LONDON SCHOOL OF ECONOMICS

Estimating Parameters in the Presence of Many Nuisance Parameters

by

Billy Wu

Thesis submitted in fulfilment for the
MPhil Degree

in the
Department of Statistics

Supervisor: Prof. Qiwei Yao

January 2013

Declaration of Authorship

I, BILLY WU, declare that this report titled, 'Estimating Parameters in the Presence of Many Nuisance Parameters' and the work presented in it are my own.

Signed: ____________________________________

Date: ____________________________________

Abstract

This paper considers estimation of parameters for high-dimensional time series with the presence of many nuisance parameters. In particular we are interested in data consisting of p time series of length n, with p to be as large or even larger than n. Here we consider the composite-likelihood estimation and the profile quasi-likelihood estimation. The asymptotic properties of these methodologies are investigated. Simulations are used to illustrate our both of these methods and explore the performance of these methods.

Key words: composite likelihood, nuisance parameter, profile likelihood, quasi-likelihood, root-n convergence, time series.

Acknowledgements

I would like to thank my advisor Professor Qiwei Yao, he gave me a huge amount of support and guidance. He has a seemingly endless well of knowledge and an incredible amount of patience. I hope to one day repay his immeasurable amount of help.

I would also like to thank my family for their all their support. I owe them a great deal for everything in my life, and for making me the person I am today. My parents and my grandparents raised me and nurtured me, I have had an incredible life experience. They always had my best interests at heart and worked extremely hard to give me the best opportunities available, I will forever be grateful.

Last but not least I would also like to thank all my friends. Without their moral support and friendship, my entire life would not be the same. I am fortunate to have such amazing friends.

Contents

Chapter 1

Introduction

1.1 Introduction

Rapid developments in technology in this information age has led to data collection in an unprecedentedly large scale. This brings a new opportunity with challenge to statistics. The availability of large data sets enable statisticians to look into complex structures using sophisticated models. In this paper we consider a class of models in which the number of parameters of interest is finite while the number of nuisance parameters is large or excessively large in relation to the sample size. Those models arise in various statistical applications. For example, in a longitudinal model with a large number of sites the primary interest lies in a small number of parameters representing the common effects while the individual levels of different sites are treated as nuisance parameters. For a large panel of time series data, one is often interested in a few common factors which drives the dynamics of all the component series and treats the parameters representing each idiosyncratic components as nuisance parameters. In the attempts to model the volatilities of large number of financial securities, it is often to assume that the dynamic volatilities are controlled by a small number of parameters in the presence of large number of nuisance parameters representing marginal covariance matrices.

In this paper we consider two methods to obtain the estimators of a fixed number of parameters of interest in presence of a large number of nuisance parameters. The methods concerned are the maximum profile quasi-likelihood estimation (MPQLE) and the maximum composite quasi-likelihood estimation (MCQLE). With an initial estimator for the nuisance parameter vector, the MPQLE maximises a profile quasi-likelihood function to obtain the estimation. This is in line with more conventional approach. By plugging in an initial estimator for the nuisance parameters, we avoid a maximisation problem with a large number of variables. However it is intuitively clear that the quality of the initial estimator impacts on the ultimate outcome of the procedure.

Another method to be considered is the composite likelihood, the name coined by Lindsay (1988). See also a recent survey Varin et al. (2011). A composite likelihood is a function derived by multiplying a collection of, typically two- or there-dimensional, marginal density functions. In our context, each low dimensional density function only depends on a small number of nuisance parameters, hence can be easily profiled. The resulting composite profile likelihood function depends on those parameters of interest only, can be solved to obtain the estimator without running into high-dimensional optimisation problems. Because the marginal densities are multiplied together, ignoring the original distribution structure, the MCQLE can be viewed as derived from a (seriously) misspecified model.

The major contribution of this paper is the establishment of the asymptotic properties for both the MCQLE and the MPQLE under the condition which is relevant to the settings concerned. The conventional asymptotic theory is typically under the assumption that the sample size goes to infinity while everything else remains fixed. For our setting, the number of nuisance parameters is of a comparable magnitude of the sample size. Hence it is more pertinent to consider the asymptotics when both the sample size and the number of nuisance parameters go to infinity together. Though bearing a similar banner, our theory is different from large literature on the theory for the so-called 'large p and small n' regression problem; see, among the others, Zou (2006), Fan and Lv (2008), Huang et al (2008), Zhang and Huang (2008), Bickel et al (2009) and Zhang (2009).

The name of 'composite likelihood' was introduced by Lindsay (1988), although the idea of using 'submodels' or 'marginal models' had appeared before. As the full likelihood with complex models are often computationally infeasible, The composite likelihood methods have been used in different regression with dependent errors (Eicher 1967), problems including modelling spatial processes (Besag 1974), case control studies (Liang 1987), inference for nonlinear dynamic models (Gallant and White 1988), correlated binary data (Kuk and Nott 2000), grouped data (deLeon 2005), longitudinal studies (Molenberghs and Verbeke 2005), multivariate volatility modeling (Engle et al. 2008), bioinformatice (Larribe and Fearnhead 2011). The asymptotic theory under the assumption that only sample size tends to infinity has been studies by, for example, Cox (1961), Eicher (1967), White (1982), Gallant and White (1988), and Cox and Reid (2004). To our knowledge, no results have been derived under our setting when both the sample size and the number of nuisance parameters go to infinity together. For more comprehensive survey in the composite likelihood methodology, we refer to the first issue *Statistica Sinica* (2011) vol.21 which contains a collection of the papers on this topic.

The rest of the paper is organised as follows. Section 2 deals with the MCQLE and section 3 is on MPQLE. In each of those two sections, we outline the method and state the asymptotic normality results. Both the methods are illustrated in simulation reported in section 4. All technical proofs are given in section 5.

Chapter 2

Methodology

2.1 Composite-likelihood estimation

Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be $p \times 1$ observations from a stationary process with the underlying distribution depending on parameter $(\boldsymbol{\theta}, \boldsymbol{\omega}) \in \Theta \times \Omega \subset \mathcal{R}^{d+q}$, where $\boldsymbol{\theta}$ is a $d \times 1$ parameter of interest, and $\boldsymbol{\omega}$ is a $q \times 1$ nuisance parameter. Our goal is to estimate $\boldsymbol{\theta}$. We consider now a maximum composite quasi-likelihood estimation method for $\boldsymbol{\theta}$. We will show that such an estimator is asymptotically normal with the standard root-n convergence rate as $n, q \to \infty$ together while d is fixed, and p may also diverge to infinity.

Let $\mathbf{X}_{t1}, \cdots, \mathbf{X}_{tr}$ be r subvectors of $\mathbf{X}_t$. The lengths of those r subvectors may be different from each other, and some of those subvectors may share common components from $\mathbf{X}_t$. With the observations $\mathbf{X}_{tj}, t = 1, \cdots, n$, the log marginal quasi-likelihood function is defined as

$$l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \sum_{t=1}^{n} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j),$$

which depends on the parameter of interest $\boldsymbol{\theta}$, and a subset of nuisance parameter denoted by $\boldsymbol{\omega}_j$. Let

$$\widetilde{\boldsymbol{\omega}}_j(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\omega}_j} l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j). \tag{2.1.1}$$

We define a composite quasi-likelihood function for $\boldsymbol{\theta}$ as

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} l_j\big(\boldsymbol{\theta}, \widetilde{\boldsymbol{\omega}}_j(\boldsymbol{\theta})\big). \tag{2.1.2}$$

The maximum composite quasi-likelihood estimator (MCQLE) for $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{2.1.3}$$

We assume that $r = r(q) \to \infty$ as $q \to \infty$, while all the lengths of $\mathbf{X}_{tj}$ and $\boldsymbol{\omega}_j$ are fixed.

One implicit condition for the MCQLE defined in (2.1.3) being reasonable is that the nuisance parameters $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r$ are distinct from each other such that the maximisation (2.1.1) may be carried out independently for each j without confounding constraints from each other. This is a rather strong requirement, and may only be facilitated by selecting subvectors $\mathbf{X}_{t1}, \cdots, \mathbf{X}_{tr}$ in a restrictive manner. It is very likely there may be a heavy loss of information if we adhere to this requirement in practice. One alternative is to adopt the so-called 'variation-free' condition imposed by Engle, Hendry and Richard (1983), which ignores the links among different $\boldsymbol{\omega}_j$ and treats $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r$ as different and unconnected nuisance parameters. See also Engle, Shephard and Sheppard (2008). Of course there will be some efficiency loss in estimation of $\boldsymbol{\theta}$ resulted from neglecting the links among different $\boldsymbol{\omega}_j$. The trade-off is that we will be able to reduce an extra high-dimensional optimisation problem to many low-dimensional problems, which is the essential motivation of using composite-likelihood approach. Note that this variation-free condition also implies that $\widehat{\boldsymbol{\theta}}$ is the global maximiser in the sense that

$$(\widehat{\boldsymbol{\theta}},\ \widehat{\boldsymbol{\omega}}_1,\ \cdots,\ \widehat{\boldsymbol{\omega}}_r) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_r} \sum_{j=1}^{r} l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j),$$

where we treat $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r$ as different and independent parameters. In the rest of this section, we always adopt this assumption.

