,r_2}=\frac{d\ell(\theta_0,\bar{\lambda}_{jT}(\theta_0);Y_{jt})}{d\theta_{r_1}d\theta_{r_2}}-\mathbb{E}\left[\frac{d\ell(\theta_0,\bar{\lambda}_{jT}(\theta_0);Y_{jt})}{d\theta_{r_1}d\theta_{r_2}}\right]$. Then,*

$$
\mathbb{E}[\mathcal{T}_0]=O(1)\mathbb{E}[\ell_{\bar{r}_1}\mathcal{H}_{r_1,r_2}]=O(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}[f(Y_{is})g(Y_{jt})],
$$

which is $O(T^{-1})$ for any r_1, r_2 and $\bar{r}_1$ by Lemma B.3.

Example B.2 *Consider $\mathcal{T}_1=\ell_{\bar{r}_1}v^{\bar{r}_1,r_2}v^{r_1,q}\frac{1}{N}\sum_{j=1}^{N}\ell_{j;r_1,r_2,m}\ell_{j;p_1}v_j^{p_1,m}$. Again, we have both zero-mean likelihood derivatives and $O(1)$ expected values. For example, $v_j^{p_1,m}$ is the row p_1 column m entry of $\left\{\mathbb{E}\left[\frac{\partial^2}{\partial\lambda\partial\lambda'}\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}\right]\right\}^{-1}$ while $v^{\bar{r}_1,r_2}$ and $v^{r_1,q}$ are different entries of $\frac{1}{N}\sum_{j=1}^{N}$ $\left\{\mathbb{E}\left[\frac{d^2}{d\theta d\theta'}\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}\right]\right\}^{-1}$. All these expressions are $O(1)$. Now, consider the zero-mean terms. A typical element of the three-dimensional array given by $\ell_{j;r_1,r_2,m}$ would be $\frac{d^2}{d\theta_{r_1}d\theta_{r_2}}\frac{\partial}{\partial\lambda_m}\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}$. Similarly, $\ell_{j;p_1}$ represents $\frac{\partial}{\partial\lambda}\frac{1}{T}\sum_{t=1}^{T}\ell_{jt}$ while $\ell_{\bar{r}_1}$ is the same as in Example B.1. Now, let $f(Y_{is})=\ell_{is;\bar{r}_1}=\frac{d\ell(\theta_0,\bar{\lambda}_{iT}(\theta_0);Y_{is})}{d\theta_{\bar{r}_1}}$, $g(Y_{jt})=\ell_{jt;r_1,r_2,m}=\frac{d^3\ell(\theta_0,\bar{\lambda}_{jT}(\theta_0);Y_{jt})}{d\theta_{r_1}d\theta_{r_2}d\lambda_{j;m}}$ and $h(Y_{jw})=\ell_{jw;p_1}=\frac{d\ell(\theta_0,\bar{\lambda}_{jT}(\theta_0);Y_{jw})}{d\lambda_{j;p_1}}$. Notice that all of these functions are zero-mean and they retain the α-mixing properties of the underlying dataset. Then,*

$$
\begin{aligned}
\mathcal{T}_1 &= \left(\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}f(Y_{js})\right)v^{\bar{r}_1,r_2}v^{r_1,q}\frac{1}{NT^2}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{w=1}^{T}g\left(Y_{jt}\right)h(Y_{jw})v_j^{p_1,m}\\
&= O\left(1\right)\frac{1}{N^2T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{w=1}^{T}f(Y_{is})g\left(Y_{jt}\right)h(Y_{jw}).
\end{aligned}
$$

By (11) in Lemma B.3, $\sum_{s=1}^T \sum_{t=1}^T \sum_{w=1}^T \mathbb{E}\left[f(Y_{is}) g\left(Y_{jt}\right) f(Y_{jw})\right] = O(T)$ *for any* i *and* j*. Therefore,* $\mathbb{E}[\mathcal{T}_1] = O(T^{-2})$*.*

Example B.3 *The third example is,*

$$\mathcal{T}_2 = \frac{1}{2N} \upsilon^{r_1,q} \sum_{j=1}^N \ell_{j;p_1} \ell_{j;a'} \ell_{j;d'} \upsilon_{j;r_1,m,n} \upsilon_j^{p_1,m} \upsilon_{j;p_1',p_2',p_3'} \upsilon_j^{p_1',n} \upsilon_j^{a',p_2'} \upsilon_j^{d',p_3'}$$

