Estimation in the Presence of Many Nuisance Parameters: composite likelihood and plug-in likelihood

Billy Wu[⋆] Qiwei Yao[⋆,*] Shiwu Zhu[‡]

[⋆]Department of Statistics, London School of Economics, London, UK

[‡]School of Economics and Management, Tsinghua University, Beijing, China

Email: q.yao@lse.ac.uk

6 March 2013

Abstract

We consider the incidental parameters problem in this paper, i.e. the estimation for a small number of parameters of interest in the presence of a large number nuisance parameters. By assuming that the observations are taken from a multiple strictly stationary process, the two estimation methods, namely the maximum composite quasi-likelihood estimation (MCQLE) and the maximum plug-in quasi-likelihood estimation (MPQLE) are considered. For the MCQLE, we profile out nuisance parameters based on lower-dimensional marginal likelihoods, while the MPQLE is based on some initial estimators for nuisance parameters. The asymptotic normality for both the MCQLE and the MPQLE is established under the assumption that the number of nuisance parameters and the number of observations go to infinity together, and both the estimators for the parameters of interest enjoy the standard root-n convergence rate. Simulation with a spatial-temporal model illustrates the finite sample properties of the two estimation methods.

Key words: Composite likelihood, incidental parameters problem, nuisance parameter, panel data, profile likelihood, quasi-likelihood, root-n convergence, spatial autoregressive model, stationary process, time series, U-statistic.

[*] Partially supported by an EPSRC research grant.

1 Introduction

Rapid developments in technology in this information age have lead to data collection in an unprecedently large scale. This brings new opportunities with challenges to statistics. The availability of large data sets enables 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 small 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 or a panel data 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 (Baltagi 2005, Chapter 2). For a large panel of time series data, one is often interested in a few common factors which drive the dynamics of all the component series and treat the parameters representing each idiosyncratic components as nuisance parameters. In the attempt to model the volatilities of a large number of financial securities, it is often assumed that the dynamic volatilities are controlled by a small number of parameters in the presence of a large number of nuisance parameters for marginal covariance matrices (Engle *et al.* 2008). For a spatio-temporal study focussing on the spatial correlation, the parameters determining the temporal dynamics at each location are treated as nuisance parameters (see, for example, the example in section 4 below).

In this paper we consider two methods of estimating a small number of parameters of interest in the presence of a large number of nuisance parameters, namely the maximum composite quasi-likelihood estimation (MCQLE) and the maximum plug-in quasi-likelihood estimation (MPQLE). The composite likelihood, the name coined by Lindsay (1988), is a function derived by multiplying a collection of, typically two- or three-dimensional, marginal density functions. Its composition is often dictated by, among other things, the computational feasibility. See a recent survey by Varin *et al.* (2011). 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 a high-dimensional optimization problem. Because the marginal densities are multiplied together, ignoring the original distribution structure, the MCQLE can be viewed as derived from a (seriously) misspecified model. On the other hand, the MPQLE maximizes a quasi-likelihood function with a plug-in estimator for the nuisance parameter vector. Therefore we avoid a maximization 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. When the number of nuisance parameters is large, the estimation for all of them collectively is typically poor. A case in point is the estimation for large covariance matrices; see, for example, Figure 1 of Tao *et al.* (2011).

The major contribution of this paper includes the asymptotic properties for both the MCQLE and the MPQLE under the assumption which is relevant to the problem 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 to 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. We adopt the setting under which the observations are taken from a multiple strictly stationary process and the dimension of the process may also go to infinity together with the sample size. The setting is generic and the results are applicable to the relevant inference problems in, for example, multiple time series, panel data and spatio-temporal models. Though bearing a similar banner, our theory is different from the large body of literature on the so-called 'large p and small n' regression problem; see, among others, Fan and Lv (2008), Zhang and Huang (2008), and Bickel *et al.* (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 problems including, for example, regression with dependent errors (Eicker 1967), modeling for 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), bioinformatics (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), Eicker (1967), White (1982), Gallant and White (1988), and Cox and Reid (2004). For more comprehensive survey on the composite likelihood methodology, we refer to the first issue of *Statistica Sinica* (2011) vol.21 which contains a collection of the papers on this topic.

The estimation problem concerned in this paper was initially termed as an incidental parameters problem by Neyman and Scott (1948). See also the survey by Lancaster (2000). Parameters of interest are 'structural' and nuisance parameters are 'incidental' in Neyman and Scott's terminology. (The word 'nuisance' suggests that those parameters are burdensome or even annoying while 'incidental' is much milder. Barndorff-Nielsen (1978, p.33) prefers incidental to nuisance, finding the latter 'somewhat emotional'.) One of the classical examples of the incidental parameters problem concerns the estimation for the common variance parameter σ^2 of $n \times r$ independent and normal observations with n different mean values $\mu_1, \cdots, \mu_n$. Then when $n \to \infty$ but r fixed, the maximum likelihood estimator for the 'structural' parameter σ^2 is not ever consistent due to the inconsistent estimators for the incidental parameters $\mu_1, \cdots, \mu_n$, though a consistent and efficient estimator for σ^2 exists. See Example 7.9 on p.482 of Lehmann and Casella (1998). This is because that the information on each incidental parameter μ_i does not increase when n increases. Neyman and Scott (1948) labels the data in such a scenario as 'partially consistent observations', as one only can estimate structural parameters consistently but not the incidental parameters. More traditional likelihood approaches for incidental parameter problems include, for example, a conditional likelihood method based on a conditional distribution which is free from incidental parameters (Basu 1977, Barndorff-Nielsen 1978), a partial likelihood method based on a statistic of which the density function is free from the incidental parameters (Cox 1975), and a profile likelihood obtained by replacing incidental parameters by their maximum likelihood estimators (Cox and Reid 1987). On the other hand, a Bayesian treatment may integrate the incidental parameters with respect to a prior distribution (Berger *et al.* 1999). See also Reid (1996).