Let $\boldsymbol{\beta} = (\boldsymbol{\theta}', \boldsymbol{\omega}_1', \cdots, \boldsymbol{\omega}_r')'$, and $l(\boldsymbol{\beta}) = \sum_{j=1}^{r} l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j)$. In practice we take $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\omega}}_1', \cdots, \widehat{\boldsymbol{\omega}}_r')'$ as a solution of the likelihood equation

$$\dot{l}(\widehat{\boldsymbol{\beta}}) \ \equiv\ \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \ =\ 0. \tag{2.1.4}$$

Let

$$\boldsymbol{\beta}_o \equiv (\boldsymbol{\theta}_o', \boldsymbol{\omega}_{1o}', \cdots, \boldsymbol{\omega}_{ro}')' = \arg\max_{\boldsymbol{\theta},\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_r} E\{\sum_{j=1}^{r} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j)\} \tag{2.1.5}$$

be the true value of the parameter, which is assumed to be an inner point of the parameter space. Put

$$\mathbf{a}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j), \quad \mathbf{b}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \frac{\partial}{\partial \boldsymbol{\omega}_j} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j),$$

$$\mathbf{A}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j), \quad \mathbf{B}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}_j'} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j),$$

$$\mathbf{C}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) = \frac{\partial^2}{\partial \boldsymbol{\omega}_j \partial \boldsymbol{\omega}_j'} \log f_j(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j).$$

We simply write $\mathbf{a}_{tj} = \mathbf{a}_{tj}(\boldsymbol{\beta}_o, \boldsymbol{\omega}_{jo})$, and $\mathbf{b}_{tj}, \mathbf{A}_{tj}, \mathbf{B}_{tj}$ and $\mathbf{C}_{tj}$ in the same manner. Put

$$\mathbf{M}_1 = -\begin{pmatrix} \sum_{j=1}^r E\mathbf{A}_{tj} & E\mathbf{B}_{t1} & \cdots & E\mathbf{B}_{tr} \\ E\mathbf{B}'_{t1} & E\mathbf{C}_{t1} & & \\ \vdots & & \ddots & \\ E\mathbf{B}'_{tr} & & & E\mathbf{C}_{tr} \end{pmatrix}, \tag{2.1.6}$$

$$\mathbf{M}_2 = -\begin{pmatrix} \frac{1}{r}\sum_{j=1}^r E\mathbf{A}_{tj} & \frac{1}{\sqrt{r}}E\mathbf{B}_{t1} & \cdots & \frac{1}{\sqrt{r}}E\mathbf{B}_{tr} \\ \frac{1}{\sqrt{r}}E\mathbf{B}'_{t1} & E\mathbf{C}_{t1} & & \\ \vdots & & \ddots & \\ \frac{1}{\sqrt{r}}E\mathbf{B}'_{tr} & & & E\mathbf{C}_{tr} \end{pmatrix}, \tag{2.1.7}$$

and the elements at the blank places in the above matrices are 0.

We introduce some regularity conditions first.

A1 $\{\mathbf{X}_t\}$ satisfies the mixing condition stated in C3 in the Appendix.

A2 f_j are smooth enough such that all the required derivatives exist and are continuous and integrable whenever necessary.

A3 Denote by ξ_{tj} any component of $\mathbf{a}_{tj}$, and η_{tj} any component of $\mathbf{b}_{tj}$. For $\nu > 2$ given in A1 above, it holds that

$$\overline{\lim_{r\to\infty}}\, E\{|\frac{1}{r}\sum_{j=1}^r \xi_{tj}|^\nu\} < \infty, \tag{2.1.8}$$

$$\overline{\lim_{r\to\infty}}\, \frac{1}{r}\sum_{j=1}^r [E(\eta_{tj}^2) + \{E(|\eta_{tj}|^\nu)\}^{2/\nu}] < \infty. \tag{2.1.9}$$

By Hlder's inequality this is equivalent to

$$\overline{\lim_{r\to\infty}}\, \frac{1}{r}\sum_{j=1}^r [E(|\eta_{tj}|^\nu)^{2/\nu}] < \infty. \tag{2.1.10}$$

A4 Denote by η_{tj} any element of $\mathbf{A}_{tj} - E(\mathbf{A}_{tj})$, $\mathbf{B}_{tj} - E(\mathbf{B}_{tj})$ or $\mathbf{C}_{tj} - E(\mathbf{C}_{tj})$. Then (2.1.9) holds.

A5 The matrix $\mathbf{M}_1$ is positive-definite. Furthermore all the eigenvalues of the matrix $\mathbf{M}_2$ are bounded above from ∞ and below from 0, as $r \to \infty$.

A6 There exists a constant $c_1 > 0$ and positive functions $\lambda_j(\cdot)$ such that $|\frac{\partial^3}{\partial\beta_\ell\partial\beta_i\partial\beta_k}\log f_j(\mathbf{x}_j;\boldsymbol{\theta},\boldsymbol{\omega}_j)| \le \lambda_j(\mathbf{x}_j)$ for any $||\boldsymbol{\theta}-\boldsymbol{\theta}_o|| \le c_1$ and $||\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo}|| \le c_1$. Furthermore $\overline{\lim}_{r\to\infty}\sup_{1\le j\le r} E\{\lambda_j(\mathbf{X}_{tj})\} < \infty$,, and (2.1.9) holds with with $\eta_{tj} = \lambda_j(\mathbf{X}_{tj}) - E\{\lambda_j(\mathbf{X}_{tj})\}$.

A7 (2.1.9) holds with η_{tj} being any component of $\boldsymbol{\zeta}_{tj} \equiv \mathbf{a}_{tj} - E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}\mathbf{b}_{tj}$. Furthermore the limits of the convariance
$\mathbf{W}_k = \lim_{r\to\infty} \frac{1}{r^2}\big(\sum_{j=1}^r \boldsymbol{\zeta}_{1j},\ \sum_{j=1}^r \boldsymbol{\zeta}_{k+1,j}\big), \qquad k = 0, 1, \cdots, n$. Exists

Remark 1. (i) Note that $\mathbf{M}_1 = -E\Big\{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \sum_{j=1}^r \log f_j(\mathbf{X}_{tj};\boldsymbol{\theta},\boldsymbol{\omega}_j)\Big\}$. The condition that $\mathbf{M}_1 > 0$ in A5 implies that $\boldsymbol{\beta}_o$, defined in (2.1.5), is an isolated maximiser. It also implies that $\mathbf{M}_2$ is positive-definite as $\mathbf{M}_2 = \boldsymbol{\Lambda}\mathbf{M}_1\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is an appropriate full-ranked diagonal matrix.

(ii) If $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are independent observations, conditions A3, A4 and A6 may be reduced to those with $\nu = 2$ only.

2.1.1 Theorem 1

Theorem 1. Let conditions A1 – A6 hold. Then there exists a solution of the likelihood equation (2.1.4) for which

$$m\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o||^2 + \frac{1}{r}\sum_{j=1}^r ||\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}||^2\} \xrightarrow{P} 0$$

for any $m \to \infty$, $r/m \to 0$ and $r^2 m/n \to 0$.

Remark 2. The convergence rates in Theorem 1 are not optimal; see, for example, Theorem 2 below which indicates that the convergence rate for $\widehat{\boldsymbol{\theta}}$ is root-n. The important message here is the difference in the convergence rates between $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{\boldsymbol{\omega}}_j,\ j = 1, \cdots, r\}$. As $r \to \infty$ together with n, the rate for the uniform convergence of $\widehat{\boldsymbol{\omega}}_1, \cdots, \widehat{\boldsymbol{\omega}}_r$ is slower. It also imposes some restrictions on the number of parameters which can be consistently estimated, although the implied rates such as $r = o(n^{1/3})$ is presumably too restrictive. In region of the true parameter if there is a unique solution then this would be the true min.

2.1.2 Theorem 2

Theorem 2. Let conditions A1 – A7 hold, matrices $E(\mathbf{C}_{1j})$, $j = 1, \cdots, r$, be invertible, and the limit of M_2, defined in (2.1.7), exist (as $r \to \infty$). Furthermore, let $r/n \to 0$. For any consistent solution of the likelihood equation (2.1.4) in the sense that

$$||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o||^2 + \sum_{j=1}^r ||\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}||^2 \xrightarrow{P} 0, \tag{2.1.11}$$

it holds that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\Big(0,\ \mathbf{L}^{-1}\big(\mathbf{W}_0 + 2\sum_{k=1}^{\infty}\mathbf{W}_k\big)\mathbf{L}^{-1}\Big),$$

where $\mathbf{W}_k$ are defined in A7, and $\mathbf{L} = \lim_{r\to\infty} r^{-1}\sum_{j=1}^{r}\{E(\mathbf{A}_{1j}) - E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}E(\mathbf{B}'_{1j})\}$.

Remark 3. (i) The consistence condition (2.1.10) is weaker than that identified in Theorem 1, as $m/r \to \infty$.

(ii) The limit which defines the matrix $\mathbf{L}$ exists. This is implied by the existence of the limit of M_2.

2.2 Profile quasi-likelihood estimation

We consider now the asymptotic properties of a qMLE for $\boldsymbol{\theta}$, obtained based on a reasonable initial estimator for the nuisance parameter $\boldsymbol{\omega}$. We will show that the qMLE is asymptotically normal with the standard root-n convergence rate in spite that the number of nuisance parameters q goes to ∞.

We use a log quasi-likelihood function

$$l(\boldsymbol{\theta},\ \boldsymbol{\omega}) = \sum_{t=1}^{n} \log f(\mathbf{X}_t;\ \boldsymbol{\theta},\ \boldsymbol{\omega}), \tag{2.2.12}$$

where f is a density function defined on $\mathcal{R}^p$. With an initial estimator $\widehat{\boldsymbol{\omega}}$ for the nuisance parameter $\boldsymbol{\omega}$, a profile quasi-likelihood function for $\boldsymbol{\theta}$ is defined as

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log f(\mathbf{X}_t;\ \boldsymbol{\theta},\ \widehat{\boldsymbol{\omega}}),$$

and the maximum profile quasi-likelihood estimator (MPQLE) is defined as

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\ l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{n} \log f(\mathbf{X}_t;\ \boldsymbol{\theta},\ \widehat{\boldsymbol{\omega}}).$$

Let $(\boldsymbol{\theta}_o, \boldsymbol{\omega}_o) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\omega}} E\{\log f(\mathbf{X}_t; \boldsymbol{\theta}, \boldsymbol{\omega})\}$ be the true parameter values. Since $\dot{l}(\widetilde{\boldsymbol{\theta}}) = 0$, it follows a Taylor expansion that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)\ =\ -\big\{\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^\star)\big\}^{-1}\frac{1}{m\sqrt{n}}\,\dot{l}(\boldsymbol{\theta}_o), \tag{2.2.13}$$

where $\boldsymbol{\theta}^\star$ is between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, and m is a normalised constant depending on q.

We introduce the regularity conditions first. Let

$$\dot{l}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \qquad \mathbf{a}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}),$$

$$\mathbf{B}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}), \qquad \mathbf{C}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}'} \log f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}),$$

and $D(\boldsymbol{\theta}, \boldsymbol{\omega}) = E\{\mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}, \boldsymbol{\omega})\}$.

B1 The initial estimator $\widehat{\boldsymbol{\omega}} = (\widehat{\omega}_1, \cdots, \widehat{\omega}_q)'$ is asymptotically linear in the sense that for each $1 \leq j \leq q$, $\widehat{\omega}_j - \omega_{jo} = \frac{1}{n}\sum_{t=1}^n g_j(\mathbf{X}_t) + o_P(n^{-1/2})$, where $E\{g_j(\mathbf{X}_t)\} = 0$, $\text{Var}\{g_j(\mathbf{X}_t)\} \leq c < \infty$, and $c > 0$ is a constant independent of j. Furthermore $||\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o||^2 = O_P(q/n)$, and $q/n \to 0$.

