Lodz
Economics
Working
Papers

Improved asymptotic analysis of Gaussian QML estimators in spatial models

Jakub Olejnik
Alicja Olejnik

9/2017

FACULTY OF ECONOMICS AND SOCIOLOGY
University of LODZ

Improved asymptotic analysis of Gaussian QML estimators in spatial models
Working Paper

Jakub Olejnik
Department of Mathematics and Computer Science
University of Lodz
Banacha 22, 90-238 Łódź, Poland

Alicja Olejnik*
Department of Economics and Sociology
University of Lodz
3/5 POW Street, 90-255 Łódź, Poland

December 7, 2017

Abstract

This paper presents a fundamentally improved statement on asymptotic behaviour of the well-known Gaussian QML estimator of parameters in high-order mixed regressive/autoregressive spatial model. We generalize the approach previously known in the econometric literature by considerably weakening assumptions on the spatial weight matrix, distribution of the residuals and the parameter space for the spatial autoregressive parameter. As an example application of our new asymptotic analysis we also give a statement on the large sample behaviour of a general fixed effects design.

KEYWORDS: spatial autoregression, quasi-maximum likelihood estimation, high-order SAR model, asymptotic analysis, fixed effects model
JEL CLASSIFICATION: C21, C23, C51

1 Introduction

Spatial econometrics constitutes an important chapter in the econometric literature. In the case of cross-sectional and panel data, researchers often have to deal with spatial interactions which are embodied through spatial autocorrelation, see, e.g. Cliff, A. D. and Ord, J. K. (1973), Paelinck, J. *et. al.* (1979), Anselin, L. (1988), and Cressie, N. (1993). Through the years, applied spatial econometrics attracted attention of researchers in the fields of regional science, urban and real estate economics. In the past decade, a dynamic development of the theory of spatial econometrics could also be observed. Importantly, many contributions aimed at providing rigorous arguments on asymptotic properties of estimators for spatial models.

Our work has been largely motivated by the seminal paper of Lee, L. F. (2004), where consistency and asymptotic distribution of the Gaussian Quasi-Maximum Likelihood estimators

*Corresponding author, e-mail: olejnika@uni.lodz.pl, phone: +48 42 6355131

for spatial autoregressive models are investigated. In particular, the paper describes conditions imposed on, inter alia, asymptotic behaviour of the spatial weight matrix and distribution of the residuals which imply consistency and asymptotic normality of estimates.[1] Many significant theoretical studies have already applied the asymptotic theory of Lee, L. F. (2004) in extended model specifications. Among other things, spatial models with various forms of fixed effects have been considered. For example, Lee, L. F. and Yu, J. (2010a) develop a QML estimator for a spatial autoregressive model with time and individual fixed effects. In addition, they suggest a bias correction for estimates of the error variance. In Lee, L. F. *et. al.* (2010) a further extension to a non-balanced panel case is described, together with an application to social networking models. Furthermore, Shi, W. and Lee, L. F. (2017) consider a QML estimation scheme of a dynamic spatial panel data model with interactive fixed-effects. The ideas of Lee, L. F. (2004) have also been used to study model specifications with dynamic dependence of the explained variable, see e.g. Yu, J. *et. al.* (2008). Finally, a very recent paper of Qu, X. and Lee, L. F. (2017) extends the standard asymptotic analysis to the case of dynamic spatial models with time-varying endogenous spatial weight matrix.

In recent years, high-order spatial autoregressive models have been gaining popularity in applied spatial econometrics. Therefore, they are also an increasingly interesting subject for theoretical considerations. For example, Gupta, A. and Robinson, P. (2015) develop a high-order specification where the number of autoregressive terms grows to infinity and Elhorst, J. P. *et. al.* (2012) provide insight on the problem of high-order autoregressive parameter space definition. Also, alternative estimation procedures for high-order models have been investigated. Most notably, Han, X. *et. al.* (2017) exercise the Bayesian approach. Badinger, H. and Egger, P. (2013), in turn, refine GMM estimation of spatial high-order error component models. Moreover, recently Li, K. (2017) studies impulse response analysis in the case of fixed effects dynamic high-order spatial panel models, using the standard asymptotic analysis.

The large sample analyses presented in the papers based on Lee, L. F. (2004) inherit some of the limitations of the original argument. In particular, the set of possible spatial weight matrices is restricted to those which are row and column summable.[2] Unfortunately, this prerequisite excludes from theoretical considerations some of the spatial interaction patterns in which the number of spatial units influenced by any given unit grows to infinity. In particular, under infill asymptotics, if spatial units are assumed to interact with other units within a given distance, then the number of non-zero elements in a row or column of the spatial weight matrix grows with the sample size. Similarly, if the increasing domain asymptotics is considered, then certain spatial weight matrices based on inverted distance also lead to non-summable interaction patterns.[3] Moreover, in cases where the original specification is transformed, e.g. by linear filtering or demeaning, it is necessary to ensure that the applied transformation preserves the summability of rows and columns of the spatial weight matrix. As a result, this requirement narrows the class of possible transformations of the model.[4] Furthermore, the standard approach stipulates for the components of the model residuals to be independent

[1]For the purpose of the present paper we will refer to the asymptotic theory presented therein as the standard asymptotic analysis.

[2]See Assumption 5 in Lee, L. F. (2004). Moreover, the inverse spatial difference operator ($\Delta^{-1} = (\mathrm{I} - \lambda W)^{-1}$) is also required to be row and column summable for each possible value of the spatial autoregressive parameter λ.

[3]An example of a spatial weight matrix which describes an interaction pattern not covered by the standard analysis of Lee, L. F. (2004) is given in Section 2.6.

[4]An example of a theoretical argument which would not be possible under the standard asymptotic analysis is presented in Section 3. Somewhat similar situation can be observed in Lee, L. F. and Yu, J. (2013), where a transformation of spatial weight matrix is introduced to eliminate its near unit root eigenvalues. The analysis could be more general and the assumptions made therein could be simpler and less restrictive if our improved asymptotic analysis were employed.

identically distributed. Also, the existence of high-order moments (higher than four) for the distribution is required. Likewise, as will be shown in the present paper, these prerequisites unnecessarily reduce the usability of the QML estimation scheme.