The terms $\upsilon_j^{p_1',n}$, $\upsilon_j^{a',p_2'}$, $\upsilon_j^{d',p_3'}$ *and* $\upsilon_j^{p_1,m}$ *represent different entries of the matrix* $\frac{\partial^2}{\partial\lambda\partial\lambda'}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{jt}\right]$*. Further,* $\upsilon^{r_1,q}$ *which stands for the row* r_1 *and column* q *entry of* $\frac{d^2}{d\theta d\theta'}\mathbb{E}\left[\frac{1}{NT}\sum_{j=1}^N \sum_{t=1}^T \ell_{jt}\right]$*. The arrays given by* $\upsilon_{j;r_1,m,n}$ *and* $\upsilon_{j;p_1',p_2',p_3'}$ *have typical entries given by* $\frac{d}{d\theta_r}\frac{\partial}{\partial\lambda_m\partial\lambda_n}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{jt}\right]$ *and* $\frac{\partial^3}{\partial\lambda_{p_1'}\partial\lambda_{p_2'}\partial\lambda_{p_3'}}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{jt}\right]$*, respectively. These terms are all* $O(1)$*. As before,* $\ell_{j;p_1}$*,* $\ell_{j;a'}$ *and* $\ell_{j;d'}$ *are different entries of the score with respect to* λ*. Using the same ideas as above, the asymptotic order of* $\mathbb{E}[\mathcal{T}_2]$ *depends on the zero-mean likelihood terms and, specifically, on the order of* $\mathbb{E}\left[\ell_{j;p_1}\ell_{j;a'}\ell_{j;d'}\right]$*. Let* $f(Y_{js}) = \ell_{js;p_1}$*,* $g(Y_{jt}) = \ell_{jt;a'}$ *and* $h(Y_{jw}) = \ell_{jw;d'}$ *be defined similarly as in Examples B.1 and B.2. Then,*

$$\mathbb{E}[\mathcal{T}_2] \;=\; O(1)\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\ell_{j;p_1}\ell_{j;a'}\ell_{j;d'}\right] = O(1)\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\frac{1}{T^3}\sum_{s=1}^T\sum_{t=1}^T\sum_{w=1}^T f(Y_{js})g(Y_{jt})h(Y_{jw})\right],$$

which is $O(T^{-2})$ *by Lemma B.3 (remember that (11) holds for any* $i, j, k = 1, ..., N$*).*

Example B.4 *Finally, consider* $\mathcal{T}_3 = \upsilon_{r_1,r_2,r_3}\upsilon^{r_1,q}\ell_{\bar{r}_1}\upsilon^{\bar{r}_1,r_2}\ell_{\bar{r}_1}\upsilon^{\bar{r}_1,\tilde{r}_2}\ell_{\bar{r}_1}\upsilon^{\bar{r}_1,\tilde{r}_3}\upsilon_{\tilde{r}_1,\tilde{r}_2,\tilde{r}_3}\upsilon^{\tilde{r}_1,r_3}$*. The pattern we have been using so far is now clear: the order of the expected value of any term depends directly on the order of the expected value of the product of the zero-mean likelihood derivatives. If a given term,* $\mathcal{T}$*, contains only one zero-mean likelihood derivative, then* $\mathcal{T}$ *is exactly equal to zero. If there are two zero-mean likelihood terms involved, as with* $\mathcal{T}_0$*, then* $\mathbb{E}[\mathcal{T}] = O(T^{-1})$*. If, on the other hand, there are three zero-mean terms, then* $\mathbb{E}[\mathcal{T}] = O(T^{-2})$*, as with* $\mathcal{T}_1$ *and* $\mathcal{T}_2$*. In this specific example, the terms comprising* $\upsilon_{r_1,r_2,r_3}\upsilon^{r_1,q}\upsilon^{\bar{r}_1,r_2}\upsilon^{\bar{r}_1,\tilde{r}_2}\upsilon^{\bar{r}_1,\tilde{r}_3}\upsilon_{\tilde{r}_1,\tilde{r}_2,\tilde{r}_3}\upsilon^{\tilde{r}_1,r_3}$ *are all* $O(1)$*. Hence,* $\mathbb{E}[\mathcal{T}_3] = O(1)\mathbb{E}[\ell_{\bar{r}_1}\ell_{\bar{r}_1}\ell_{\bar{r}_1}]$ *and by using the same arguments as above, one can easily show that, by (13) in Lemma B.3,* $\mathbb{E}[\ell_{\bar{r}_1}\ell_{\bar{r}_1}\ell_{\bar{r}_1}]$ *and, therefore,* $\mathbb{E}[\mathcal{T}_3]$ *are* $O(T^{-2})$*.*