B2 $f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega})$ is smooth such that all the required partial derivatives exists and are continuous. Denoted by a_j the j-th component of $\mathbf{a}$. There exists a positive number c_1 and a positive function $\lambda_1(\cdot)$

such that

$$\left|\mathbf{u}' \frac{\partial^2 a_j(\mathbf{x}; \boldsymbol{\theta}_o, \boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'} \mathbf{u}\right| \leq \lambda_1(\mathbf{x}) ||\mathbf{u}||^2 \quad \text{for any } ||\boldsymbol{\omega} - \boldsymbol{\omega}_o|| \leq c_1, \ \mathbf{u} \in \mathcal{R}^q \ \text{ and } \ 1 \leq j \leq q,$$

and $E\{\lambda_1(\mathbf{X}_t)\}$ is bounded (as $q \to \infty$). Furthermore $q/(m\sqrt{n}) \to 0$.

B3 $\{\mathbf{X}_t\}$ satisfies condition C1 in the Appendix, and

$$\psi_n(\mathbf{X}_t, \mathbf{X}_s) = \{\mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s) + \mathbf{C}(\mathbf{X}_s; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)\}/m$$

satisfies condition C2.

B4 For some $\gamma > 2$ and $\gamma > \delta'$ given in C1, $\overline{\lim}_{q \to \infty} E\{||\mathbf{a}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o) + 2\mathbf{D}(\boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)||^\gamma\}/m^\gamma < \infty$. Furthermore

$$\boldsymbol{\Sigma}_j \equiv \lim_{q \to \infty} \frac{1}{m^2} \text{Cov}\{\mathbf{a}(\mathbf{X}_1; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o) + 2\mathbf{D}(\boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_1), \ \mathbf{a}(\mathbf{X}_{1+j}; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o) + 2\mathbf{D}(\boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_{1+j})\}$$

exists for all $j \geq 0$.

B5 Let $b_{ij}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega})$ be the (i, j)-th element of $\mathbf{B}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega})$. There exist a positive number c_2 and a positive function $\lambda_2(\cdot)$ such that $||\frac{\partial}{\partial \boldsymbol{\theta}} b_{ij}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega})|| + ||\frac{\partial}{\partial \boldsymbol{\omega}} b_{ij}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega})|| \leq \lambda_2(\mathbf{x})$ for any $||\boldsymbol{\theta} - \boldsymbol{\theta}_o|| \leq c_2$, $||\boldsymbol{\omega} - \boldsymbol{\omega}_o|| \leq c_2$ and $1 \leq i, j \leq d$, the limit of $E\{b_{ij}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\}/m$ exists, and both $E\{\lambda_2(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)^\nu\}/m^\nu$ and $E\{b_{ij}(\mathbf{X}_t; \theta_o, \boldsymbol{\omega}_o)^\nu\}/m^\nu$ are bounded (as $q \to \infty$), where $\nu > 2$ is given as in C3. Furthermore, $\widetilde{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.

2.2.1 Theorem 3

Theorem 3. Under conditions B1-B5, $\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ is asymptotically normal with mean 0 and covariance matrix $\mathbf{M}^{-1}(\boldsymbol{\Sigma}_0 + 2\sum_{j=1}^{\infty} \boldsymbol{\Sigma}_j)\mathbf{M}^{-1}$, where $\mathbf{M} = \lim_{q\to\infty} E\{\mathbf{B}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\}/m > 0$, and $\boldsymbol{\Sigma}_j$ is defined in B4.

Chapter 3

Numerical Properties

3.1 Example 1

One-way error component model for panel data (Baltagi 2005, Chapter 2). Let

$$Y_{tj} = \mu + \mathbf{X}'_{tj}\boldsymbol{\beta} + \mu_j + \varepsilon_{tj}, \qquad t = 1, \cdots, n;\ j = 1, \cdots, r.$$

In the above model Y_{tj} is the observation on the j-th individual at the time t, $\mathbf{X}_{tj}$ is the $k \times 1$ observation on k explanatory variables, μ_j denotes the individual-specific effect, ε_{tj} are i.i.d. noise with mean 0 and variance σ^2. To make the parameters identifiable, we assume that $\sum_j \mu_j = 0$. Suppose that we are interested in the effect of the explanatory variables on individuals. Therefore we would like to estimate $\boldsymbol{\beta}$ and μ, treating $\mu_1, \cdots, \mu_r$ as nuisance parameters. We consider the case that r is large in relation to n while k is fixed.

The conventional approach is to treat μ_j i.i.d. from an unknown distribution. Then the MCQLE may be viewed as a conditional inference on μ_j. It is interesting to compare the two approaches.

As an example, Y_{tj} could be a country's GDP. X_{tj} are explanatory variables for each country j at year t such as population, literacy rate, unemployment rate etc. We are interested in the parameters μ and β for the linear regression $Y_{tj} = \mu + \mathbf{X}'_{tj}\boldsymbol{\beta}$. The time-invariant μ_j accounts for any idiosyncratic domestic productivity. An example of μ_j could be the amount the j-th country's GDP is boosted by the timber which its sub-tropical climate produces, assuming sustainable cutting and re-planting of trees this constant boost is time-invariant and specific to this country. We are not interested in such country specific factors and treat it as a nuisance parameter.

We will conduct a Monte Carlo simulation with 1000 repetitions. Newton Raphson algorithm will be used to solve the likelihood equations for this and all subsequent examples. First we generate $\boldsymbol{\beta}$ as random numbers from $\mathcal{U}(-1,1)$. Next we generate $\mu_j \sim \mathcal{N}(0,1)$, we then construct $\mu = \frac{1}{r}\sum_{j=1}^{r}(\mu_j)$. We transform $\mu_j \to \mu_j - \mu$ so that the condition $\sum_j \mu_j = 0$ is satisfied. We then generate X_{tj} where each element is a random number from $\mathcal{N}(0,1)$, let $\bar{X}_j = \frac{1}{n}\sum_{t=1}^{n}(X_{tj})$. We will transform $X_{tj} \to X_{tj} - \bar{X}_j$ so that the new X_{tj} satisfy the condition $\frac{1}{n}\sum_{t=1}^{n}(X_{tj}) = 0$. Finally we generate the dependent variable $Y_{tj} = \mu + \mathbf{X}'_{tj}\boldsymbol{\beta} + \mu_j + \varepsilon_{tj}$ by adding the i.i.d noise $\varepsilon_{tj} \sim \mathcal{N}(0,1)$.

We wish to minimise the log likelihood function of the form

$$l(\mu, \boldsymbol{\beta}, \widehat{\mu}) = -\frac{1}{2}\sum_j \sum_t (Y_{tj} - \mu - X'_{tj}\boldsymbol{\beta} - \mu_j)^2,$$

FOR MCQLE we will use 2 different approaches. First we pick all subsets $(\mathbf{X}_{ti}, \mathbf{X}_{tj})$ for $1 \le i \ne j \le r$. The number of subsets is $r(r-1)/2$ and will grow rapidly as r increases, which is computationally costly. Our second approach we will pick consecutive subsets of $(\mathbf{X}_{ti}, \mathbf{X}_{tj})$ with $1 \le i \le r-1, andj = i+1$.

For MPQLE we'll need to find an initial estimator $\widehat{\mu}_j$.

$$\frac{1}{n}\sum_{t=1}^{n}(Y_{tj}) = \mu + \frac{1}{n}\sum_{t=1}^{n}(X'_{tj}\boldsymbol{\beta}) + \mu_j + \frac{1}{n}\sum_{t=1}^{n}(\varepsilon_{tj}),$$

Since we constructed X_{tj} such that $\frac{1}{n}\sum_{t=1}^{n}(X_{tj}) = 0$ and $\varepsilon_{tj} \sim \mathcal{N}(0,1)$ we have:

$$\frac{1}{n}\sum_{t=1}^{n}(Y_{tj}) = \mu + \mu_j,$$

$$\mu = \frac{1}{n}\frac{1}{r}\sum_{t=1}^{n}\sum_{j=1}^{r}(Y_{tj}),$$

$$\widehat{\mu}_j = \frac{1}{n}\sum_{t=1}^{n}(Y_{tj}) - \frac{1}{n}\frac{1}{r}\sum_{t=1}^{n}\sum_{j=1}^{r}(Y_{tj}),$$

Table 1: The root mean square error of the methods over 1000 replications. n is the number of observation, and r the number of nuisance parameters.

	RMSE											
	MPQLE				MCQLE				MCQLE consecutive			
n	μ	β1	β2	β3	μ	β1	β2	β3	μ	β1	β2	β3
					r=10							
20	0.0703	0.0726	0.0745	0.0720	0.0691	0.0758	0.0714	0.0707	0.0679	0.0729	0.0735	0.0710
50	0.0459	0.0453	0.0456	0.0454	0.0498	0.0431	0.0474	0.0515	0.4210	0.0466	0.0447	0.0426
100	0.0319	0.0309	0.0318	0.0316	0.0323	0.0365	0.0325	0.0334	0.0308	0.0320	0.0316	0.0323
200	0.0224	0.0230	0.0213	0.0207	0.0238	0.0212	0.0239	0.0175	0.0231	0.0227	0.0220	0.0221
					r=20							
20	0.0499	0.0508	0.0516	0.0511	0.0525	0.0531	0.0451	0.0606	0.0501	0.0530	0.0542	0.0518
50	0.0308	0.0318	0.0321	0.0316	0.0310	0.0318	0.0335	0.0313	0.0309	0.0302	0.0309	0.0293
100	0.0229	0.0230	0.0223	0.0221	0.0211	0.0232	0.0210	0.0229	0.0228	0.0229	0.0227	0.0221
200	0.0152	0.0161	0.0157	0.0167	0.0130	0.0159	0.0151	0.0147	0.0167	0.0153	0.0155	0.0160
					r=50							
20	0.0324	0.0338	0.0327	0.0345	0.0328	0.0324	0.0348	0.0334	0.0308	0.0326	0.0317	0.0330
50	0.0199	0.0201	0.0199	0.0201	0.0201	0.0209	0.0203	0.0180	0.0194	0.0209	0.0200	0.0186
100	0.0142	0.0145	0.0138	0.0144	0.0138	0.0152	0.0136	0.0148	0.0143	0.0150	0.0133	0.1420
200	0.0107	0.0095	0.0098	0.0099	0.0111	0.0096	0.0099	0.1040	0.0100	0.0096	0.0102	0.1010
					r=100							
20	0.0229	0.0218	0.0223	0.0233	0.0222	0.0210	0.0214	0.0219	0.0225	0.0214	0.0228	0.0227
50	0.0138	0.0143	0.0148	0.0139	0.0133	0.0135	0.0125	0.0131	0.0136	0.0137	0.0131	0.0141
100	0.0099	0.0099	0.0099	0.0103	0.0104	0.0102	0.0096	0.0088	0.0113	0.0091	0.0110	0.0095
200	0.0072	0.0071	0.0071	0.0070	0.0069	0.0068	0.0070	0.0067	0.0070	0.0073	0.0067	0.0071

The results above illustrate the nice "blessings of dimensionality". As r increases the number of nuisance parameteres μ_j also increases, however the additional "information" available improves our estimates for μ and β. As n the number of observations increase we see improvements in our estimators as we would expect. The main hinderance to the accuracy of our estimator would come from an increase in k, the number of explanatory variables.