The standard asymptotic analysis of the Gaussian QML estimators requires the true value of the spatial autoregressive parameter $\lambda = (\lambda_j)_{j \leq d}$ to lie within the interior of its parameter space. In many applications this prerequisite on its own is not overly restrictive. In practice, the space is assumed to be the interior of a compact subset of the pre-image of $[0, 1)$ under the continuous map $\mathbb{R}^d \ni \lambda \mapsto \| \sum_{j=1}^{d} \lambda_j W_j \|_0$, where W_j are the spatial weight matrices and $\| \cdot \|_0$ is an arbitrary matrix norm. However, it is still beneficial to search for asymptotic theories which alleviate such restrictions. Firstly, they could allow one to consider more flexible definitions of the parameter space, see Elhorst, J. P. *et. al.* (2012) for a pioneering approach. Secondly, under model specifications in which further restrictions on elements of λ are imposed, such theories could provide immediate implications on the consistency and asymptotic distribution of QML estimates.

Therefore, the aim of this paper is to present a refined asymptotic analysis of the Gaussian QML estimator for high-order spatial autoregressive models. Our argument, as compared to the standard analysis originating in Lee, L. F. (2004), features the following improvements.

- It accounts for a larger class of spatial weight matrices, i.e. not necessarily having all rows and columns absolutely summable As a result, models with a greater amount of spatial interaction can be considered.
- It allows for the consideration of a larger class of distributions for the vector of model residuals. In particular, individual innovations do not have to follow the same distribution.
- Consistency of the Gaussian QML estimator is proven with minimal assumptions.
- The parameter space for the autoregressive parameter is considered in a fully general form – it is merely assumed to be compact. The space is not required to be connected, nor to contain the true parameter as its internal point.
- It is possible to apply a larger class of transformations to model specification in theoretical arguments, while preserving the asymptotic properties of the Gaussian QML estimators.

Furthermore, we will present an extension of the general fixed effects model described in Olejnik, A. and Olejnik, J. (2017). Our approach to elimination of fixed effects from the spatial process also generalises the approaches of Lee, L. F. and Yu, J. (2010a) and Lee, L. F. *et. al.* (2010). As an example of possible applications of our improved analysis, we will develop theorems on consistency and asymptotic normality of this estimation technique which would not be possible with the standard approach.

The paper is organized as follows. Section 2.1 introduces the basic notation used throughout the paper. Section 2.2 describes the Gaussian QML estimator for the high-order spatial autoregressive model. Our improved statements on its consistency and asymptotic normality are presented in Sections 2.3 and 2.4 respectively. Section 2.5 discusses the assumptions adopted in this paper and compares them with the prerequisites for the standard asymptotic analysis. Finally, Section 3 develops an estimator for a high-order spatial autoregressive general fixed effects model, together with an analysis of its large sample behaviour.

2 Improved asymptotic analysis of high-order SAR model

This section presents the foundations of our improved asymptotic analysis. Let us note that there are two important tools used in the subsequent arguments. The first is a more or less standard lemma used for obtaining consistency, namely Lemma 3.1 in Pötscher, M. B. and Prucha, I. R. (1997). The second is a new CLT for linear-quadratic forms (our Theorem C.1). We prove it in Appendix C by the use of an argument based on bounds originally developed in Bhansali, R. J. *et. al.* (2007), where a CLT for quadratic forms of i.i.d. vectors is shown.

2.1 Basic notation and the SAR model

Throughout the paper we assume the following natural notation. Let $d \in \mathbb{N}$ be fixed. Suppose $A_1, \dots, A_d$ are matrices of the same, arbitrary dimension and let $\mathbf{A} = \langle A_1, \dots, A_d \rangle^{\mathrm{T}}$ be a d element vector of those matrices. Let $\lambda \in \mathbb{R}^d$, $\lambda = (\lambda_1, \dots, \lambda_d)^{\mathrm{T}}$. Then, the formula $\lambda^{\mathrm{T}}\mathbf{A}$ denotes the matrix $\sum_{r=1}^{d} \lambda_r A_r$. Furthermore, if B and C are arbitrary matrices for which the products $A_1 B$ and CA_1 are defined, then we set

$$\lambda^{\mathrm{T}}\mathbf{A}B = \left(\lambda^{\mathrm{T}}\mathbf{A}\right) B \quad \text{and} \quad C\lambda^{\mathrm{T}}\mathbf{A} = C\left(\lambda^{\mathrm{T}}\mathbf{A}\right).$$

Clearly, the newly defined product $\lambda^{\mathrm{T}}\mathbf{A}$ is associative in a manner that is consistent with the ordinary product of matrices.

Unless stated differently, vectors, i.e. elements of $\mathbb{R}^m$, for some $m \in \mathbb{N}$, are $m \times 1$ column vectors. The symbol $\|x\|$ for a non-stochastic $x \in \mathbb{R}^m \cup (\mathbb{R}^m)^{\mathrm{T}}$, denotes the usual vector norm (i.e. $\sqrt{x^{\mathrm{T}}x}$ or $\sqrt{xx^{\mathrm{T}}}$). The same symbol, when used for matrices, denotes the induced operator norm, i.e. the spectral norm. Namely, for an $m_1 \times m_2$ matrix A we have $\|A\| = \sup_{x \in \mathbb{R}^{m_2} :\, \|x\|=1} \|Ax\|$. Alternatively, $\|A\|$ is the largest singular value of A. We will also use the matrix norms $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{tr}\,(A^{\mathrm{T}}A)}$ and $\|A\|_{\mathrm{K}} = \max\{\|A\|_1, \|A\|_\infty\}$, with $\|A\|_1 = \max_{i \le m_1} \{\sum_{j=1}^{m_2} |a_{ij}|\}$, $\|A\|_\infty = \|A^{\mathrm{T}}\|_1$, $A = [a_{ij}]_{ij}$. The symbol I_m denotes the identity matrix of size $m \times m$ and $n \in N$ always refers to the sample size. Lastly, for any set S, the symbol $\mathbb{I}_S$ denotes its indicator function on the implied domain.

Let $\mathbf{W}_n = \langle W_{n,1}, \dots, W_{n,d} \rangle^{\mathrm{T}}$ be a vector of $n \times n$ arbitrary matrices.[5] The high-order SAR model is described by the following equation

$$Y_n = \lambda^{\mathrm{T}}\mathbf{W}_n Y_n + X_n \beta + \varepsilon_n, \tag{2.1}$$

where $\lambda \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^k$ are model parameters. Furthermore, Y_n is a vector of n observations on the dependent variable, X_n is the matrix of k explanatory variables and ε_n is the error term, for which the assumptions given below hold. See also Section 2.5 for a detailed discussion.