The proposed methods in this paper are designed for complex applications with large and high-dimensional data when incidental-parameter-free conditional or partial likelihoods do not exist, profiling a likelihood directly leads to a high-dimensional optimization problem which is computationally infeasible. On the other hand, the information on incidental parameters also increases in our asymptotic framework in the sense that each of those incidental parameters can be estimated consistently at least in principle. Hence our setting is different from the setting with 'partially consistent observations' in Neyman and Scott's terminology.

The rest of the paper is organized as follows. Section 2 deals with the MCQLE and section 3 is on MPQLE. We outline the estimation methods and state the asymptotic normality results. In section 4 the finite sample properties of both the methods are illustrated in a small scale simulation with a simple spatio-temporal model. It reveals the advantages of using the MCQLE when the number of nuisance parameters is large in relation to the sample size, the phenomenon observed in Engle *et al.* (2008) with a high-dimensional volatility model. All technical proofs are given in sections 5 and

6. An extension on the U-statistic, which plays a key role in establishing the asymptotic normality, is presented in the Appendix.

2 Composite-likelihood estimation

Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be $p \times 1$ observations from a strictly 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 an 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, some of those subvectors may share common components of $\mathbf{X}_t$. 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 parameters denoted by $\boldsymbol{\omega}_j$. Let

$$\widetilde{\boldsymbol{\omega}}_j(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\omega}_j} l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j). \tag{2.1}$$

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

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} l_j\big(\boldsymbol{\theta}, \widetilde{\boldsymbol{\omega}}_j(\boldsymbol{\theta})\big). \tag{2.2}$$

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

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{2.3}$$

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 as in (2.3) being reasonable is that the nuisance parameters $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r$ are distinct from each other such that the maximization (2.1) may be carried out independently for each j without confounding constraints from each other. This is a strong requirement, and may only be facilitated by selecting subvectors $\mathbf{X}_{t1}, \cdots, \mathbf{X}_{tr}$ in a restrictive manner. It may make this approach infeasible or lead to a heavy loss of information. One alternative is to adopt the so-called ‘variation-free’ condition imposed by Engle, Hendry and Richard (1983), which 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 for $\boldsymbol{\theta}$ resulted from neglecting the links among different $\boldsymbol{\omega}_j$. The trade-off is that we will be able to reduce a high-dimensional optimization problem to many low-dimensional problems, which is the essential motivation of using the composite-likelihood approach. Note that this variation-free condition also implies that $\widehat{\boldsymbol{\theta}}$ is the global maximizer 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 $l(\boldsymbol{\beta}) = \sum_{j=1}^{r} l_j(\boldsymbol{\theta}, \boldsymbol{\omega}_j)$. We take $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\omega}}_1', \cdots, \widehat{\boldsymbol{\omega}}_r')'$ as a solution of the likelihood equation

$$\dot{l}(\widehat{\boldsymbol{\beta}}) \;\equiv\; \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \;=\; 0. \tag{2.4}$$

Let

$$\boldsymbol{\beta}_o \equiv (\boldsymbol{\theta}_o', \boldsymbol{\omega}_{1o}', \cdots, \boldsymbol{\omega}_{ro}')' = \arg \max_{\boldsymbol{\theta}, \boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_r} E\{\sum_{j=1}^{r} \log f(\mathbf{X}_{tj}; \boldsymbol{\theta}, \boldsymbol{\omega}_j)\} \tag{2.5}$$

be the true value of the parameter, which is assumed to be an inner point of the (expanded) 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

$$\mathbf{M}_1 = -\begin{pmatrix} \sum_{j=1}^{r} E\mathbf{A}_{tj} & E\mathbf{B}_{t1} & \cdots & E\mathbf{B}_{tr} \\ E\mathbf{B}_{t1}' & E\mathbf{C}_{t1} & & \\ \vdots & & \ddots & \\ E\mathbf{B}_{tr}' & & & E\mathbf{C}_{tr} \end{pmatrix}, \tag{2.6}$$

$$\mathbf{M}_2 = -\begin{pmatrix} \frac{1}{r}\sum_{j=1}^r E\mathbf{A}_{tj} & \frac{1}{\sqrt{r}}E\mathbf{B}_{t1} & \cdots & \frac{1}{\sqrt{r}}E\mathbf{B}_{tr} \\ \frac{1}{\sqrt{r}}E\mathbf{B}'_{t1} & E\mathbf{C}_{t1} & & \\ \vdots & & \ddots & \\ \frac{1}{\sqrt{r}}E\mathbf{B}'_{tr} & & & E\mathbf{C}_{tr} \end{pmatrix}, \tag{2.7}$$

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

We introduce some regularity conditions first.