The performance of MPQLE and MCQLE is very similar at all n and r, it is important to remember that the performance of MPQLE depends heavily on the quality of the initial estimator, for this example our initial estimator is very good.

The quality of estimators from MCQLE and MCQLE consecutive are very close across all n and r, however the computational cost of MCQLE is very high compared to MCQLE consecutive. MCQLE is far more time efficient in practice.

3.2 Example 2

Scalar BEKK model (Engle, Shephard and Sheppard 2008). Let us consider $p \times 1$ return series $\mathbf{X}_t$ defined by

$$\mathbf{X}_t = \mathbf{H}_t^{1/2} \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim_{i.i.d.} (0,\ \mathbf{I}_p), \tag{3.2.1}$$

$$\mathbf{H}_t = (1 - \alpha - \beta)\boldsymbol{\Sigma} + \alpha \mathbf{X}_{t-1}\mathbf{X}_{t-1}' + \beta \mathbf{H}_{t-1}, \tag{3.2.2}$$

where $\alpha, \beta > 0$ are dynamic parameters, $\alpha + \beta < 1$, $\boldsymbol{\Sigma} \equiv (\sigma_{ij}) > 0$ is the unconditional covariance matrix of $\mathbf{X}_t$. Note that the model admits a strictly stationary solution. Our interest is to estimate the dynamic parameters α and β while σ_{ij} play the role of nuisance parameters. It may be shown that (3.2.2) admits the solution

$$\mathbf{H}_t = \frac{1 - \alpha - \beta}{1 - \beta}\boldsymbol{\Sigma} + \alpha \sum_{j=1}^{\infty} \beta^{j-1} \mathbf{X}_{t-j}\mathbf{X}_{t-j}'. \tag{3.2.3}$$

Assuming $\boldsymbol{\varepsilon}_t \sim N(0, \mathbf{I}_p)$, the log-likelihood function is of the form

$$l(\alpha, \beta, \boldsymbol{\Sigma}) = -\frac{1}{2}\sum_t (\log|\mathbf{H}_t| + \mathbf{X}_t' \mathbf{H}_t^{-1} \mathbf{X}_t),$$

which involves both the inverse and the determinant of $p \times p$ matrices $\mathbf{H}_t$.

For MCQLE, we may consider two options: using all binary pairs (X_{ti}, X_{tj}) for all $1 \le i \ne j \le p$, or using only the consecutive pairs $(X_{ti}, X_{t,i+1})$ for $i = 1, \cdots, p-1$. For MPQLE, we use the initial estimator

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n} (\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})',$$

where $\bar{\mathbf{X}} = n^{-1}\sum_t \mathbf{X}_t$.

First we generate a random unconditional covariance matrix σ, steps are made to ensure it is positive semi-definite. We explicitly choose true values of α and β in the region of the empirical values of α and β when using the BEKK model on equity indices such as DAX30, FTSE100, S&P500, in a similar fashion to Engle, Shephard and Sheppard 2008.

$(\alpha, \beta) = (0.1, 0.8), (0.05, 0.93), (0.02, 0.97)$ and $p = 5, 10, 50, 100$, $n = 2000$.

Then we will generate $\mathbf{X}_t$ and $\mathbf{H}_t$ stepwise. During our estimation we use the constraint that $0 < \alpha, \beta$ and $\alpha + \beta < 1$. We will do 500 repetitions in our simulation.

For MCQLE, we may consider two options: using all binary pairs (X_{ti}, X_{tj}) for all $1 \leq i \neq j \leq p$; we call this MCQLEA, or using only the consecutive pairs $(X_{ti}, X_{t,i+1})$ for $i = 1, \cdots, p-1$ we call this MCQLEB. For MPQLE, we use the MLE of the covariance matrix as the initial estimator for $\widehat{\Sigma}$. Since we use the sample covariance matrix as an initial estimator the conditions for theorem 2 is met.

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})',$$

where $\bar{\mathbf{X}} = n^{-1}\sum_t \mathbf{X}_t$.

Table 2: Root mean square error of the methods over 500 replications. Where the number of observation n is fixed at 1000, and different values are taken for the number of nuisance parameters p. 3 Sub-tables with different true values for the parameters we wish to estimate.

	RMSE					
	MCQLEA		MCQLEB		MPQLE	
P	α	β	α	β	α	β
	α = 0.02		β = 0.97			
5	0.0063	0.0395	0.0068	0.0594	0.0060	0.0388
10	0.0024	0.0062	0.0028	0.0075	0.0023	0.0058
50	0.0017	0.0040	0.0025	0.0060	0.0058	0.0139
100	0.0011	0.0031	0.0014	0.0054	0.0112	0.0203
	α = 0.05		β = 0.93			
5	0.0049	0.0142	0.0116	0.0228	0.0048	0.0144
10	0.0021	0.0074	0.0028	0.0085	0.0029	0.0069
50	0.0012	0.0050	0.0014	0.0046	0.0062	0.0151
100	0.0007	0.0036	0.0010	0.0025	0.0108	0.0231
	α = 0.10		β = 0.80			
5	0.0188	0.0397	0.0171	0.0495	0.0164	0.0365
10	0.0027	0.0148	0.0094	0.0201	0.0060	0.0121
50	0.0015	0.0087	0.0037	0.0129	0.0130	0.0150
100	0.0012	0.0064	0.0034	0.0105	0.0170	0.0193

From the table above we can see for both MCQLEA and MCQLEB we gain performance with the increase of dimensionality in p, however for MPQLE when p increases and gets closer to the size of n the quality of estimators starts to get worse. We will further investigate the impact on all 3 approaches when the size of p is large compared to n with another simulation shown below.

MCQLEB does not perform as well as MCQLEA when the dimension p of $\mathbf{X}_t$ is small, but in the cases where p is large there are no significant differences in performance. However the number of subset pairings of MCQLEA is $p(p-1)/2$ while MCQLEB only have $p-1$ pairs. The huge increase of computational cost of MCQLEA is not justified for reasonably large p. In some situations, for example indices derivatives trading, a decision or price quotation could be extremely time sensitive, it may be worth investigating what impact reducing the number of pairings even further could have. We saw similar results in example 1 in terms of the computational cost benefits of not choosing subsets efficiently and not excessively.

Table 3: We perform 500 replications where each sub-table has a fixed number of nuisance parameters p, with the number of observations T ranging from 100 to 1000.

	RMSE					
	MCQLEA		MCQLEB		MPQLE	
T	α	β	α	β	α	β
P = 10						
100	0.0225	0.2219	0.0284	0.3463	0.0224	0.3711
200	0.0091	0.0251	0.0126	0.0440	0.0192	0.1884
500	0.0062	0.0110	0.0070	0.0174	0.0093	0.0198
1,000	0.0027	0.0090	0.0036	0.0093	0.0064	0.0108
p=50						
100	0.0147	0.1042	0.0182	0.1463	0.0333	0.6528
200	0.0054	0.0250	0.0083	0.0263	0.0194	0.0292
500	0.0034	0.0093	0.0059	0.0098	0.0117	0.0194
1,000	0.0026	0.0041	0.0034	0.0054	0.0066	0.0075
p=100						
100	0.0126	0.0548	0.0158	0.1218	0.0591	0.8949
200	0.0037	0.0215	0.0061	0.0227	0.0344	0.0427
500	0.0029	0.0105	0.0037	0.0109	0.0283	0.0237
1,000	0.0023	0.0032	0.0027	0.0039	0.0118	0.0160

We perform further simulations to cross examine how the estimation methods compare with each other when p is large in relation to n. We will perform 500 replications for each simulation. We use the values $\alpha = 0.03$ and $\beta = 0.95$, $p = 10, 50, 100$ and $t = 100, 200, 500, 1000$

The results above show that the increase in n improves our estimators for all 3 approaches as we would expect, since more information from an increase of observations would yield better estimates. For the same n both MCQLE approaches give better estimators as p increases. For MPQLE this is not the case and the increased p has a heavy negative impact on the quality of estimators.

Chapter 4

Proofs

We use the same notation as in chapter 2.

4.1 Proof of Theorem 1

The basic idea in the proof of Theorem 1 is the same as that of Theorem 6.5.1 of Lehmann and Casella (1998), although it becomes technically much more involved in order to handle the increasing number of parameters as $n \to \infty$.

Let

$$Q_\delta = \big\{(\boldsymbol{\theta}, \boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r) \;\big|\; ||\boldsymbol{\theta} - \boldsymbol{\theta}_o||^2 + \frac{1}{r}\sum_{j=1}^{r} ||\boldsymbol{\omega}_j - \boldsymbol{\omega}_{jo}||^2 = \delta^2/m\big\}.$$

We will show that for any $\delta > 0$ fixed, $l(\boldsymbol{\beta}) < l(\boldsymbol{\beta}_o)$, for all $\boldsymbol{\beta} \in Q_\delta$, with probability converging to 1. Therefore with probability arbitrarily close to 1 $l(\boldsymbol{\beta})$ attains a local maximum in the interior of Q_m for all sufficiently large n. Let $\widehat{\boldsymbol{\beta}}$ be the local maximum closest to $\boldsymbol{\beta}_0$. By the above argument, $\widehat{\boldsymbol{\beta}}$ must lie in the interior of Q_δ for any $\delta > 0$. This entails the required assertion.