Assumption 1 Each X_n, $n \in \mathbb{N}$, is a non-stochastic $n \times k$ matrix. Moreover, the matrix $X_n^{\mathrm{T}} X_n$ is invertible and we have

$$\sup_{n \in \mathbb{N}} \left\| n^{-1} \cdot X_n^{\mathrm{T}} X_n \right\| < \infty \quad \text{and} \quad \sup_{n \in \mathbb{N}} \left\| n \cdot (X_n^{\mathrm{T}} X_n)^{-1} \right\| < \infty.$$

Assumption 2 For each $n \in \mathbb{N}$ there exists $\bar{n} \in \mathbb{N}$, $\bar{n} = \bar{n}(n)$,[6] for which the vector of residuals $\varepsilon_n = (\varepsilon_{n,i})_{i \le n}$ is of the form $\varepsilon_n = \mathrm{E}_n \bar{\varepsilon}_n$, where E_n is an $n \times \bar{n}$ non-random matrix with orthogonal rows of unit norm and $\bar{\varepsilon}_n = (\bar{\varepsilon}_{n,i})_{i \le \bar{n}}$ is a random vector with quadruple independent elements[7] satisfying $\mathbb{E}\,\bar{\varepsilon}_{n,i} = 0$, $\mathbb{E}\,\bar{\varepsilon}_{n,i}^2 = \sigma^2$, for some $\sigma^2 > 0$, and $\sup_{n,i} \mathbb{E}\,\bar{\varepsilon}_{n,i}^4 < \infty$.[8]

Finally, the estimated model parameter is $\theta = \left(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \sigma^2\right)^{\mathrm{T}}$.

[5] Typically, in practice, the spatial weight matrices $W_{n,r}$, $r \le d$, have non-negative elements and zero diagonals, i.e. $W_{n,r,ij} = 0$ for $i = j$, which facilitates interpretation of model parameters. However, for the purpose of the argument of Section 3, we deliberately do not assume so.

[6] That is, $\bar{n}$ is a function of n.

[7] Namely, for each $n \in \mathbb{N}$, any two, three or four elements of $(\bar{\varepsilon}_{n,i})_{i \le n}$ are independent.

[8] In particular, we have $\mathbb{E}\,\varepsilon_n = 0$ and $\mathbb{E}\,\varepsilon_n \varepsilon_n^{\mathrm{T}} = \sigma^2 \mathrm{I}_n$. Elements of ε_n do not need to be pairwise independent, although, in typical applications it will be assumed that $\varepsilon_n = \bar{\varepsilon}_n$.

2.2 The Gaussian QML estimator

Let $\Lambda \subset \mathbb{R}^d$ be the parameter space for the autoregressive parameter λ. In order to describe the Gaussian QML estimation procedure we make the following assumption.

Assumption 3 For every $\lambda \in \Lambda$ and $n \in \mathbb{N}$ the matrix $\Delta_n(\lambda) = \mathrm{I}_n - \lambda^{\mathrm{T}}\mathbf{W}_n$ is invertible.

Under the assumption of normally distributed innovations (i.e. $\varepsilon_n \sim \mathbf{N}(0, \sigma^2\mathrm{I}_n)$), it can be shown that the log-likelihood of θ is

$$\ln L_n(\theta) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln|\det \Delta_n(\lambda)| - \frac{1}{2\sigma^2}\left\|\Delta_n(\lambda)Y_n - X_n\beta\right\|^2. \tag{2.2}$$

It is generally known that, provided that certain regularity conditions are met, a consistent estimator can be based on the maximisation of $\ln L_n$, even if the model residuals do not follow normal distribution. In such case $\hat\theta_n = \arg\max_\theta [\ln L_n(\theta)]$ is referred to as the Gaussian QML estimator. A standard approach in obtaining the value for $\hat\theta_n$ is to first concentrate out the parameters β and σ^2. That is, substituting

$$\begin{aligned} \hat\beta_n(\lambda) &= \left(X_n^{\mathrm{T}}X_n\right)^{-1}X_n^{\mathrm{T}}\Delta_n(\lambda)Y_n, \\ \hat\sigma_n^2(\lambda) &= \tfrac{1}{n}\|\Delta_n(\lambda)Y_n - X_n\hat\beta_n(\lambda)\|^2, \end{aligned} \tag{2.3}$$

which are implied by first order optimality conditions, into (2.2), we get the concentrated log-likelihood

$$\ln L_n^{\mathrm{c}}(\lambda) = -\frac{n}{2}\left(\ln\left(2\pi\cdot\hat\sigma_n^2(\lambda)\right) + 1\right) + \ln|\det\Delta_n(\lambda)|. \tag{2.4}$$

Maximising $\ln L_n^{\mathrm{c}}(\lambda)$ with respect to the autoregressive parameter yields $\hat\lambda_n = \arg\max_\lambda [\ln L_n^{\mathrm{c}}(\lambda)]$. Finally, $\hat\theta_n = \left(\hat\beta_n(\hat\lambda_n)^{\mathrm{T}}, \hat\lambda_n^{\mathrm{T}}, \hat\sigma_n^2(\hat\lambda_n)\right)^{\mathrm{T}}$ is the QML estimator for θ, which we will denote in short $\left(\hat\beta_n^{\mathrm{T}}, \hat\lambda_n^{\mathrm{T}}, \hat\sigma_n^2\right)^{\mathrm{T}}$.

2.3 Consistency of $\hat\theta_n$

Our result on consistency of $\hat\theta_n$ requires the following boundedness assumption.

Assumption 4 The set Λ is compact in R^d. There exists a universal constant K_Λ such that for all $n \in \mathbb{N}$, $\lambda \in \Lambda$, $r = 1, \ldots, d$ the matrix norms $\|W_{n,r}\|$ and $\|\Delta_n(\lambda)^{-1}\|$ do not exceed K_Λ.

The following Remark 2.1 is a non-trivial implication of Assumption 4. It turns out to be crucial to our argument. We prove it in Appendix A.

Remark 2.1 If $\mathbf{W}_n$ and Λ satisfy Assumptions 3 and 4, then there exists a bounded open set $U \subset \mathbb{R}^d$, $U \supset \Lambda$,[9] such that $\Delta_n(\lambda)$ is invertible for each $\lambda \in U$ and the norms $\|\Delta_n(\lambda)^{-1}\|$, $\|\Delta_n(\lambda)\|$ do not exceed $K_\Lambda + 1$.

Throughout the paper, U_Λ will denote one fixed set U satisfying the statement of Remark 2.1. The following identification assumption guarantees that the Gaussian QML estimator is able to identify the true value of the spatial autoregressive parameter λ.