By using the same methods, one can finally show that

$$\begin{aligned}\mathbb{E}[\delta_q] \;=\;& \upsilon^{r_1,q}\frac{1}{N}\sum_{j=1}^N \mathbb{E}[\ell_{j;r_1,m}\ell_{j;p_1}\upsilon_j^{p_1,m}] + \upsilon^{\bar{r}_1,r_2}\upsilon^{r_1,q}\mathbb{E}\left[\ell_{\bar{r}_1}\mathcal{H}_{r_1,r_2}\right] \\ & -\frac{1}{2}\mathbb{E}\left[\ell_{\bar{r}_1}\upsilon^{\bar{r}_1,r_2}\ell_{\tilde{r}_1}\upsilon^{\tilde{r}_1,r_3}\upsilon_{r_1,r_2,r_3}\upsilon^{r_1,q}\right] - \upsilon^{r_1,q}\frac{1}{2N}\sum_{j=1}^N \mathbb{E}[\upsilon_{j;r_1,m,n}\ell_{j;p_1}\upsilon_j^{p_1,m}\ell_{j;p_1'}\upsilon_j^{p_1',n}] \\ & +O(T^{-2}), \qquad (27)\end{aligned}$$

where all terms other than the remainder are $O(T^{-1})$. This characterises the first-order bias terms.

B.3.5 Third step: from index to matrix notation

We will now prove Theorems 4.2 and 4.3, and show that

$$\mathcal{A}_{NT}(\theta_0, \lambda_{10}, ..., \lambda_{N0}) \;=\; \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}$$

$$\times\left[\frac{1}{N}\sum_{j=1}^{N}\left(\left(\frac{d}{d\theta}\frac{\partial \ell_{jT}}{\partial \lambda_j'}\right)\left\{\mathbb{E}\left[\frac{\partial^2 \ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\frac{\partial \ell_{jT}}{\partial\lambda_j}\right) - \frac{1}{2N}\sum_{j=1}^{N} M_j'\right.$$

$$\left. + \left\{\frac{d^2\ell_{NT}}{d\theta d\theta'} - \mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{d\ell_{NT}}{d\theta} - \frac{1}{2}M''\right] \tag{28}$$

where

$$M_j' = \begin{bmatrix} \frac{\partial \ell_{jT}}{\partial\lambda_j'}\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\mathbb{E}\left[\frac{d}{d\theta_1}\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\frac{\partial\ell_{jT}}{\partial\lambda_j} \\ \vdots \\ \frac{\partial \ell_{jT}}{\partial\lambda_j'}\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\mathbb{E}\left[\frac{d}{d\theta_r}\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\frac{\partial\ell_{jT}}{\partial\lambda_j} \end{bmatrix},$$

$$M'' = \begin{bmatrix} \frac{d\ell_{NT}}{d\theta'}\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\mathbb{E}\left[\frac{d^3\ell_{NT}}{d\theta_1 d\theta d\theta'}\right]\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{d\ell_{NT}}{d\theta} \\ \vdots \\ \frac{d\ell_{NT}}{d\theta'}\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\mathbb{E}\left[\frac{d^3\ell_{NT}}{d\theta_r d\theta d\theta'}\right]\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{d\ell_{NT}}{d\theta} \end{bmatrix}.$$

Remember that, as mentioned in Section B.3.1, the indices q and r are used to denote differentiation with respect to θ while the indices $a, b, c, d, e, f, l, m, n, o$ and p denote differentiation with respect to λ. First, $\ell_{r_1}v^{r_1,q} = \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{d\ell_{NT}}{d\theta}$, an $(R \times 1)$ vector. Next,

$$\ell_{q_1}\mathcal{H}_{r_1,r_2}v^{r_1,q}v^{q_1,r_2} = \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\left\{\frac{d^2\ell_{NT}}{d\theta d\theta'} - \mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{d\ell_{NT}}{d\theta}. \tag{29}$$

To see this, notice that $\ell_{q_1}v^{q_1,r_2}$ is the same as $\ell_{r_1}v^{r_1,q}$ while,

$$\mathcal{H}_{r_1,r_2}v^{r_1,q} = \underbrace{\left\{\frac{d^2\ell_{NT}}{d\theta d\theta'} - \mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}\left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}}_{(R\times R)}.$$