To establish the needed fact concerning the behaviour of $l(\boldsymbol{\beta})$ on Q_δ, we evoke a Taylor expansion:

$$\begin{aligned} \frac{1}{nr}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_o)\} &= \frac{1}{nr}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\dot{l}(\boldsymbol{\beta}_o) + \frac{1}{2nr}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\ddot{l}(\boldsymbol{\beta}_o)(\boldsymbol{\beta} - \boldsymbol{\beta}_o) \\ + \frac{1}{6nr}\sum_{\ell,i,k}(\beta_\ell - \beta_{\ell o})(\beta_i - \beta_{io})(\beta_k - \beta_{ko})&\frac{\partial^3}{\partial\beta_\ell\partial\beta_i\partial\beta_k} l(\boldsymbol{\beta}^\star) \equiv S_1 + S_2 + S_3, \end{aligned} \tag{4.1.1}$$

where $\boldsymbol{\beta}^\star$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_o$.

For $\boldsymbol{\beta} \in Q_\delta$, write $\boldsymbol{\theta} - \boldsymbol{\theta}_o = \frac{\delta}{\sqrt{m}}\boldsymbol{\gamma}$ and $\boldsymbol{\omega}_j - \boldsymbol{\omega}_{jo} = \delta\sqrt{\frac{r}{m}}\boldsymbol{\gamma}_j$. Then all the elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_j$ are between -1 and 1. Furthermore,

$$S_1 = \frac{\delta\boldsymbol{\gamma}'}{n\sqrt{m}}\sum_{t=1}^{n}\frac{1}{r}\sum_{j=1}^{r}\mathbf{a}_{tj} + \frac{\delta\sqrt{r}}{n\sqrt{m}}\sum_{t=1}^{n}\frac{1}{r}\sum_{j=1}^{r}\boldsymbol{\gamma}_j'\mathbf{b}_{tj}. \tag{4.1.2}$$

Let ξ_{tj} denote any component of $\mathbf{a}_{tj}$. Since $E(\sum_j \mathbf{a}_{tj}) = 0$, it holds for any $\epsilon > 0$ that

$$\begin{aligned} P(\frac{\sqrt{m}}{n}\sum_{t=1}^{n}|\frac{1}{r}\sum_{j=1}^{r}\xi_{tj}| > \epsilon) &\leq \frac{m}{n\epsilon^2}\{\mathrm{Var}(\zeta_{tr}) + 2\sum_{t=1}^{n-1}(1-\frac{t}{n})\mathrm{Cov}(\zeta_{1r},\zeta_{1+t,r})\} \\ &\leq \frac{m}{n\epsilon^2}\{\mathrm{Var}(\zeta_{tr}) + 2E(|\zeta_{tr}|^{\nu})^{2/\nu}\sum_{t=1}^{\infty}\alpha(t)^{1-2/\nu}\} \to 0, \end{aligned} \tag{4.1.3}$$

where $\zeta_{tr} = r^{-1}\sum_{1\leq j\leq r}\xi_{tj}$. The last inequality follows from Proposition 2.5 of Fan and Yao (2003); see also conditions A1 and A3. Hence the first sum on the RHS of (4.1.2) is of the order $o_P(m^{-1})$, and the convergence is uniform for $\boldsymbol{\gamma}$ in any compact subset of $\mathcal{R}^d$.

To estimate the second term on the RHS of (4.1.2), let d_j denotes the length of $\mathbf{b}_{tj} \equiv (b_{tj1},\cdots,b_{tjd_j})'$. Then $\max_{1\leq j\leq r} d_j$ are bounded (as $r \to \infty$). Note

$$\sup_{\{\boldsymbol{\gamma}_j\}}|\sum_{t=1}^{n}\sum_{j=1}^{r}\boldsymbol{\gamma}_j'\mathbf{b}_{tj}| = \sup_{\{\boldsymbol{\gamma}_j\}}|\sum_{j=1}^{r}\boldsymbol{\gamma}_j'\sum_{t=1}^{n}\mathbf{b}_{tj}| \leq \sum_{j=1}^{r}\sum_{i=1}^{d_j}|\sum_{t=1}^{n}b_{tji}|.$$

Hence

$$\begin{aligned} P\{\sup_{\{\boldsymbol{\gamma}_j\}}\frac{\sqrt{rm}}{n}|\sum_{t=1}^{n}\frac{1}{r}\sum_{j=1}^{r}\boldsymbol{\gamma}_j'\mathbf{b}_{tj}| > \epsilon\} &\leq P\{\frac{\sqrt{rm}}{n}\sum_{j=1}^{r}\sum_{i=1}^{d_j}|\sum_{t=1}^{n}b_{tji}| > \epsilon r\} \\ \leq \sum_{j=1}^{r}P\{\frac{\sqrt{rm}}{n}\sum_{i=1}^{d_j}|\sum_{t=1}^{n}b_{tji}| > \epsilon\} &\leq \sum_{j=1}^{r}\sum_{i=1}^{d_j}P\{\frac{\sqrt{rm}}{n}|\sum_{t=1}^{n}b_{tji}| > \epsilon/d_j\} \\ \leq \frac{rm(\max_j d_j)^2}{n\epsilon^2}\sum_{j=1}^{r}\sum_{i=1}^{d_j}&\{\mathrm{Var}(b_{tji}) + 2(E|b_{tji}|^{\nu})^{2/\nu}\sum_{t=1}^{\infty}\alpha(t)^{1-2/\nu}\} \to 0, \end{aligned} \tag{4.1.4}$$

as $r^2m/n \to 0$ and condition A3 stands. The last inequality in the above expression follows the same argument as for (4.1.3). This shows that the second sum on the RHS of (4.1.2) is also $o_P(m^{-1})$. Therefore $S_1 = o_P(m^{-1})$, and the convergence is uniform for $\boldsymbol{\beta} \in Q_\delta$.

To calculate S_2, we first note that similar to (4.1.4), condition A4 implies that

$$\begin{aligned} &\frac{1}{nr}\sum_{t=1}^{n}\sum_{j=1}^{r}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'\mathbf{A}_{tj}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)-\frac{1}{r}\sum_{j=1}^{r}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'E(\mathbf{A}_{1j})(\boldsymbol{\theta}-\boldsymbol{\theta}_o)\\ = &\frac{1}{n}\sum_{t=1}^{n}\frac{1}{r}\sum_{j=1}^{r}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'(\mathbf{A}_{tj}-E\mathbf{A}_{tj})(\boldsymbol{\theta}-\boldsymbol{\theta}_o)=o_P(m^{-1}),\end{aligned}$$

$$\frac{1}{nr}\sum_{t=1}^{n}\sum_{j=1}^{r}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'\mathbf{B}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})-\frac{1}{r}\sum_{j=1}^{r}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'E(\mathbf{B}_{1j})(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})=o_P(m^{-1}),$$

$$\frac{1}{nr}\sum_{t=1}^{n}\sum_{j=1}^{r}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})'\mathbf{C}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})-\frac{1}{r}\sum_{j=1}^{r}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})'E(\mathbf{C}_{1j})(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{tj})=o_P(m^{-1}).$$

Furthermore, all the convergences above are uniform for $\boldsymbol{\beta}\in Q_\delta$, as the sizes of all the matrices on the LHS of in the above expressions are fixed, and the the uniform convergence may be established in the same manner as in (4.1.4). Now

$$\begin{aligned} S_2 &= \frac{1}{2nr}\sum_{t=1}^{n}\sum_{j=1}^{r}\{(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'\mathbf{A}_{tj}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)+2(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'\mathbf{B}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})+(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})'\mathbf{C}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})\}\\ &= \frac{1+o_P(1)}{2r}\sum_{j=1}^{r}\{(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'E\mathbf{A}_{tj}(\boldsymbol{\theta}-\boldsymbol{\theta}_o)+2(\boldsymbol{\theta}-\boldsymbol{\theta}_o)'E\mathbf{B}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})+(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})'E\mathbf{C}_{tj}(\boldsymbol{\omega}_j-\boldsymbol{\omega}_{jo})\}\\ &= -\frac{1}{2r}(\boldsymbol{\beta}-\boldsymbol{\beta}_o)'\mathbf{M}_1(\boldsymbol{\beta}-\boldsymbol{\beta}_o)\{1+o_P(1)\}=-\frac{1}{2}\boldsymbol{\beta}_r'\mathbf{M}_2\boldsymbol{\beta}_r\{1+o_P(1)\},\end{aligned}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$ are defined in (2.1.6) and (2.1.7), and

$$\boldsymbol{\beta}_r=((\boldsymbol{\theta}-\boldsymbol{\theta}_o)',(\boldsymbol{\omega}_1-\boldsymbol{\omega}_{1o})'/\sqrt{r},\cdots,(\boldsymbol{\omega}_r-\boldsymbol{\omega}_{ro})'/\sqrt{r})'.$$

For $\boldsymbol{\beta}\in Q_\delta$, $||\boldsymbol{\beta}_r||^2=\delta^2/m$. Since all the eigenvalues of $\mathbf{M}_2$ are bounded between 0 and ∞ (see condition A5), $\boldsymbol{\beta}_r'\mathbf{M}_2\boldsymbol{\beta}_r=2c||\boldsymbol{\beta}_r||^2=2c\delta^2/m$, where $c>0$ is a constant. Hence $S_2=-c\delta^2/m\{1+o_P(1)\}$ uniformly for all $\boldsymbol{\beta}\in Q_\delta$.