A1 $\{\mathbf{X}_t\}$ is α-mixing and satisfies the mixing condition 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

$$\overline{\lim_{r\to\infty}} E\{|\frac{1}{r}\sum_{j=1}^r \xi_{tj}|^\nu\} < \infty, \tag{2.8}$$

$$\overline{\lim_{r\to\infty}} \frac{1}{r}\sum_{j=1}^r [E(\eta_{tj}^2) + \{E(|\eta_{tj}|^\nu)\}^{2/\nu}] < \infty. \tag{2.9}$$

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 (2.9) 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 exist 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 (2.9) holds with $\eta_{tj} = \lambda_j(\mathbf{X}_{tj}) - E\{\lambda_j(\mathbf{X}_{tj})\}$.

A7 (2.9) 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

$$\mathbf{W}_k = \lim_{r\to\infty}\frac{1}{r^2}\Big(\sum_{j=1}^r \zeta_{1j}, \sum_{j=1}^r \zeta_{k+1,j}\Big), \qquad k = 0, 1, \cdots, n.$$

exist.

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 (2.5), is an isolated maximizer. It also implies $\mathbf{M}_2$ is positive-definite as $\mathbf{M}_2 = \boldsymbol{\Lambda}\mathbf{M}_1\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a 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.

Theorem 1. Let conditions A1 – A6 hold. Then there exists a solution of the likelihood equation (2.4) 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 restriction on the number of the (nuisance) the parameters which can be consistently estimated, although the implied rates such as $r = o(n^{1/3})$ is presumably too restrictive.

Theorem 2. Let conditions A1 – A7 hold, matrices $E(\mathbf{C}_{1j})$, $j = 1, \cdots, r$, be invertible, and the limit of $\mathbf{M}_2$, defined in (2.7), exist (as $r \to \infty$). Furthermore, let $r/n \to 0$. For any consistent solution of the likelihood equation (2.4) in the sense that

$$||\widehat{\boldsymbol\theta} - \boldsymbol\theta_o||^2 + \sum_{j=1}^{r} ||\widehat{\boldsymbol\omega}_j - \boldsymbol\omega_{jo}||^2 \xrightarrow{P} 0, \tag{2.10}$$

it holds that

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

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 (2.10) 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 $\mathbf{M}_2$.

3 Plug-in quasi-likelihood estimation

We consider now the asymptotic properties of a plug-in 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 function

$$l(\boldsymbol{\theta},\ \boldsymbol{\omega}) = \sum_{t=1}^{n} \log f(\mathbf{X}_t;\ \boldsymbol{\theta},\ \boldsymbol{\omega}), \tag{3.1}$$

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 plug-in 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 plug-in 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 from a Taylor expansion that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \ = \ -\big\{\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^\star)\big\}^{-1}\frac{1}{m\sqrt{n}}\,\dot{l}(\boldsymbol{\theta}_o), \tag{3.2}$$

where $\boldsymbol{\theta}^\star$ is between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, $\dot{l}$ and $\ddot{l}$ are defined in (3.3) below, and m is a normalized constant depending on q and determined by conditions B3 and B4 below.

We introduce regularity conditions first. Let

$$\dot{l}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \qquad \mathbf{a}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}), \tag{3.3}$$

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

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 $1 \le j \le q$, $\widehat{\omega}_j - \omega_{jo} = \frac{1}{n}\sum_{t=1}^{n} g_j(\mathbf{X}_t) + o_P(n^{-1/2})$, where $E\{g_j(\mathbf{X}_t)\} = 0$, $\text{Var}\{g_j(\mathbf{X}_t)\} \le c < \infty$, and $c > 0$ is a constant independent of j. Furthermore $||\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o||^2 = O_P(\tau_{n,q})$, where $\tau_{n,q} \to 0$ and $\tau_{n,q}\sqrt{n}/m \to 0$.

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

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

and $E\{\lambda_1(\mathbf{X}_t)\}$ is bounded as $q \to \infty$.

B3 $\{\mathbf{X}_t\}$ is β-mixing and 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

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

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, $\widetilde{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.

Theorem 3. Under condition 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.

Remark 4. The collective quality of the estimation for all nuisance parameters is reflected by the condition $||\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o||^2 = O_P(\tau_{n,q})$ in B1. With n observations and q (nuisance) parameters in total, the average number of observations available for estimating each parameter may be regarded as in the order of n/q. This suggests $|\widehat{\omega}_j - \omega_{jo}|^2 = O_P(q/n)$ for all $1 \leq j \leq q$ and, consequently, $q/n \leq \tau_{n,q} \leq q^2/n$. In the case $m = q$, B1 implies $q = o(\sqrt{n})$ if $\tau_{n,q} = q^2/n$, and $q = o(n)$ if $\tau_{n,q} = q/n$. Hence the maximum number of nuisance parameters allowed in Theorem 3 depends on the quality of the initial plug-in estimator $\widehat{\boldsymbol{\omega}}$: the faster $\tau_{n,q} \to 0$, the larger q can be.