Similarly,

$$v^{r_1,q}\frac{1}{N}\sum_{j=1}^{N}\ell_{j;p_1}\ell_{j;r_1,m}v_j^{p_1,m} = \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\frac{1}{N}\sum_{j=1}^{N}\left(\left(\frac{d}{d\theta}\frac{\partial\ell_{jT}}{\partial\lambda_j'}\right)\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\frac{\partial\ell_{jT}}{\partial\lambda_j}\right).$$

Next, consider $v^{r_1,q}\sum_{j=1}^{N}\ell_{j;p_1}\ell_{j;p_2}v_{j;r_1,m,n}v_j^{p_1,m}v_j^{p_2,n}$. Observe first that

$$\ell_{j;p_1}v_j^{p_1,m} = \ell_{j;p_2}v_j^{p_2,n} = \underbrace{\left\{\mathbb{E}\left[\frac{\partial^2\ell_{jT}}{\partial\lambda_j\partial\lambda_j'}\right]\right\}^{-1}\frac{\partial\ell_{jT}}{\partial\lambda_j}}_{(P\times 1)}.$$

Then, the $(R \times 1)$ array $v_{j;r_1,m,n}\ell_{j;p_1}v_j^{p_1,m}\ell_{j;p_2}v_j^{p_2,n}$ is the same as M_j'. Therefore,

$$v^{r_1,q}\sum_{j=1}^{N}\ell_{j;p_1}\ell_{j;p_2}v_{j;r_1,m,n}v_j^{p_1,m}v_j^{p_2,n} = \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}\sum_{j=1}^{N}M_j'.$$

Following a similar pattern, one can show that, $\ell_{q_1}\ell_{q_4}v_{r_1,r_2,r_3}v^{r_1,q}v^{q_1,r_2}v^{q_4,r_3} = \left\{\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1}M''$.

Proof of Theorem 4.2. By (26), $\hat{\theta} - \theta_0 = \left\{ -\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right]\right\}^{-1} \frac{d\ell_{NT}}{d\theta} + o_p(1)$ where

$$\begin{aligned}
\frac{d\ell_{NT}}{d\theta} &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\left\{\frac{\partial \ell_{jt}}{\partial\theta} - \mathbb{E}\left[\frac{\partial^2 \ell_{jT}}{\partial\theta\partial\lambda'}\right]\left\{\mathbb{E}\left[\frac{\partial^2 \ell_{jT}}{\partial\lambda\partial\lambda'}\right]\right\}^{-1}\frac{\partial \ell_{jt}}{\partial\lambda}\right\} \\
\mathbb{E}\left[\frac{d^2\ell_{NT}}{d\theta d\theta'}\right] &= \mathbb{E}\left[\frac{\partial^2\ell_{NT}}{\partial\theta\partial\theta'}\right] - \left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\frac{\partial \ell_{jT}}{\partial\theta\partial\lambda'}\right]\right\}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\frac{\partial^2 \ell_{jT}}{\partial\lambda\partial\lambda'}\right]\right\}^{-1}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\frac{\partial^2 \ell_{jT}}{\partial\lambda\partial\theta'}\right]\right\}.
\end{aligned}$$

Let $\mathcal{I} = \lim_{N,T\to\infty}\mathcal{I}_{NT}$ where $\mathcal{I}_{NT} = Var\left(\sqrt{T}\frac{d\ell_{NT}}{d\theta}\right)$. Then, for any $(R \times 1)$ vector γ such that $\gamma'\gamma = 1$, $\gamma'\sqrt{T}\mathcal{I}_{NT}^{-1/2}\sum_{j=1}^{N}\sum_{t=1}^{T}\frac{d}{d\theta}\ell_{jt}$ is a linear combination of mixing processes and, therefore, is a mixing process itself. Moreover, $Var\left(\gamma'\sqrt{T}\mathcal{I}_{NT}^{-1/2}\frac{d\ell_{NT}}{d\theta}\right) = 1$. Therefore, $\gamma'\sqrt{T}\mathcal{I}_{NT}^{-1/2}\frac{d\ell_{NT}}{d\theta} \overset{d}{\to} \mathcal{N}(0,1)$ and by the Cramér-Wold device $\sqrt{T}\frac{d\ell_{NT}}{d\theta} \overset{d}{\to} \mathcal{N}(0,\mathcal{I})$. Hence, by Slutsky's Theorem, $\hat{\theta} - \theta_0 \overset{d}{\to} \mathcal{N}(0,\mathcal{D}^{-1}\mathcal{I}\mathcal{D}^{-1})$. ■

Proof of Theorem 4.3. Rewriting (27) in matrix notation by using the results of section B.3.5 and remembering that all terms in (27) other than the remainder are $O(T^{-1})$ proves (28) and the theorem. ■