Finally we deal with S_3. Note that $\frac{\partial^2}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j'} l(\boldsymbol{\beta}) = 0$ for any $i \neq j$. Similar to the above, it may be proved using condition A6 that

$$\begin{aligned} |S_3| &\leq \frac{1+o_P(1)}{6r}\Big(\big|\sum_{\ell,i,k}(\theta_\ell - \theta_{\ell o})(\theta_i - \theta_{io})(\theta_k - \theta_{ko})\big| \sum_{j=1}^{r} E\{\lambda_j(\mathbf{X}_{tj})\} \\ &+ \big|\sum_{i,k}(\theta_i - \theta_{io})(\theta_k - \theta_{ko})\big| \sum_{j=1}^{r} \big|\sum_{\ell}(\omega_{j\ell} - \omega_{j\ell o})\big| E\{\lambda_j(\mathbf{X}_{tj})\} \\ &+ \big|\sum_{k}(\theta_k - \theta_{ko})\big| \sum_{j=1}^{r} \big|\sum_{\ell,i}(\omega_{j\ell} - \omega_{j\ell o})(\omega_{ji} - \omega_{jio})\big| E\{\lambda_j(\mathbf{X}_{tj})\} \\ &+ \sum_{j=1}^{r} \big|\sum_{\ell,i,k}(\omega_{j\ell} - \omega_{j\ell o})(\omega_{ji} - \omega_{jio})(\omega_{jk} - \omega_{jko})\big| E\{\lambda_j(\mathbf{X}_{tj})\}\Big) \\ &\equiv (S_{31} + S_{32} + S_{33} + S_{34})\{1 + o_P(1)\}. \end{aligned}$$

Note that $E\{\lambda_j(\mathbf{X}_{tj})\}$ is bounded by a constant for $1 \leq j \leq r$, $|\theta_i - \theta_{io}| \leq \delta/\sqrt{m}$ and $|\omega_{jk} - \omega_{jko}| \leq \delta\sqrt{r/m}$ for all $\boldsymbol{\beta} \in Q_\delta$, and all the lengths of $\boldsymbol{\omega}_j$ are bounded. It is easy to see $S_{31} = O(m^{-3/2}) = o(m^{-1})$ and $S_{32} = O(m^{-3/2} r^{1/2}) = o(m^{-1})$. On the other hand,

$$\begin{aligned} S_{33} &\leq \frac{c_2}{r\sqrt{m}} \sum_{j=1}^{r} \big|\sum_{\ell,i}(\omega_{j\ell} - \omega_{j\ell o})(\omega_{ji} - \omega_{jio})\big| \;=\; \frac{c_2}{r\sqrt{m}} \sum_{j=1}^{r} \big|\sum_{i}(\omega_{ji} - \omega_{jio})\big|^2 \\ &\leq \frac{c_3}{r\sqrt{m}} \sum_{j=1}^{r} ||\boldsymbol{\omega}_j - \boldsymbol{\omega}_{jo}||^2 \;\leq\; \frac{c_3}{m^{3/2}} \;=\; o(m^{-1}), \end{aligned}$$

$$S_{34} \leq \frac{c_4}{\sqrt{mr}} \sum_{j=1}^{r} \big|\sum_{i}(\omega_{ji} - \omega_{jio})\big|^2 \leq \frac{c_5 r^{1/2}}{m^{3/2}} = o(m^{-1}).$$

This concludes that $S_3 = o_P(m^{-1})$.

Combining the above asymptotic approximations for S_1, S_2 and S_3 together, we have shown that uniformly for $\boldsymbol{\beta} \in Q_\delta$

$$\frac{1}{nr}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_o)\} = -c\,\delta^2/m + o_P(m^{-1}),$$

where $c > 0$ is a constant. This completes the proof.

4.2 Proof of Theorem 2

Since $\dot{l}(\widehat{\boldsymbol{\beta}}) = 0$, it follows from a simple Taylor expansion that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o = -\{\ddot{l}(\boldsymbol{\beta}^\star)\}^{-1}\dot{l}(\boldsymbol{\beta}_o), \tag{4.2.5}$$

where $\ddot{l} = \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$, and $\boldsymbol{\beta}^\star$ lies on the line between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_o$. Note

$$\ddot{l}(\boldsymbol{\beta}) = \sum_{t=1}^{n} \begin{pmatrix} \sum_{j=1}^{r} \mathbf{A}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) & \mathbf{B}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_1) & \cdots & \mathbf{B}_{tr}(\boldsymbol{\theta}, \boldsymbol{\omega}_r) \\ \mathbf{B}_{t1}(\boldsymbol{\theta}, \boldsymbol{\omega}_1)' & \mathbf{C}_{t1}(\boldsymbol{\theta}, \boldsymbol{\omega}_1) & & \\ \vdots & & \ddots & \\ \mathbf{B}_{tr}(\boldsymbol{\theta}, \boldsymbol{\omega}_r)' & & & \mathbf{C}_{tr}(\boldsymbol{\theta}, \boldsymbol{\omega}_r) \end{pmatrix},$$

where the entries at the blank places are all 0. We partition the above matrix into 2×2 blocks with $\sum_t \sum_j \mathbf{A}_{tj}(\boldsymbol{\theta}, \boldsymbol{\omega}_j)$ as the $(1,1)$-th block. By inverting this partitioned matrix, the first d components of (4.2.5) may now be expressed as

$$\begin{aligned} & \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \\ = \; & -\Big\{ \frac{1}{nr} \sum_{j=1}^{r} \Big(\sum_{t=1}^{n} \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\{ \sum_{t=1}^{n} \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\}^{-1} \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)' \Big) \Big\}^{-1} \\ \times \; & \frac{1}{\sqrt{n}\, r} \sum_{j=1}^{r} \Big(\sum_{t=1}^{n} \mathbf{a}_{tj} - \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\{ \sum_{t=1}^{n} \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\}^{-1} \sum_{t=1}^{n} \mathbf{b}_{tj} \Big). \end{aligned} \tag{4.2.6}$$

For any matrix $\mathbf{B}$, denote by $|\mathbf{B}|_a$ the sum of the absolute values of all the elements of $\mathbf{B}$. Note that all the sizes of the matrices $\mathbf{A}_{tj}$, $\mathbf{B}_{tj}$ and $\mathbf{C}_{tj}$ are bounded. It follows from condition A6 that

$$\begin{aligned} & \max_{1 \le j \le r} \Big| \frac{1}{n} \sum_{t=1}^{n} \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{A}_{1j}) \Big|_a \qquad (4.2.7) \\ \le \; & \max_{1 \le j \le r} \frac{1}{n} \Big| \sum_{t=1}^{n} \{ \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - \mathbf{A}_{tj} \} \Big|_a + \max_{1 \le j \le r} \Big| \frac{1}{n} \sum_{t=1}^{n} \mathbf{A}_{tj} - E(\mathbf{A}_{1j}) \Big|_a \\ \le \; & \{ |\boldsymbol{\theta}^\star - \boldsymbol{\theta}_o|_a + \max_{1 \le j \le r} |\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a \} \max_{1 \le j \le r} \frac{1}{n} \sum_{t=1}^{n} \lambda_j(\mathbf{X}_{tj}) + \max_{1 \le j \le r} \Big| \frac{1}{n} \sum_{t=1}^{n} \mathbf{A}_{tj} - E(\mathbf{A}_{1j}) \Big|_a. \end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned} & P\Big\{ \max_{1 \le j \le r} \Big| \frac{1}{n} \sum_{t=1}^{n} \mathbf{A}_{tj} - E(\mathbf{A}_{1j}) \Big|_a > \epsilon \Big\} \; \le \; \sum_{j=1}^{r} P\Big\{ \Big| \frac{1}{n} \sum_{t=1}^{n} \mathbf{A}_{tj} - E(\mathbf{A}_{1j}) \Big|_a > \epsilon \Big\} \qquad (4.2.8) \\ \le \; & \frac{c}{n} \sum_{\eta_{tj}} \sum_{j=1}^{r} \Big[\mathrm{Var}(\eta_{tj}) + 2\{ E(|\eta_{tj}|^\nu) \}^{2/\nu} \sum_{k=1}^{\infty} \alpha(k)^{1-2/\nu} \Big] \; \to \; 0. \end{aligned}$$

The limit above is guaranteed by condition A4 and the fact that $r/n \to 0$. In the above expression, η_{tj} denotes a generic element of $\mathbf{A}_{tj}$, and the sum $\sum_{\eta_{tj}}$ is taken over all the elements of $\mathbf{A}_{tj}$. The last inequality follows the same argument as in (4.1.3). In the same way we may show that $\max_j |\frac{1}{n}\sum_{t=1}^n[\lambda_j(\mathbf{X}_{tj}) - E\{\lambda_j(\mathbf{X}_{tj})\}]| \xrightarrow{P} 0$, and therefore

$$\max_{1\le j\le r} \frac{1}{n}\sum_{t=1}^n \lambda_j(\mathbf{X}_{tj}) = O_P(1). \tag{4.2.9}$$

Now we show that

$$\max_{1\le j\le r} |\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a \xrightarrow{P} 0. \tag{4.2.10}$$

It follows from (2.1.10) that for any $\epsilon > 0$, it holds for all sufficiently large n that

$$P\{\sum_{j=1}^r ||\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}||^2 \le \epsilon^2/k_0^2\} > 1-\epsilon,$$

where k_0 is the maximum length of the vectors $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r$, which is fixed. Since $\boldsymbol{\omega}_j^\star$ lies between $\widehat{\boldsymbol{\omega}}_j$ and $\boldsymbol{\omega}_{jo}$, $|\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a \le |\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}|_a$. Hence

$$\begin{aligned} &P\{\max_{1\le j\le r} |\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a \le \epsilon\} \;\ge\; P\{\max_{1\le j\le r} |\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}|_a \le \epsilon\} \\ &\ge\; P\{\sum_{j=1}^r ||\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_{jo}||^2 \le \epsilon^2/k_0^2\} > 1-\epsilon. \end{aligned}$$

Therefore (4.2.10) holds. Combining (4.2.7) – (4.2.10), we conclude

$$\max_{1\le j\le r} \Big|\frac{1}{n}\sum_{t=1}^n \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{A}_{1j})\Big|_a \xrightarrow{P} 0. \tag{4.2.11}$$

It may be established in the same manner that

$$\max_{1\le j\le r} \Big|\frac{1}{n}\sum_{t=1}^n \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{B}_{1j})\Big|_a \xrightarrow{P} 0, \quad \max_{1\le j\le r} \Big|\frac{1}{n}\sum_{t=1}^n \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{C}_{1j})\Big|_a \xrightarrow{P} 0,$$

which implies that

$$\max_{1\le j\le r} \Big|\frac{1}{n}\sum_{t=1}^n \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)\{\sum_{t=1}^n \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)\}^{-1}\sum_{t=1}^n \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)' - E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}E(\mathbf{B}_{1j}')\Big|_a \xrightarrow{P} 0.$$

Combining this with (4.2.11), we obtain that

$$\begin{aligned} &\frac{1}{nr}\sum_{j=1}^r \Big(\sum_{t=1}^n \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - \sum_{t=1}^n \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)\Big\{\sum_{t=1}^n \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)\Big\}^{-1}\sum_{t=1}^n \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)'\Big) \\ &= \frac{1}{r}\sum_{j=1}^r \{E(\mathbf{A}_{1j}) - E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}E(\mathbf{B}_{1j}')\} + o_P(1) \to \mathbf{L}. \end{aligned}$$

Using the similar arguments, we may show that

$$\frac{1}{\sqrt{n}\,r}\sum_{j=1}^{r}\sum_{t=1}^{n}\mathbf{B}_{tj}(\boldsymbol{\theta}^\star,\boldsymbol{\omega}_j^\star)\{\sum_{t=1}^{n}\mathbf{C}_{tj}(\boldsymbol{\theta}^\star,\boldsymbol{\omega}_j^\star)\}^{-1}\sum_{t=1}^{n}\mathbf{b}_{tj}-\frac{1}{\sqrt{n}\,r}\sum_{j=1}^{r}E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}\sum_{t=1}^{n}\mathbf{b}_{tj}\xrightarrow{P}0.$$

Now it follows from (4.2.6) that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)=\mathbf{L}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{1}{r}\sum_{j=1}^{y}\{\mathbf{a}_{tj}-E(\mathbf{B}_{1j})(E\mathbf{C}_{1j})^{-1}\mathbf{b}_{tj}\}\{1+o_P(1)\}.$$

The required asymptotic normality follows from Proposition 2 in the Appendix now; see condition A7. This concludes the proof.