4 Numerical properties

We consider a simple spatio-temporal model

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{Z}_t, \qquad \mathbf{Z}_t = \rho \mathbf{H}\mathbf{Z}_t + \boldsymbol{\varepsilon}_t, \tag{4.1}$$

where $\mathbf{Y}_t$ is a $p \times 1$ vector, representing the values at time t over p locations, $\mathbf{A} = \mathrm{diag}(\omega_1, \cdots, \omega_p)$ is a diagonal coefficient matrix, the innovation $\mathbf{Z}_t$ in the AR equation is unobservable and its components are correlated with each other. The correlation structure is defined by the second equation above, in which $\mathbf{H}$ is a known $p \times p$ matrix with the main diagonal elements equal to 0 and all the other elements equal to 1, ρ is an unknown parameter, and $\boldsymbol{\varepsilon}_t$ are independent $N(0, \sigma^2 \mathbf{I}_p)$ random vectors, where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. The second equation in (4.1) is a simple example of spatial autoregressive models in spatial econometrics literature; see, e.g. LeSage and Pace (2009).

Under the above setting, each component of $\mathbf{Y}_t$ follows an AR(1) model. However those components are correlated due to the spatial dependence in $\mathbf{Z}_t$. Based on observations $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$, we are interested in estimating the parameter $\boldsymbol{\theta} = (\rho, \sigma^2)'$ which determines the spatial correlations among different locations, treating the temporal autoregressive parameters $\omega_1, \cdots, \omega_p$ as nuisance parameters. We conduct a simulation to compare the performance of the MPQLE and MCQLE for $\boldsymbol{\theta}$. Note for this example, there are $q = p$ nuisance parameters $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)'$.

For the MPQLE, we estimate nuisance parameter ω_j by the ordinary least squares estimation using the j-th component series of $\mathbf{Y}_t = (Y_{t1}, \cdots, Y_{tp})'$ only, i.e.

$$\widehat{\omega}_j = \sum_{t=2}^{n} Y_{tj} Y_{t-1,j} \Big/ \sum_{t=2}^{n} Y_{t-1,j}^2, \qquad j = 1, \cdots, p. \tag{4.2}$$

Let $\widehat{\mathbf{A}} = \mathrm{diag}(\widehat{\omega}_1, \cdots, \widehat{\omega}_p)$. This leads to the residuals $\widetilde{\mathbf{Z}}_t = \mathbf{Y}_t - \widehat{\mathbf{A}} \mathbf{Y}_{t-1}$. It follows from the second equation in (4.1) that

$$\mathbf{Z}_t \sim N\big(0, \sigma^2 (\mathbf{I}_p - \rho \mathbf{H})^{-2}\big). \tag{4.3}$$

Hence the MPQLE is defined as

$$(\widetilde{\rho}, \widetilde{\sigma}^2) = \arg\min_{\rho, \sigma^2} \Big\{ p \log(\sigma^2) - \log(|\mathbf{I}_p - \rho \mathbf{H}|^2) + \frac{1}{\sigma^2 (n-1)} \sum_{t=2}^{n} \widetilde{\mathbf{Z}}_t' (\mathbf{I}_p - \rho \mathbf{H})^2 \widetilde{\mathbf{Z}}_t \Big\}. \tag{4.4}$$

Note that the determinant $|\mathbf{I}_p - \rho \mathbf{H}|$ admits an explicit formula:

$$|\mathbf{I}_p - \rho \mathbf{H}| = (1 + \rho)^{p-1} \{1 - (p-1)\rho\}, \qquad p = 1, 2, \cdots.$$

To construct an MCQLE, we first calculate the profile likelihood for (ρ, σ^2) by maximizing the likelihood based on the component observations $\{(Y_{t,j-1}, Y_{tj})\}$ over (ω_{j-1}, ω_j), for $j = 2, \cdots, p$. This leads to

$$\begin{aligned} \{\omega_{j-1}(\rho), \omega_j(\rho)\} = \arg\min_{\omega_i, \omega_j} \sum_{t=2}^{n} \big\{ & \tau (Y_{t,j-1} - \omega_{j-1} Y_{t-1,j-1})^2 + \tau (Y_{tj} - \omega_j Y_{t-1,j})^2 \\ & - 2\nu (Y_{t,j-1} - \omega_{j-1} Y_{t-1,j-1})(Y_{tj} - \omega_j Y_{t-1,j}) \big\}, \end{aligned} \tag{4.5}$$

where $\tau \equiv \tau(\rho) = \mathrm{Var}(Z_{tj})/\sigma^2$ and $\nu = \nu(\rho) = \mathrm{Cov}(Z_{tj}, Z_{t,j-1})/\sigma^2$. Note that $\omega_j(\rho)$ obtained from the pairing with $\omega_{j-1}(\rho)$ above differs from that obtained from the pairing with $\omega_{j+1}(\rho)$, as we adhere the 'variation-free' condition discussed in section 2 above. Now let

$$\widehat{Z}_{t,j-1} \equiv \widehat{Z}_{t,j-1}(\rho) = Y_{t,j-1} - \omega_{j-1}(\rho)Y_{t-1,j-1}, \quad \widehat{Z}_{t,j} \equiv \widehat{Z}_{t,j}(\rho) = Y_{t,j} - \omega_j(\rho)Y_{t-1,j}.$$