Assumption 5 For every $\lambda_1, \lambda_2 \in \Lambda$, such that $\lambda_1 \neq \lambda_2$, at least one of the statements **(a)**, **(b)** below is satisfied:

(a) $\liminf_{n\to\infty} \frac{1}{\sqrt{n}D_n(\lambda)}\|\Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1}\|_{\mathrm{F}} > 1$,

(b) $\liminf_{n\to\infty} \frac{1}{\sqrt{n}}\|M_{X_n}\Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1}X_n\beta\| > 0$, for every $\beta \in \mathbb{R}^k$,

[9] Importantly, the set U does not depend on n.

where $D_n(\lambda) = |\det(\Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1})|^{1/n}$, $M_{X_n} = \mathrm{I}_n - X_n\left(X_n^{\mathrm{T}} X_n\right)^{-1} X_n^{\mathrm{T}}$.

Theorem 2.1. *Under* Assumptions 1 – 5 *the Gaussian QML estimator* $\hat{\theta}_n = \left(\hat{\beta}_n^{\mathrm{T}}, \hat{\lambda}_n^{\mathrm{T}}, \hat{\sigma}_n^2\right)^{\mathrm{T}}$ *described in* Section 2.2 *is consistent.*

Proof. Let $\theta_0 = \left(\beta_0^{\mathrm{T}}, \lambda_0^{\mathrm{T}}, \sigma_0^2\right)^{\mathrm{T}}$ be the true value of parameter θ. Let $S_n(\lambda) = \Delta_n(\lambda)\Delta_n(\lambda_0)^{-1}$, $\lambda \in \Lambda$. Moreover, let P_{X_n} denote the projection matrix $X_n\left(X_n^{\mathrm{T}} X_n\right)^{-1} X_n^{\mathrm{T}}$ and let $M_{X_n} = \mathrm{I}_n - P_{X_n}$.

Let us set $R_n(\lambda) = \frac{1}{n}\ln L_n^{\mathrm{c}}(\lambda) + \frac{1}{2}(\ln(2\pi) + 1)$, for $\lambda \in \Lambda$, cf. (2.4). Naturally, maximising $\ln L_n^{\mathrm{c}}$ is equivalent to maximising R_n. Let us define the non-random function $\bar{R}_n\colon U_\Lambda \to \mathbb{R}$, $n \in \mathbb{N}$, by the following formula

$$\bar{R}_n(\lambda) = \frac{1}{n}\ln|\det \Delta_n(\lambda)| - \frac{1}{2}\ln(\chi_n(\lambda)), \tag{2.5}$$

where $\chi_n(\lambda) = \frac{1}{n}\left\|M_{X_n} S_n(\lambda) X_n \beta_0\right\|^2 + \frac{\sigma_0^2}{n}\left\|S_n(\lambda)\right\|_{\mathrm{F}}^2$.

Now, we will show that $\sup_{\lambda\in\Lambda}\left|R_n(\lambda) - \bar{R}_n(\lambda)\right|$ converges to 0 in probability. As

$$Y_n = \Delta_n(\lambda_0)^{-1} X_n \beta_0 + \Delta_n(\lambda_0)^{-1}\varepsilon_n, \tag{2.6}$$

we have, cf. (2.3),

$$\begin{aligned}\hat{\sigma}_n^2(\lambda) = \frac{1}{n}\left\|M_{X_n}\Delta_n(\lambda) Y_n\right\|^2 &= \frac{1}{n}\left\|M_{X_n} S_n(\lambda) X_n \beta_0 + M_{X_n} S_n(\lambda)\varepsilon_n\right\|^2 \\ &= \chi_n(\lambda) + \xi_n^{(1)}(\lambda) - \xi_n^{(2)}(\lambda) + \xi_n^{(3)}(\lambda),\end{aligned}$$

where

$$\begin{aligned}\xi_n^{(1)}(\lambda) &= 2n^{-1}\cdot\left(S_n(\lambda) X_n \beta_0\right)^{\mathrm{T}} M_{X_n} S_n(\lambda)\varepsilon_n, \\ \xi_n^{(2)}(\lambda) &= n^{-1}\varepsilon_n^{\mathrm{T}} S_n(\lambda)^{\mathrm{T}} P_{X_n} S_n(\lambda)\varepsilon_n, \\ \xi_n^{(3)}(\lambda) &= n^{-1}\varepsilon_n^{\mathrm{T}} S_n(\lambda)^{\mathrm{T}} S_n(\lambda)\varepsilon_n - n^{-1}\sigma_0^2\left\|S_n(\lambda)\right\|_{\mathrm{F}}^2.\end{aligned}$$

The fact that quantities $\xi_n^{(1)}(\lambda)$, $\xi_n^{(2)}(\lambda)$ and $\xi_n^{(3)}(\lambda)$ converge to 0 in probability, uniformly in $\lambda \in \Lambda$, is a consequence of Lemma B.2 **(a)**, **(c)** and **(b)** respectively, with the use of Assumptions 1 and 4.[10] We have $\chi_n(\lambda) \geq \frac{\sigma_0^2}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2 \geq \sigma_0^2\|S_n(\lambda)^{-1}\|^{-2}$. Thus, by Assumption 4, χ_n is uniformly separated from 0 on Λ. Lastly, we conclude that

$$2\left(\bar{R}_n(\lambda) - R_n(\lambda)\right) = \ln\frac{\hat{\sigma}_n^2(\lambda)}{\chi_n(\lambda)} = \ln\left(1 + \frac{\xi_n^{(1)}(\lambda) + \xi_n^{(2)}(\lambda) + \xi_n^{(3)}(\lambda)}{\chi_n(\lambda)}\right)$$

converges in probability to 0, uniformly in λ.

Since $\bar{R}_n(\lambda_0) = \frac{1}{n}\ln|\det\Delta_n(\lambda_0)| - \frac{1}{2}\ln\sigma_0^2$ (cf. (2.5)), we have

$$\begin{aligned}2\left(\bar{R}_n(\lambda_0) - \bar{R}_n(\lambda)\right) &= \ln\frac{\frac{1}{n\sigma_0^2}\left\|M_{X_n} S_n(\lambda) X_n\beta_0\right\|^2 + \frac{1}{n}\left\|S_n(\lambda)\right\|_{\mathrm{F}}^2}{\left|\det S_n(\lambda)\right|^{2/n}} \\ &\geq \ln\left(\frac{C}{n\sigma_0^2}\left\|M_{X_n} S_n(\lambda) X_n\beta_0\right\|^2 + \frac{\frac{1}{n}\left\|S_n(\lambda)\right\|_{\mathrm{F}}^2}{\left|\det S_n(\lambda)\right|^{2/n}}\right) \geq 0,\end{aligned}$$

[10] Also notice that by Assumption 1 we have $\|X_n\| = \mathrm{O}(\sqrt{n})$.

with $C = K_\Lambda^{-2}(1+K_\Lambda \sup_{\lambda\in\Lambda}\|\lambda\|)^{-2} > 0$. Indeed, by elementary AM-GM inequality, $\frac{1}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2 \geq |\det S_n(\lambda)|^{2/n}$ for $\lambda \in \Lambda$. Furthermore, Assumption 5 implies that $\liminf_{n\to\infty} \bar{R}_n(\lambda_0) - \bar{R}_n(\lambda) > 0$, for any $\lambda \in \Lambda$.