4.3 Proof of Theorem 3

Using the notation in section 3, we have

$$\begin{aligned}\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}_o)-\frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o) &= \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\{\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\widehat{\boldsymbol{\omega}})-\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\} \quad (4.3.12)\\
&= \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)+\frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)'\frac{\partial^2 a_1(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}^\star)}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)\\ \vdots\\ (\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)'\frac{\partial^2 a_d(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}^\star)}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)\end{pmatrix}\\
&= \frac{1}{n^{3/2}m}\sum_{t,s=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s)+O_P(\frac{q}{m\sqrt{n}}),\end{aligned}$$

where $\boldsymbol{\omega}^\star$ is between $\widehat{\boldsymbol{\omega}}$ and $\boldsymbol{\omega}_o$, and $\mathbf{g}=(g_1,\cdots,g_q)'$. The last equality in the above expression follows from conditions B1 and B2. Note that

$$\begin{aligned}\sum_{t,s=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s) &= 2\sum_{1\le t<s\le n}\{\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s) \quad (4.3.13)\\
&+ \mathbf{C}(\mathbf{X}_s;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)\} + \sum_{t=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t).\end{aligned}$$

By applying the Hoeffding decompostion (A.1) (with $m=2$) to the first sum on the RHS of (4.3.13), it follows from (4.3.12) and (4.3.13) that

$$\begin{aligned}\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}) &= \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)+\frac{2(n-1)}{n^{3/2}m}\sum_{t=1}^{n}\mathbf{D}(\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t) \quad (4.3.14)\\
&+ L_n+\frac{1}{n^{3/2}m}\sum_{t=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)+O_P(\frac{q}{m\sqrt{n}}),\end{aligned}$$

where

$$L_n = \frac{2}{n^{3/2}m} \sum_{1\le t<s\le n} \big[\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s)+\mathbf{C}(\mathbf{X}_s;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)-\mathbf{D}(\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\{\mathbf{g}(\mathbf{X}_t)+\mathbf{g}(\mathbf{X}_s)\}\big].$$

By Proposition 1 in the Appendix, $E\{(n^{-1/2}L_n)^2\} = O(n^{-1-\gamma})$. Hence it holds for any constant c,

$$P(|L_n| \ge c) = P\big\{n(n^{-1/2}L_n)^2 > c\big\} = n \cdot O(n^{-1-\gamma}) = O(n^{-\gamma}) \to 0;$$

see condition B3. We may also show in the similar (but simpler) manner that

$$\frac{1}{n^{3/2}m}\sum_{t=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t) = O_P(n^{-1/2}).$$

Therefore it follows from (4.3.14) that

$$\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}_o) = \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\{\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o) + 2\mathbf{D}(\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)\} + o_P(1).$$

Note conditions B4 and B3 imply conditions C3 and C4. By Proposition 2,

$$\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}_o) \xrightarrow{D} N(0,\ \boldsymbol{\Sigma}_0 + 2\sum_{j=1}^{\infty}\boldsymbol{\Sigma}_j). \tag{4.3.15}$$

Furthermore, the convergence of the sum $\sum_{j\ge1}\boldsymbol{\Sigma}_j$ is guaranteed by condition B4.

On the other hand,

$$\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^\star) = \frac{1}{nm}\sum_{t=1}^{n}\mathbf{B}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o) + \frac{1}{nm}\sum_{t=1}^{n}\mathbf{G}(\mathbf{X}_t;\boldsymbol{\theta}^{\star\star},\boldsymbol{\omega}^\star,\boldsymbol{\theta}^\star-\boldsymbol{\theta}_o,\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o), \tag{4.3.16}$$

where $(\boldsymbol{\theta}^{\star\star},\boldsymbol{\omega}^\star)$ lies between $(\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\omega}})$ and $(\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)$, and $\mathbf{G}$ is a $d\times d$ matrix with the (i,j)-th element

$$(\boldsymbol{\theta}^\star-\boldsymbol{\theta}_o)'\frac{\partial}{\partial\boldsymbol{\theta}}b_{ij}(\mathbf{X}_t;\boldsymbol{\theta}^{\star\star},\boldsymbol{\omega}^\star) + (\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)'\frac{\partial}{\partial\boldsymbol{\omega}}b_{ij}(\mathbf{X}_t;\boldsymbol{\theta}^{\star\star},\boldsymbol{\omega}^\star), \tag{4.3.17}$$

and b_{ij} denotes the (i,j)-th element of $\mathbf{B}$. Write $\mu_{ij,m} = E\{b_{ij}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\}/m$. Then for any $\epsilon > 0$,

$$P\{\big|\frac{1}{nm}\sum_{t=1}^{n}b_{ij}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o) - \mu_{ij,m}(\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\big| > \epsilon\} \le \frac{1}{\epsilon^2 n^2}\mathrm{Var}\{\frac{1}{m}\sum_{t=1}^{n}b_{ij}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\} \to 0.$$

The limit is guaranteed by B5 and the mixing condition on $\mathbf{X}_t$; see Proposition 2.5 of Fan and Yao (2003). Hence

$$\frac{1}{nm}\sum_{t=1}^{n}\mathbf{B}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\xrightarrow{P}\mathbf{M},$$

where $\mathbf{M}$ is a $d\times d$ matrix with the limit of $\mu_{ij,m}$ as its (i,j)-th element. Note that the absolute value of the expression in (4.3.17) is bounded from the above by

$$\lambda_2(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\{||\boldsymbol{\theta}^\star-\boldsymbol{\theta}_o||+||\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o||\}.$$

Condition B5 implies that there exists a positive and finite constant c for which

$$P\Big\{\ \frac{1}{nm}\sum_{t=1}^{n}\lambda_2(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\le c\ \Big\}\to 1.$$

Since $||\boldsymbol{\theta}^\star-\boldsymbol{\theta}_o||+||\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o||\xrightarrow{P}0$, the second term on the RHS of (4.3.16) converges to 0 in probability. Therefore $\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^\star)\xrightarrow{P}\mathbf{M}$. This, together with (4.3.15), concludes the theorem.

Chapter 5

Appendix: U-statistics

Let $\boldsymbol{\xi}_t$ be a $p \times 1$ strictly stationary process, $\boldsymbol{\xi}_t$ is $\mathcal{F}_t$-measurable, and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ is a sequence of σ-algebra. Let $\psi_n(\mathbf{x}_1, \cdots, \mathbf{x}_m)$ be a real-valued function defined on $(\Re^p)^m$, and it is symmetric in its $m(\geq 2)$ arguments. A U-statistic based on n observations $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n$ is defined as

$$U_n = \frac{m!(n-m)!}{n!} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \psi_n(\boldsymbol{\xi}_{i_1}, \cdots, \boldsymbol{\xi}_{i_m}).$$

For $k = 1, \cdots, m-1$, let

$$\psi_{n,k}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \int \psi_n(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}, \cdots, \mathbf{x}_m) \prod_{j=k+1}^{n} F(d\mathbf{x}_j),$$

where $F(\cdot)$ denotes the marginal distribution of $\boldsymbol{\xi}_t$. For the simplicity in presentation, we assume that $E\{\psi_{n,1}(\boldsymbol{\xi}_t)\} = 0$. (Otherwise we replace ψ_n by $\psi_n - E\{\psi_{n,1}(\boldsymbol{\xi}_t)\}$.) Put

$$\begin{aligned}
h_{n,1}(\mathbf{x}_1) &= \psi_{n,1}(\mathbf{x}_1), \\
h_{n,2}(\mathbf{x}_1, \mathbf{x}_2) &= \psi_{n,2}(\mathbf{x}_1, \mathbf{x}_2) - h_{n,1}(\mathbf{x}_1) - h_{n,1}(\mathbf{x}_2), \\
h_{n,3}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) &= \psi_{n,3}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) - \sum_{j=1}^{3} h_{n,1}(\mathbf{x}_j) - \sum_{1 \leq i < j \leq 3} h_{n,2}(\mathbf{x}_i, \mathbf{x}_j), \\
& \cdots\cdots \\
h_{n,m}(\mathbf{x}_1, \cdots, \mathbf{x}_k) &= \psi_n(\mathbf{x}_1, \cdots, \mathbf{x}_k) - \sum_{j=1}^{m} h_{n,1}(\mathbf{x}_j) - \sum_{1 \leq i < j \leq m} h_{n,2}(\mathbf{x}_i, \mathbf{x}_j) - \cdots \\
& \quad - \sum_{1 \leq i_1 < \cdots i_{m-1} \leq m} h_{n,m-1}(\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_k}).
\end{aligned}$$

The Hoeffding decomposition (Lemma A, pp. 178 in Serfling 1980) is of the form

$$U_n = \frac{m}{n}\sum_{j=1}^{n}\psi_{n,1}(\boldsymbol{\xi}_j) + \sum_{k=2}^{m}\frac{m!}{(m-k)!}S_{n,k}, \tag{A.1}$$

where

$$S_{n,k} = \frac{(n-k)!}{n!}\sum_{1\le i_1<\cdots<i_k\le n} h_{n,k}(\boldsymbol{\xi}_{i_1},\cdots,\boldsymbol{\xi}_{i_k}). \tag{A.2}$$

As long as the variance of $\psi_{n,1}(\boldsymbol{\xi}_j)$ does not diminish to 0, the asymptotic property of U_n is determined by that of the first sum on the RHS of (A.1). The lemma below shows indeed that the remainder term (i.e. the other sum) is asymptotically negligible. Different from conventional setting, we allow the kernel function ψ_n to vary with respect to the sample size n. Furthermore, we allow the dimension p of $\boldsymbol{\xi}_j$ to diverge to ∞ together with n. We first introduce some regularity conditions.