Our MCQLE is defined as

$$\begin{aligned}(\widehat{\rho}, \widehat{\sigma}^2) = \arg\min_{\rho,\sigma^2} \Big\{ & \log(\sigma^2) + \frac{1}{2}\log(\tau^2 - \nu^2) \\ & + \frac{1}{2(n-1)(p-1)(\tau^2-\nu^2)} \sum_{j=2}^{p}\sum_{t=2}^{n} (\tau \widehat{Z}_{t,j}^2 + \tau \widehat{Z}_{t,j-1}^2 - 2\nu \widehat{Z}_{t,j}\widehat{Z}_{t,j-1}) \Big\}.\end{aligned} \tag{4.6}$$

Note that both τ and ν in (4.6) and (4.5) can be explicitly expressed as functions of ρ. To this end, let $\mathbf{G} \equiv (g_{ij}) = (\mathbf{I}_p - \rho\mathbf{H})^{-1}$. Then

$$g_{ii} = \frac{1-(p-2)\rho}{\{1-(p-1)\rho\}(1+\rho)}, \qquad g_{ij} = \frac{\rho}{\{1-(p-1)\rho\}(1+\rho)} \quad (i \neq j).$$

It follows from (4.3) that τ is the main-diagonal element of $\mathbf{G}^2$, and ν is the off-main-diagonal element of $\mathbf{G}^2$. Hence

$$\tau = \frac{\{1-(p-2)\rho\}^2 + (p-1)\rho^2}{\{1-(p-1)\rho\}^2(1+\rho)^2}, \qquad \nu = \frac{2\rho\{1-(p-2)\rho\} + (p-2)\rho^2}{\{1-(p-1)\rho\}^2(1+\rho)^2}.$$

We conducted a simulation to compare the performance of the MPQLE (4.4) and the MCQLE (4.6). We set the sample size $n = 100$ or 300, the parameter $\rho = 0.1$, 0.5 or 0.9 and $\sigma^2 = 1$. For $n = 100$, we set the number of locations $p = 10$, 50 or 100. For $n = 300$, we set $p = 30$, 150 or 300. For each setting, we drew 200 samples. For each sample, nuisance parameters ω_j were drawn independent from the uniform distribution on the interval $[-0.9, 0.9]$. Table 1 lists the mean absolute estimation errors (i.e. in the form $0.5(|\widehat{\rho}-\rho|+|\widehat{\sigma}^2-\sigma^2|)$) over the 200 samples for all different settings.

Since the composite likelihood is a wrong likelihood, it is not surprising to see that the MPQLE, which was calculated based on the correct likelihood or marginal likelihood functions, outperforms the MCQLE under the 'normal' circumstances (i.e. when p is relatively small with respect to n). However when p is large in relation to n, the MPQLE suffers from too many initial estimates $\widehat{\omega}_j$ defined in (4.2); some of them are bound to be poor or very poor. In contrast, the MCQLE profiles out the nuisance parameters ω_j in (4.5), which makes the use of the pairwise correlations. Although the form of likelihood function in (4.6) is wrong, it does not involve any initial estimates. Table 1 indicates that the MCQLE provides more accurate estimates than the MPQLE when, for example,

Table 1: The mean absolute errors of the MCQLE and the MPQLE over 200 replications.

		$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
n	p	MCQLE	MPQLE	MCQLE	MPQLE	MCQLE	MPQLE
100	10	0.031	0.029	0.412	0.082	0.482	0.106
	50	0.355	0.017	0.485	0.370	0.532	0.596
	100	0.372	0.187	0.442	0.449	0.599	0.729
300	30	0.226	0.038	0.398	0.081	0.399	0.641
	150	0.408	0.219	0.407	0.902	0.436	0.703
	300	0.442	0.880	0.479	0.960	0.596	0.981

$p = n = 300$, and also $p = n = 100$ and $\rho \geq 0.5$. This is also the cases when the spatial correlation is strong (e.g. $\rho = 0.9$ or 0.5) and p is moderately large (e.g. $p = 150$ and $n = 300$). Note that for the MPQLE the spatial correlations were completely ignored in estimating nuisance parameters ω_j (see (4.2)). In contrast the pairwise correlation structure was utilized in (4.5) in deriving the MCQLE.

5 Proofs of Theorems 1 and 2

We use the same notation as in section 2.

5.1 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 more involved in order to handle the increasing number of parameters as $n \to \infty$.