We will show that $(\lambda_0)_{n\in\mathbb{N}}$ is an identifiably unique sequence of maximisers of $\bar{R}_n$, see Definition 3.1 in Pötscher, M. B. and Prucha, I. R. (1997). To this end, let us assume the contrary. Then, there is a number $\epsilon > 0$ for which, for some increasing sequence $\{k(n)\}_{n\in\mathbb{N}} \subset \mathbb{N}$ and some sequence $\{\tilde{\lambda}_n\}_{n\in\mathbb{N}} \subset C_\epsilon = \{\lambda \in \Lambda\colon \|\lambda - \lambda_0\| \geq \epsilon\}$, we have $\lim_{n\to\infty} \bar{R}_{k(n)}(\lambda_0) - \bar{R}_{k(n)}(\tilde{\lambda}_n) \leq 0$. Since C_ϵ is closed in compact Λ, the sequences $\{\tilde{\lambda}_n\}_{n\in\mathbb{N}}$ and $\{k(n)\}_{n\in\mathbb{N}} \subset \mathbb{N}$ can be chosen in such a way that $\tilde{\lambda}_n \to \tilde{\lambda}$, for some $\tilde{\lambda} \neq \lambda_0$. Let

$$\delta = \liminf_{n\to\infty} \bar{R}_n(\lambda_0) - \bar{R}_n(\tilde{\lambda}) > 0.$$

According to Remark A.6 and Lemma B.1, all $\bar{R}_n$ are Lipschitz continuous on Λ with a uniform constant K_{L}. We can choose $n_0 \in \mathbb{N}$ such that $\|\tilde{\lambda}_m - \tilde{\lambda}\| < \frac{\delta}{3K_{\mathrm{L}}}$ for all $m \geq n_0$. Finally, the contradiction follows from the inequality

$$\delta \leq \liminf_{n\to\infty} \left(\bar{R}_{k(n)}(\lambda_0) - \bar{R}_{k(n)}(\tilde{\lambda}_n) + |\bar{R}_{k(n)}(\tilde{\lambda}_n) - \bar{R}_{k(n)}(\tilde{\lambda})|\right) \leq \frac{\delta}{3}.$$

Lastly, the convergence $\|\hat{\lambda}_n - \lambda_0\| = \mathrm{o}_{\mathbb{P}}(1)$, follows from Lemma 3.1 in Pötscher, M. B. and Prucha, I. R. (1997) as, by definition, for each $n \in \mathbb{N}$, $\hat{\lambda}_n$ is a maximiser of R_n.

Notice that, by (2.3) and (2.6), we have $\hat{\beta}_n(\hat{\lambda}_n) = \beta_0 - \zeta_n^{(1)} + \zeta_n^{(2)} - \zeta_n^{(3)}$, where

$$\begin{aligned}
\zeta_n^{(1)} &= \left(X_n^{\mathrm{T}} X_n\right)^{-1} X_n^{\mathrm{T}}(\hat{\lambda}_n - \lambda_0)^{\mathrm{T}} \mathbf{W}_n \Delta_n(\lambda_0)^{-1} X_n \beta_0,\\
\zeta_n^{(2)} &= \left(X_n^{\mathrm{T}} X_n\right)^{-1} X_n^{\mathrm{T}} \varepsilon_n,\\
\zeta_n^{(3)} &= \left(X_n^{\mathrm{T}} X_n\right)^{-1} X_n^{\mathrm{T}}(\hat{\lambda}_n - \lambda_0)^{\mathrm{T}} \mathbf{W}_n \Delta_n(\lambda_0)^{-1} \varepsilon_n.
\end{aligned}$$

From Assumptions 1 and 4, using submultiplicativity of the norm, we have $\|\zeta_n^{(1)}\| = \mathrm{O}(\|\hat{\lambda}_n - \lambda_0\|) = \mathrm{o}_{\mathbb{P}}(1)$. The convergence $\|\zeta_n^{(3)}\| = \mathrm{o}_{\mathbb{P}}(1)$ can be deduced from Chebyshev inequality, as, for a constant C, we have $(\mathbb{E}_{\theta_0}\|\zeta_n^{(3)}\|)^2 \leq C\sigma_0^2\, \mathbb{E}_{\theta_0}\|\hat{\lambda}_n - \lambda_0\|^2$, by Schwartz inequality. Lastly, $\|\operatorname{Var}_\theta \zeta_n^{(2)}\| = \mathrm{O}(1/n)$ by Assumption 1. Thus, $\hat{\beta}_n$ is consistent.

Using (2.3) and consistency of $\hat{\lambda}_n$ and $\hat{\beta}_n$, we have

$$\begin{aligned}
\hat{\sigma}_n^2(\hat{\lambda}_n) = \frac{1}{n}\left\|\Delta_n\left(\lambda_0 + \mathrm{o}_{\mathbb{P}}^{d\times 1}(1)\right) Y_n - X_n\left(\beta_0 + \mathrm{o}_{\mathbb{P}}^{k\times 1}(1)\right)\right\|^2&\\
= \frac{1}{n}\left\|\Delta_n(\lambda_0) Y_n - X_n\beta_0 - \left(\mathrm{o}_{\mathbb{P}}^{d\times 1}(1)\right)^{\mathrm{T}} \mathbf{W}_n Y_n - X_n \mathrm{o}_{\mathbb{P}}^{k\times 1}(1)\right\|^2&.
\end{aligned}$$

Similar arguments imply that $\frac{1}{n}\left\|\left(\mathrm{o}_{\mathbb{P}}^{d\times 1}(1)\right)^{\mathrm{T}} \mathbf{W}_n Y_n + X_n \mathrm{o}_{\mathbb{P}}^{k\times 1}(1)\right\|^2 = \mathrm{o}_{\mathbb{P}}(1)$. As $\operatorname{plim}_{n\to\infty} \frac{1}{n}\|\varepsilon_n\|^2 = \sigma_0^2$, by statement **(b)** of Lemma B.2, we also have $\operatorname{plim}_{n\to\infty} \hat{\sigma}_n^2 = \sigma_0^2$. □

2.4 Asymptotic normality of $\sqrt{n} \cdot \hat{\theta}_n$

In order to establish the asymptotic distribution of the QML estimator $\hat{\theta}_n$ we need to adopt a number of additional assumptions. First, however, let us introduce the following definition.