C1. $\{\boldsymbol{\xi}_t\}$ is a strictly stationary and β-mixing (i.e. absolutely regular) process with the β-mixing coefficients satisfying the condition $\beta(n) = O(n^{-(2+\delta')/\delta'})$, where $\delta' \in (0,\delta)$ is a constant.

C2. It holds for all n, p and $1 \le i_1 < \cdots < i_m \le n$ that $E\{|\psi_n(\boldsymbol{\xi}_{i_1},\cdots,\boldsymbol{\xi}_{i_m})|^{2+\delta}\} \le M$, and

$$\int \left|\psi_n(\mathbf{x}_1,\cdots,\mathbf{x}_m)\right|^{2+\delta}\prod_{j=1}^{m}F(d\mathbf{x}_j) \le M,$$

where $\delta > 0,\ M > 0$ are fixed constants.

Proposition 1. Under conditions C1 and C2, it holds that $E(S_{n,k}^2) = O(n^{-1-\gamma})$ for $k = 2,\cdots,m$, where $S_{n,k}$ is defined as in (A.2) and $\gamma = \min\{1,\ \frac{2(\delta-\delta')}{\delta'(2+\delta)}\}$.

5.1 Proposition 1

Proposition 1 is essentially Lemma 2 of Yoshihara (1976). The only difference here is to allow ψ_n to vary with n and the dimension p to grow. Nevertheless the original proof is still applicable. However it was an error to define $\gamma = \frac{2(\delta-\delta')}{\delta'(2+\delta)}$ in Yoshihara (1976), as the optimal rate for $E(S_{n,k}^2)$ is n^{-2}. Therefore it must hold that $\gamma \le 1$. Note that this optimal rate is attainable when, for example, $\{\boldsymbol{\xi}_t\}$ is a sequence of independent r.v.s, or the rate of the mixing coefficients is strengthened to satisfy the condition

$$\sum_{k=1}^{\infty} k\beta(k)^{\delta/(2+\delta)} < \infty.$$

Now we turn to the asymptotic normality of the first term on the RHS of (A.1). We state the required regularity conditions separately below.

C3. $\{\boldsymbol{\xi}_t\}$ is a strictly stationary and α-mixing (i.e. strong mixing) process with α-mixing coefficients satisfying the condition $\sum_{k\geq 1}\alpha(k)^{1-2/\nu}<\infty$, where $\nu>2$ is a constant.

C4. For $\nu>2$ given in C3 above, $\overline{\lim}_{n\to\infty} E\{|\psi_{n,1}(\boldsymbol{\xi}_1)|^\nu\}<\infty$. Furthermore, the limit of $\mathrm{Cov}\{\psi_{n,1}(\boldsymbol{\xi}_1),\ \psi_{n,1}(\boldsymbol{\xi}_j)\}$ exists for any $1\leq j\leq n$.

Put

$$B_n^2 \;=\; \frac{1}{n}\mathrm{Var}\Big\{\sum_{t=1}^{n}\psi_{n,1}(\boldsymbol{\xi}_t)\Big\} \;=\; \mathrm{Var}\{\psi_{n,1}(\boldsymbol{\xi}_1)\}+2\sum_{j=1}^{n-1}\big(1-\frac{j}{n}\big)\mathrm{Cov}\{\psi_{n,1}(\boldsymbol{\xi}_1),\ \psi_{n,1}(\boldsymbol{\xi}_{1+j})\}.$$

5.2 Proposition 2

Proposition 2. Under conditions C3 and C4, it holds that

$$\frac{1}{\sqrt{n}B_n}\sum_{t=1}^{n}\psi_{n,1}(\boldsymbol{\xi}_t)\xrightarrow{D} N(0,\,1).$$

5.2.1 Proof

Proof. By Proposition 2.5 of Fan and Yao (2003) with $p=q=\nu$,

$$|\mathrm{Cov}\{\psi_{n,1}(\boldsymbol{\xi}_1),\ \psi_{n,1}(\boldsymbol{\xi}_{1+j})\}|\leq 8\alpha(j)^{1-\frac{2}{\nu}}\{E|\psi_{n,1}(\boldsymbol{\xi}_1)|^\nu\}^{2/\nu},$$

see condition C4. Hence it follows from condition C3 that

$$\lim_{n\to\infty}\sum_{j=1}^{n-1}|\mathrm{Cov}\{\psi_{n,1}(\boldsymbol{\xi}_1),\ \psi_{n,1}(\boldsymbol{\xi}_{1+j})\}|\leq 8\,\overline{\lim_{n\to\infty}}\{E|\psi_{n,1}(\boldsymbol{\xi}_1)|^\nu\}^{2/\nu}\sum_{j=1}^{\infty}\alpha(j)^{1-2/\nu}<\infty.$$

Now by the Lebesgue dominated convergence theorem, it holds that

$$\lim_{n\to\infty}B_n^2=\lim_{n\to\infty}\frac{1}{n}\mathrm{Var}\Big\{\sum_{t=1}^{n}\psi_{n,1}(\boldsymbol{\xi}_t)\Big\}=\sigma^2\in(0,\infty), \tag{A.3}$$

where σ^2 is a constant.

Now we partition the set $\{1,\cdots,n\}$ into $2k_n+1$ subsets with large blocks of size l_n, small blocks of size s_n and the last remaining set of size $n-k_n(l_n+s_n)$, where l_n and

s_n are selected such that

$$s_n \to \infty, \quad s_n/l_n \to 0, \quad l_n/n \to 0, \quad \text{and} \quad k_n = [n/(l_n + s_n)] = O(s_n).$$

For example, we may choose $l_n = O(n^{\frac{a-1}{a}})$ and $s_n = O(n^{1/a})$ for any $a > 2$. Then $k_n = O(n^{1/a})$ too. For $j = 1, \cdots, k_n$, define

$$\eta_j = \sum_{i=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \psi_{n,1}(\boldsymbol{\xi}_i), \quad \zeta_j = \sum_{i=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} \psi_{n,1}(\boldsymbol{\xi}_i), \quad \chi = \sum_{i=k_n(l_n+s_n)+1}^{n} \psi_{n,1}(\boldsymbol{\xi}_i).$$

Similar to (A.3), it may be proved that

$$\lim_{n\to\infty} \frac{1}{n} \text{Var}\big(\sum_{j=1}^{k_n} \zeta_j\big) = \lim_{n\to\infty} \frac{k_n s_n}{n} \frac{1}{k_n s_n} \text{Var}\big(\sum_{j=1}^{k_n} \zeta_j\big) = 0,$$

and $n^{-1}\text{Var}(\chi) \to 0$. Hence

$$\frac{1}{\sqrt{n}B_n} \sum_{t=1}^{n} \psi_{n,1}(\boldsymbol{\xi}_t) = \frac{1}{\sqrt{n}B_n} \{ \sum_{j=1}^{k_n} \eta_j + \sum_{j=1}^{k_n} \zeta_j + \chi \} = \frac{1}{\sqrt{n}B_n} \sum_{j=1}^{k_n} \eta_j + o_P(1). \tag{A.4}$$

By Proposition 2.6 of Fan and Yao (2003),

$$\Big| E\big\{ \exp\big(\frac{it}{\sqrt{n}B_n} \sum_{j=1}^{k_n} \eta_j\big)\big\} - \prod_{j=1}^{k_n} E\{\exp\big(\frac{it\eta_j}{\sqrt{n}B_n}\big)\}\Big| \le 16(k_n - 1)\alpha(s_n) \to 0, \tag{A.5}$$

see condition C3. Again similar to (A.3), it holds that $\text{Var}(\sum_{1\le j\le k_n} \eta_j)/B_n \to 1$. It follows from condition C4 that

$$\limsup_n E\big[|\psi_{n,1}(\boldsymbol{\xi}_1)|^2 I\{|\psi_{n,1}(\boldsymbol{\xi}_1)| \ge \varepsilon\sqrt{n}\}\big] \le \frac{1}{\varepsilon^{\nu-2} n^{\nu/2-1}} \lim_n E\{|\psi_{n,1}(\boldsymbol{\xi}_1)|^\nu\} \to 0,$$

for any $\varepsilon > 0$. Noticing (A.3), it follows from the theorem on page 31 of Serfling (1980) that

$$\prod_{j=1}^{k_n} E\{\exp\big(\frac{it\eta_j}{\sqrt{n}B_n}\big)\} \to e^{-t^2/2}.$$

This together with (A.5) and (A.4) entails the required result. □

Chapter 6

References

Baltagi, B.H. (2005). *Econometric Analysis of Panel Data*. Wiley, New York.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Royal Stats. Soc* **B**, **36** , 192-236.

Bickel, P., Ritov Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Danzig selector. *Ann. Statist.* **37** 1705-1732.

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

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

deLeon, A.R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Stats. and Probab. Letters*, **75**, 49-57.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Berkeley Symposium 5*, **1**, 59-82.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Stats. Soc.* **B**, **70**, 849-911.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Engle, R.F., Shephard, N. and Sheppard, K. (2008). Fitting and testing vast dimensional time-varying covariance models. *A preprint*.

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

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603-1618.

Kuk, A. and Nott, D. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Stats. Probab. Letters*, **47**, 329-335.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, **21**, 43-69.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

Liang, K.-Y. (1987). Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics*, **43**, 289-299.

Lindsay, B. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (ed. Brabhu, N.U.). pp.221-239. Providence, RI: American Mathematical Society.

Molenberghs, G. and Vervbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

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

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary, absolutely regular processes. *Z. Wahrsch. verw. Gebiete*, **35**, 237-252.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional regression. *Ann. Statist.* **36**, 1567-1594.

Zhang, T. (2009a). Some sharp performance bounds for least squares regression with L_1 regularisation. *Ann. Statist.* **37**, 2109-2144.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.