Let

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

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

$$\frac{1}{nr}\{l(\boldsymbol\beta)-l(\boldsymbol\beta_o)\} \;=\; \frac{1}{nr}(\boldsymbol\beta-\boldsymbol\beta_o)'\dot{l}(\boldsymbol\beta_o) \;+\; \frac{1}{2nr}(\boldsymbol\beta-\boldsymbol\beta_o)'\ddot{l}(\boldsymbol\beta_o)(\boldsymbol\beta-\boldsymbol\beta_o)$$
$$+\;\; \frac{1}{6nr}\sum_{\ell,i,k}(\beta_\ell-\beta_{\ell o})(\beta_i-\beta_{io})(\beta_k-\beta_{ko})\frac{\partial^3}{\partial\beta_\ell\partial\beta_i\partial\beta_k}l(\boldsymbol\beta^\star) \;\equiv\; S_1+S_2+S_3, \tag{5.1}$$

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,

$$S_1=\frac{\delta\boldsymbol\gamma'}{n\sqrt{m}}\sum_{t=1}^n\frac{1}{r}\sum_{j=1}^r\mathbf{a}_{tj}+\frac{\delta\sqrt{r}}{n\sqrt{m}}\sum_{t=1}^n\frac{1}{r}\sum_{j=1}^r\boldsymbol\gamma_j'\mathbf{b}_{tj}. \tag{5.2}$$

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

$$P(\frac{\sqrt{m}}{n}\sum_{t=1}^n|\frac{1}{r}\sum_{j=1}^r\xi_{tj}|>\epsilon) \;\le\; \frac{m}{n\epsilon^2}\{\mathrm{Var}(\zeta_{tr})+2\sum_{t=1}^{n-1}(1-\frac{t}{n})\mathrm{Cov}(\zeta_{1r},\zeta_{1+t,r})\}$$
$$\le\; \frac{m}{n\epsilon^2}\{\mathrm{Var}(\zeta_{tr})+2E(|\zeta_{tr}|^\nu)^{2/\nu}\sum_{t=1}^\infty\alpha(t)^{1-2/\nu}\}\to 0. \tag{5.3}$$

where $\zeta_{tr}=r^{-1}\sum_{1\le j\le r}\xi_{tj}$. 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 (5.2) 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 (5.2), let d_j denotes the length of $\mathbf{b}_{tj}\equiv(b_{tj1},\cdots,b_{tjd_j})'$. Then $\max_{1\le j\le r}d_j$ are bounded (as $r\to\infty$). Note

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

Hence

$$P\{\sup_{\{\boldsymbol\gamma_j\}}\frac{\sqrt{rm}}{n}|\sum_{t=1}^n\frac{1}{r}\sum_{j=1}^r\boldsymbol\gamma_j'\mathbf{b}_{tj}|>\epsilon\} \;\le\; P\{\frac{\sqrt{rm}}{n}\sum_{j=1}^r\sum_{i=1}^{d_j}|\sum_{t=1}^n b_{tji}|>\epsilon r\}$$
$$\le\;\sum_{j=1}^r P\{\frac{\sqrt{rm}}{n}\sum_{i=1}^{d_j}|\sum_{t=1}^n b_{tji}|>\epsilon\} \;\le\; \sum_{j=1}^r\sum_{i=1}^{d_j}P\{\frac{\sqrt{rm}}{n}|\sum_{t=1}^n b_{tji}|>\epsilon/d_j\}$$
$$\le\; \frac{rm(\max_j d_j)^2}{n\epsilon^2}\sum_{j=1}^r\sum_{i=1}^{d_j}\{\mathrm{Var}(b_{tji})+2(E|b_{tji}|^\nu)^{2/\nu}\sum_{t=1}^\infty\alpha(t)^{1-2/\nu}\}\to 0, \tag{5.4}$$

as $r^2m/n\to 0$ and condition A3. The last inequality in the above expression follows the same argument as for (5.3). This shows that the second sum on the RHS of (5.2) 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 (5.4), 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 in the above expressions are fixed, and the uniform convergence may be established in the same manner as in (5.4). 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 (2.6) and (2.7), 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

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

Note that $E\{\lambda_j(\mathbf{X}_{tj})\}$ is bounded by a constant for $1 \le j \le r$, $|\theta_i - \theta_{io}| \le \delta/\sqrt{m}$ and $|\omega_{jk} - \omega_{jko}| \le \delta\sqrt{r/m}$ 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,

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

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

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.

5.2 Proof of Theorem 2

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

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o = -\{\ddot{l}(\boldsymbol{\beta}^\star)\}^{-1} \dot{l}(\boldsymbol{\beta}_o), \tag{5.5}$$

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 taking the inverse of this partitioned matrix, the first d components of (5.5) may now be expressed as

$$\begin{aligned} &\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \\ = \;& -\Big\{ \frac{1}{nr} \sum_{j=1}^{r} \Big(\sum_{t=1}^{n} \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\{ \sum_{t=1}^{n} \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\}^{-1} \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star)' \Big) \Big\}^{-1} \\ \times \;& \frac{1}{\sqrt{n}\, r} \sum_{j=1}^{r} \Big(\sum_{t=1}^{n} \mathbf{a}_{tj} - \sum_{t=1}^{n} \mathbf{B}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\{ \sum_{t=1}^{n} \mathbf{C}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) \big\}^{-1} \sum_{t=1}^{n} \mathbf{b}_{tj} \Big). \end{aligned} \tag{5.6}$$