Definition 1. Let $\Xi \subset \prod_{n=1}^\infty \mathbb{R}^n$ be the linear space[11] of all sequences $(x_n)_{n\in\mathbb{N}}$, with $x_n = (x_{n,i})_{i\leq n} \in \mathbb{R}^n$, $n \in \mathbb{N}$, for which $\max_{i\leq n} x_{n,i}^2 = \mathrm{o}(n)$.

[11] Naturally, the set $\prod_{n=1}^\infty \mathbb{R}^n = \mathbb{R} \times \mathbb{R}^2 \times \ldots$ is a vector space when endowed with element-wise addition. Then, Ξ is its linear subspace.

Assumption 1' Assumption 1 holds and each column of the matrices X_n, $W_{n,r}\Delta_n(\lambda_0)^{-1}X_n\beta_0$, $r \le d$, is a member of the linear space Ξ.

Assumption 2' For each $n \in \mathbb{N}$ the components $\varepsilon_{n,i}$, $1 \le i \le n$, of the error term are independent random variables, with $\mathbb{E}\,\varepsilon_{n,i} = 0$, $\mathbb{E}\,\varepsilon_{n,i}^2 = \sigma_0^2 > 0$. Moreover, the family of random variables $\varepsilon_{n,i}^4$, $n \in \mathbb{N}$, $i \le n$, is uniformly integrable.

Assumption 6 Let θ_0 be the true value of parameter θ. For the matrices $\mathfrak{I}_n = -\mathbb{E}_{\theta_0} \frac{1}{n}\frac{\partial^2 \ln L_n}{\partial \theta^2}(\theta_0)$ and $\Sigma_{\mathcal{S},n} = \mathbb{E}_{\theta_0} \frac{1}{n}\mathcal{S}_n^{\mathrm{T}}\mathcal{S}_n$, where $\mathcal{S}_n = \frac{\partial \ln L_n}{\partial \theta}(\theta_0)$, $n \in \mathbb{N}$, the following limits exist: $\mathfrak{I} = \lim_{n\to\infty}\mathfrak{I}_n$ and $\Sigma_{\mathcal{S}} = \lim_{n\to\infty}\Sigma_{\mathcal{S},n}$.[12] Moreover, the matrix $\mathfrak{I}$ is non-singular.

Theorem 2.2. *Let Assumptions 1', 2' and 3 – 6 hold and let $\hat{\theta}_n$ be the QML estimator described in Section 2.2. If, for an orthogonal projection* P*, we have* $\mathrm{P}\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^{\mathrm{T}} = \mathrm{o}_{\mathbb{P}}(\sqrt{n})$*, then the asymptotic distribution of* $\sqrt{n}\mathrm{P}\mathfrak{I}(\hat{\theta}_n - \theta_0)$ *is multivariate normal with zero mean and variance* $\mathrm{P}\Sigma_{\mathcal{S}}\mathrm{P}$*.*

Proof. With $\mathcal{S}_n$ defined in Assumption 6, straightforward calculation, c.f. Remark A.1, shows that the consecutive entries of $\frac{1}{\sqrt{n}}\mathcal{S}_n$ are

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial \ln L_n}{\partial \beta}(\theta_0) &= \frac{1}{\sqrt{n}\sigma_0^2}\varepsilon_n^{\mathrm{T}}X_n, \\
\frac{1}{\sqrt{n}}\frac{\partial \ln L_n}{\partial \lambda}(\theta_0) &= \frac{1}{\sqrt{n}\sigma_0^2}\Big(\left[\varepsilon_n^{\mathrm{T}}W_{n,r}\Delta_n(\lambda_0)^{-1}X_n\beta_0\right]_{r\le d}^{\mathrm{T}} \\
&\qquad + \left[\varepsilon_n^{\mathrm{T}}W_{n,r}\Delta_n(\lambda_0)^{-1}\varepsilon_n - \sigma_0^2 \operatorname{tr}\left(W_{n,r}\Delta_n(\lambda_0)^{-1}\right)\right]_{r\le d}^{\mathrm{T}}\Big), \\
\frac{1}{\sqrt{n}}\frac{\partial \ln L_n}{\partial \sigma^2}(\theta_0) &= \frac{1}{2\sqrt{n}\sigma_0^4}\left(\varepsilon_n^{\mathrm{T}}\varepsilon_n - n\sigma_0^2\right).
\end{aligned}$$

We will show that $\frac{1}{\sqrt{n}}\mathcal{S}_n^{\mathrm{T}}$ converges in distribution to $\mathbf{N}(0, \Sigma_{\mathcal{S}})$.[13]