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

$$\begin{aligned}
&\max_{1\le j\le r} \Big|\frac{1}{n}\sum_{t=1}^{n} \mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{A}_{1j})\Big|_a \qquad (5.7)\\
\le\ & \max_{1\le j\le r} \frac{1}{n}\Big|\sum_{t=1}^{n}\{\mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - \mathbf{A}_{tj}\}\Big|_a + \max_{1\le j\le r}\Big|\frac{1}{n}\sum_{t=1}^{n}\mathbf{A}_{tj} - E(\mathbf{A}_{1j})\Big|_a\\
\le\ & \{|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_o|_a + \max_{1\le j\le r}|\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a\} \max_{1\le j\le r}\frac{1}{n}\sum_{t=1}^{n}\lambda_j(\mathbf{X}_{tj}) + \max_{1\le j\le r}\Big|\frac{1}{n}\sum_{t=1}^{n}\mathbf{A}_{tj} - E(\mathbf{A}_{1j})\Big|_a.
\end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned}
P\{\max_{1\le j\le r}\Big|\frac{1}{n}\sum_{t=1}^{n}\mathbf{A}_{tj} - E(\mathbf{A}_{1j})\Big|_a > \epsilon\} \ &\le\ \sum_{j=1}^{r} P\{\Big|\frac{1}{n}\sum_{t=1}^{n}\mathbf{A}_{tj} - E(\mathbf{A}_{1j})\Big|_a > \epsilon\} \qquad (5.8)\\
\le\ \frac{c}{n}\sum_{\eta_{tj}}\sum_{j=1}^{r}\Big[\mathrm{Var}(\eta_{tj}) + 2\{E(|\eta_{tj}|^\nu)\}^{2/\nu}\sum_{k=1}^{\infty}\alpha(k)^{1-2/\nu}\Big] \ &\to\ 0.
\end{aligned}$$

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 (5.3). 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

$$\max_{1\le j\le r}\frac{1}{n}\sum_{t=1}^{n}\lambda_j(\mathbf{X}_{tj}) = O_P(1). \qquad (5.9)$$

Now we show that

$$\max_{1\le j\le r}|\boldsymbol{\omega}_j^\star - \boldsymbol{\omega}_{jo}|_a \xrightarrow{P} 0. \qquad (5.10)$$

It follows from (2.10) 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 (5.10) holds. Combining (5.7) – (5.10), we conclude

$$\max_{1\le j\le r}\Big|\frac{1}{n}\sum_{t=1}^{n}\mathbf{A}_{tj}(\boldsymbol{\theta}^\star, \boldsymbol{\omega}_j^\star) - E(\mathbf{A}_{1j})\Big|_a \xrightarrow{P} 0. \qquad (5.11)$$

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 (5.11), we obtain that

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

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 (5.6) 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.

6 Proof of Theorem 3

Due to the plug-in of the nuisance parameter estimator $\widehat{\boldsymbol{\omega}}$ in the likelihood function, the proof of Theorem 3 relies on the asymptotic properties of a generalized U-statistic presented in the Appendix.

Using the notation in section 3, we have

$$\begin{aligned}\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}_o)-\frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\ &=\ \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\big\{\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\widehat{\boldsymbol{\omega}})-\mathbf{a}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\big\} \qquad (6.1)\\ =\ \frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)&+\frac{1}{m\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)'\frac{\partial^2 a_1(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}^\star)}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)\\ \vdots\\ (\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)'\frac{\partial^2 a_d(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}^\star)}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}_o)\end{pmatrix}\\ =\ \frac{1}{n^{3/2}m}\sum_{t,s=1}^{n}\mathbf{C}(\mathbf{X}_t;\boldsymbol{\theta}_o,\boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s)&+O_P\big(\frac{\tau_{n,q}\sqrt{n}}{m}\big),\end{aligned}$$

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

$$\begin{aligned}\sum_{t,s=1}^n \mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s) &= 2 \sum_{1\le t<s\le n} \{\mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_s) \\ &+ \mathbf{C}(\mathbf{X}_s; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t)\} + \sum_{t=1}^n \mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t).\end{aligned} \tag{6.2}$$

By applying the Hoeffding decomposition (A.1) (with $m = 2$) to the first sum on the RHS of (6.2), it follows from (6.1) and (6.2) that

$$\begin{aligned}\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}) &= \frac{1}{m\sqrt{n}}\sum_{t=1}^n \mathbf{a}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o) + \frac{2(n-1)}{n^{3/2}m}\sum_{t=1}^n \mathbf{D}(\boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t) \\ &+ L_n + \frac{1}{n^{3/2}m}\sum_{t=1}^n \mathbf{C}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o)\mathbf{g}(\mathbf{X}_t) + O_P\big(\frac{\tau_{n,q}\sqrt{n}}{m}\big),\end{aligned} \tag{6.3}$$

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 (6.3) 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,

$$\frac{1}{m\sqrt{n}}\dot{l}(\boldsymbol{\theta}_o) \xrightarrow{D} N(0,\ \boldsymbol{\Sigma}_0 + 2\sum_{j=1}^\infty \boldsymbol{\Sigma}_j). \tag{6.4}$$

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

On the other hand,

$$\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^\star) = \frac{1}{nm}\sum_{t=1}^n \mathbf{B}(\mathbf{X}_t; \boldsymbol{\theta}_o, \boldsymbol{\omega}_o) + \frac{1}{nm}\sum_{t=1}^n \mathbf{G}(\mathbf{X}_t; \boldsymbol{\theta}^{\star\star}, \boldsymbol{\omega}^\star, \boldsymbol{\theta}^\star - \boldsymbol{\theta}_o, \widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o), \tag{6.5}$$

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

$$(\boldsymbol{\theta}^{\star} - \boldsymbol{\theta}_o)' \frac{\partial}{\partial \boldsymbol{\theta}} b_{ij}(\mathbf{X}_t; \boldsymbol{\theta}^{\star\star}, \boldsymbol{\omega}^{\star}) + (\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o)' \frac{\partial}{\partial \boldsymbol{\omega}} b_{ij}(\mathbf{X}_t; \boldsymbol{\theta}^{\star\star}, \boldsymbol{\omega}^{\star}), \tag{6.6}$$

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 (6.6) 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

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

Since $||\boldsymbol{\theta}^{\star} - \boldsymbol{\theta}_o|| + ||\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}_o|| \xrightarrow{P} 0$, the second term on the RHS of (6.5) converges to 0 in probability. Therefore $\frac{1}{nm}\ddot{l}(\boldsymbol{\theta}^{\star}) \xrightarrow{P} \mathbf{M}$. This, together with (6.4), concludes the theorem.

7 Conclusion

In this paper we have established the asymptotic normality for the two estimation methods, namely the MCQLE and the MPQLE, for the parameter of interest in the presence of q nuisance parameters, under the assumption that q goes to infinity together with the sample size n. When q is small in relation to n, the MPQLE performs well and is typically better than the MCQLE. However when q and n are about the same (hence condition B1 no longer holds), the MPQLE suffers from the collectively poor estimation for too many nuisance parameters. Then the MCQLE provides a better alternative as it is still root-n consistent. An interesting and practical relevant question is when to use what for a given n and q. The asymptotic results provided in this paper are too complicated to give any clear indication. How to develop an effective inference method to choose between the two methods in practice remains as an unsolved challenge.

Appendix: U-statistics

Let $\boldsymbol{\xi}_t$ is 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

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

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

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

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 $\gamma = \min\{1,\ \frac{2(\delta-\delta')}{\delta'(2+\delta)}\}$.

Proposition 1 is essentially Lemma 2 of Yoshihara (1976). 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, as only the α-mixing is required now, which is weaker than the β-mixing. See section 2.6 of Fan and Yao (2003).

C3 $\{\boldsymbol{\xi}_t\}$ is a strictly stationary and α-mixing (i.e. strong mixing) process with α-mixing coefficients satisfying the condition $\sum_{k \ge 1} \alpha(k)^{1-2/\nu} < \infty$, 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 $\text{Cov}\{\psi_{n,1}(\boldsymbol{\xi}_1),\ \psi_{n,1}(\boldsymbol{\xi}_j)\}$ exists for any $1 \le j \le n$.

Put

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

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

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

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

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

see condition C4. Hence it follows from condition C3 that

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

Now by the Lebesgue dominated convergence theorem, it holds that

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

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),

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

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) entail the required result. ■

Acknowledgements

We thank reviewers for their helpful comments, especially for pointing out the link between our work and the incidental parameters problem.

References

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

Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72**, 355-366.

Berger, J.O., Liseo, B. and Wolpert, R.L. (1997). Integrated likelihood methods for eliminating nuisance parameters. *Statis. Sci.* **14**, 1-28.

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. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Stats. Soc.* **B**, **49**, 1-39.

Cox, D.R. and Reid, N. (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. *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.

Engle, R.F., Shephard, N. and Sheppard, K. (2008). Fitting and testing vast dimensional time-varying covariance models. *NYU Working Paper No.FIN-07-046.*
Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1293629`

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.

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

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

Lancaster, T. (2000). The incidental parameter problem since 1948. *J. Econometrics*, **95**, 391-413.

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* (2nd edition). Springer, New York.

LeSage, J. and Pace, R.K. (2009). *Introduction to Spatial Econometrics.* CRC/Chapman & Hall, Boca Raton.

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.

Neyman, J. and Scott, E.S. (1948). Consistent estimation from partially consistent observations. *Econometrica*, **16**, 1-32.

Reid, N. (1996). Likelihood and Bayesian approximation methods. In: Bernardo, J.M., Berger, Jo.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics, Vol. 5.* Oxford University Press, Oxford.

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

Tao, M., Wang, Y., Yao, Q. and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Amer. Statist. Assoc.* **106**, 1025-1040.

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

Yoshihara, K. (1976). Limiting behavior 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.