Let $\alpha = \left(a^{\mathrm{T}}, b^{\mathrm{T}}, c\right)^{\mathrm{T}} \in \mathbb{R}^{k+d+1}$, where $a \in \mathbb{R}^k$, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. First, assume that $\alpha^{\mathrm{T}}\Sigma_{\mathcal{S}}\alpha \ne 0$. Then, we can observe that $\frac{1}{\sqrt{n}}\alpha^{\mathrm{T}}\mathcal{S}_n^{\mathrm{T}}$ is a centred linear-quadratic form of the residual ε_n. Namely, $\frac{1}{\sqrt{n}}\alpha^{\mathrm{T}}\mathcal{S}_n^{\mathrm{T}} = Q_n - \mathbb{E}\,Q_n$, with $Q_n = x_n^{\mathrm{T}}\varepsilon_n + \varepsilon_n^{\mathrm{T}}A_n\varepsilon_n$, where $x_n = \frac{1}{\sqrt{n}\sigma_0^2}\left(X_n a + b^{\mathrm{T}}\mathbf{W}_n\Delta_n(\lambda_0)^{-1}X_n\beta_0\right)$ and $A_n = \frac{1}{\sqrt{n}\sigma_0^2}\left(\frac{c}{2\sigma_0^2}\mathrm{I}_n + b^{\mathrm{T}}\mathbf{W}_n\Delta_n(\lambda_0)^{-1}\right)$. Note that by Assumptions 1' and 6 we have $\max_{i\le n} x_{n,i}^2 = \mathrm{o}(1)$, $\|A_n\|^2 = \mathrm{O}\left(1/n\right)$, $\|x_n\|^2 + \|A_n\|_{\mathrm{F}}^2 = \mathrm{O}(1)$; and, using Assumption 6, we have $\lim_{n\to\infty}\operatorname{Var}\frac{1}{\sqrt{n}}\alpha^{\mathrm{T}}\mathcal{S}_n^{\mathrm{T}} > 0$. Thus, Theorem C.1 can be used to deduce that $\frac{1}{\sqrt{n}}\alpha^{\mathrm{T}}\mathcal{S}_n^{\mathrm{T}}$ converges in distribution to $\mathbf{N}\left(0, \alpha^{\mathrm{T}}\Sigma_{\mathcal{S}}\alpha\right)$. In the case when $\alpha^{\mathrm{T}}\Sigma_{\mathcal{S}}\alpha = 0$, the convergence holds trivially.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which ε_n, $n \in \mathbb{N}$, are defined. Let τ be the open set considered in Remark A.4 and let B^{λ_0} be an open ball centred at λ_0 contained entirely in U_Λ. Set $U_{\theta_0} = \tau \cap \left(\mathbb{R}^d \times B^{\lambda_0} \times (0, +\infty)\right)$. Also denote $\tilde{\mathfrak{I}}_n = -\frac{1}{n}\frac{\partial^2 \ln L_n}{\partial \theta^2}(\theta_0)$ and $M_n^\tau = \sup_{\tilde{\theta}\in\tau}\left\|\frac{1}{n}\frac{\partial^3 \ln L_n}{\partial \theta^3}(\tilde{\theta})\right\|$.[14] By Theorem 2.1 we have $\|\hat{\theta}_n - \theta_0\| = \mathrm{o}_{\mathbb{P}}(1)$, Remark A.4 yields $\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_0} M_n^\tau < \infty$ and by Remark A.3 it follows that $\mathbb{P}(\{\det\tilde{\mathfrak{I}}_n = 0\}) = \mathrm{o}(1)$ and $\|\tilde{\mathfrak{I}}_n^{-1}\| =$

[12] By Remark A.1, for each $n \in \mathbb{N}$, the derivative can be properly defined on the universal open set $U_\Lambda \supset \Lambda$. The pseudo-score $\mathcal{S}_n$ is treated as a row vector.

[13] Naturally, it is not sufficient to establish asymptotic normality of the above formulae, c.f. Lee, L. F. (2004). Our argument follows by considering two cases and makes use of the standard Cramér-Wald theorem (see e.g. Billingsley, P. (1995)).

[14] The symbol $\|\cdot\|$ the natural induced operator norm. However, as the matrix is of finite dimension, all norms are equivalent, so bounds for M_n^τ differ only by a factor. Also notice that M_n^τ is measurable, c.f. Remark A.4.

$O_{\mathbb{P}}(1)$. Thus, we can conclude that for the sets

$$\Omega_n = \left\{\hat{\theta}_n \in U_{\theta_0}\right\} \cap \left\{\det \tilde{\mathfrak{I}}_n \neq 0 \text{ and } M_n^\tau \|\tilde{\mathfrak{I}}_n^{-1}\| \|\hat{\theta}_n - \theta_0\| < 1\right\},$$

we have $\lim_{n\to\infty} \mathbb{P}(\Omega_n) = 1$.

For any $\omega \in \Omega_n$, by the Taylor expansion theorem, see e.g. Theorem 107 in Hájek, P. and Johanis, M. (2014), applied for the function $f_{\omega,n}(\theta) = \frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \theta} (\theta, \omega)^{\mathrm{T}}$ in $\theta = \theta_0$ we have

$$f_{\omega,n}(\theta) = \frac{1}{\sqrt{n}} \mathcal{S}_n^{\mathrm{T}} - \tilde{\mathfrak{I}}_n \cdot \left(\sqrt{n}(\theta - \theta_0)\right) + \mathcal{R}_n(\theta), \quad \theta \in U_{\theta_0},$$

where $\mathcal{R}_n$ is the expansion remainder satisfying

$$\|\mathcal{R}_n(\theta)\| \leq \frac{1}{2} \sup_{\tilde{\theta} \in U_{\theta_0}} \|f''_{\omega,n}(\tilde{\theta})\| \cdot \|\theta - \theta_0\|^2.$$

Substituting $\theta = \hat{\theta}_n(\omega)$ we obtain

$$\tilde{\mathfrak{I}}_n \cdot \left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right) = \frac{1}{\sqrt{n}} \mathcal{S}_n^{\mathrm{T}} - f_{\omega,n}(\hat{\theta}_n) + \mathcal{R}_n(\hat{\theta}_n) \tag{2.7}$$

and

$$\|\mathcal{R}_n(\hat{\theta}_n)\| \leq \frac{\sqrt{n}}{2} M_n^\tau \|\hat{\theta}_n - \theta_0\|^2. \tag{2.8}$$

With $a_n = \tilde{\mathfrak{I}}_n \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ and $b_n = \frac{1}{\sqrt{n}} \mathcal{S}_n^{\mathrm{T}} - f_{\omega,n}(\hat{\theta}_n)$, the crucial observation is that $\|a_n\| < 2\,\|b_n\|$ on the sets Ω_n. Indeed, otherwise we would have

$$\begin{aligned} \|a_n\| \leq 2\|a_n\| - 2\|b_n\| \leq 2\|a_n - b_n\| \leq 2\|\mathcal{R}_n(\theta)\| \\ \leq M_n^\tau \|\tilde{\mathfrak{I}}_n^{-1}\| \|\hat{\theta}_n - \theta_0\| \cdot \|a_n\| < \|a_n\|. \end{aligned} \tag{2.9}$$

Finally, by (2.8) and Remark A.5, we conclude that

$$\|\mathcal{R}_n(\hat{\theta}_n)\| \leq M_n^\tau \|\tilde{\mathfrak{I}}_n^{-1}\| \left\| \frac{1}{\sqrt{n}} \mathcal{S}_n^{\mathrm{T}} - f_{\omega,n}(\hat{\theta}_n) \right\| \cdot \|\hat{\theta}_n - \theta_0\| = o_{\mathbb{P}}(1). \tag{2.10}$$

Thus, by (2.7), $\mathrm{P}\tilde{\mathfrak{I}}_n \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1}{\sqrt{n}} \mathrm{P}\mathcal{S}_n^{\mathrm{T}} + o_{\mathbb{P}}(1)$. Using Remark A.3, the desired convergence in distribution follows. □

Note that under the assumption that $\lambda_0 \in \mathrm{Int}_{\mathbb{R}^d} \Lambda$ we have $\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n) = o_{\mathbb{P}}(1)$. Thus, as a special case of Theorem 2.2, we immediately obtain the following generalization of Theorem 3.2 in Lee, L. F. (2004).

Theorem 2.3. *Under* Assumptions *1', 2' and 3 – 6, for the QML estimator $\hat{\theta}_n$ described in* Section *2.2, the asymptotic distribution of $\sqrt{n} \cdot (\hat{\theta}_n - \theta_0)$ is multivariate normal with zero mean and variance $\mathfrak{I}^{-1}\Sigma_{\mathcal{S}}\mathfrak{I}^{-1}$. Furthermore, if the error term is normally distributed, then the limiting distribution is* $\mathrm{N}(0, \mathfrak{I}^{-1})$.

Additionally, Theorem 2.2 has further useful implications. Firstly, the case of the parameter space Λ locally having an empty interior can be considered. Secondly, we can address the question of the asymptotic distribution[15] of $\hat{\theta}_n$ when the true value of λ lies strictly on the boundary of Λ. The following Assumptions 7a and 7b suggest how the projection P can be chosen in two, somewhat simplified, yet illustrative cases.

[15]Let us note that although Theorem 2.2 does not explicitly provide asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ itself, inferences can be based on the distributions of coordinates of $\sqrt{n}(\mathrm{P}\mathfrak{I})^+\mathrm{P}\mathfrak{I}(\hat{\theta}_n - \theta_0)$. Here, + denotes the Moore-Penrose pseudo-inverse.

Assumption 7a Let λ_0 be the true value of parameter λ. There exist an orthogonal projection P_a in $\mathbb{R}^d$ and a number $\epsilon > 0$ for which

$$B_\epsilon \cap (\Lambda - \lambda_0) = B_\epsilon \cap \operatorname{range}(\mathrm{P}_\mathrm{a}),$$

with $B_\epsilon = \{\lambda \in \mathbb{R}^d \colon \|\lambda\| < \epsilon\}$.

Assumption 7b Let λ_0 be the true value of parameter λ. There exist a vector $\rho \in \mathbb{R}^d$ and a number $\epsilon > 0$ for which

$$B_\epsilon \cap (\Lambda - \lambda_0) = B_\epsilon \cap \{\lambda \in \mathbb{R}^d \colon \rho^\mathrm{T} \lambda \geq 0\},$$

with $B_\epsilon = \{\lambda \in \mathbb{R}^d \colon \|\lambda\| < \epsilon\}$. We set $\mathrm{P}_\mathrm{b} = \mathrm{I}_d - \rho(\rho^\mathrm{T}\rho)^{-1}\rho^\mathrm{T}$.

Remark 2.2 Under Assumption 7a we have $\tilde{\mathrm{P}}_\mathrm{a} \frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = \mathrm{o}_\mathbb{P}(1)$, with $\tilde{\mathrm{P}}_\mathrm{a} = \mathrm{I}_k \oplus \mathrm{P}_\mathrm{a} \oplus \mathrm{I}_1$.[16] Similarly, under Assumption 7b we have $\tilde{\mathrm{P}}_\mathrm{b} \frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = \mathrm{o}_\mathbb{P}(1)$, with $\tilde{\mathrm{P}}_\mathrm{b} = \mathrm{I}_k \oplus \mathrm{P}_\mathrm{b} \oplus \mathrm{I}_1$.

Proof. First, let Assumption 7b hold. Assume $\hat{\lambda}_n \in B_\epsilon$. It is enough to consider two cases. If $\rho^\mathrm{T}\hat{\lambda}_n > 0$, then $\hat{\lambda}_n$ lies in the interior of Λ, thus $\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = 0$. If $\rho^\mathrm{T}\hat{\lambda}_n = 0$, then $\hat{\theta}_n$ maximises $\ln L_n(\lambda, \beta, \sigma^2)$ subject to $\rho^\mathrm{T}\lambda = 0$, thus $\tilde{\mathrm{P}}_\mathrm{b} \frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = 0$. In any case we have $\tilde{\mathrm{P}}_\mathrm{b} \frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = \mathrm{o}_\mathbb{P}(1)$.

Similarly, under Assumption 7a, $\hat{\theta}_n$ maximises $\ln L_n(\lambda, \beta, \sigma^2)$, subject to $\mathrm{P}_\mathrm{a}\lambda = \lambda$, whenever $\hat{\lambda}_n \in B_\epsilon$. This implies that $\tilde{\mathrm{P}}_\mathrm{a} \frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)^\mathrm{T} = \mathrm{o}_\mathbb{P}(1)$. □

A natural extension of this argument allows the consideration of a continuously differentiable function F for which the equation $F(\lambda) = 0$ defines the parameter space Λ. In such case, Assumption 7b can be used through the Lagrange multipliers theorem. For example, assume that $F'(\lambda_0) \neq 0$ and set $\rho_\lambda = F'(\lambda)^\mathrm{T} \in R^d$, for any $\lambda \in \Lambda$. Then we can set $\mathrm{P} = \mathrm{P}_{\lambda_0}$ in Theorem 2.2, with $\mathrm{P}_\lambda = \mathrm{I}_k \oplus (\mathrm{I}_d - \|\rho_\lambda\|^{-1}\rho_\lambda\rho_\lambda^\mathrm{T}) \oplus \mathrm{I}_1$. Similarly, if Λ is a closure of some sufficiently regular open subset of $\mathbb{R}^d$ and λ_0 lies on the boundary of Λ, then Assumption 7b can be used by approximating ρ with a vector orthogonal to Λ at λ_0. More precisely, define $\rho_\lambda \in \mathbb{R}^d$ to be a non-zero vector orthogonal to the hyperplane tangent to Λ at $\lambda \in \Lambda \setminus \operatorname{Int}\Lambda$. In either case, the crucial observation is that $\|\mathrm{P}\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)\| \leq \|\mathrm{P} - \mathrm{P}_{\hat{\lambda}_n}\| \cdot \|\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n)\| + \mathrm{o}_\mathbb{P}(1)$. Finally, Remark A.5 and continuity of $\lambda \mapsto \mathrm{P}_\lambda$ yield the convergence $\mathrm{P}\frac{\partial \ln L_n}{\partial \theta}(\hat{\theta}_n) = \mathrm{o}_\mathbb{P}(1)$.

2.5 Discussion of the adopted assumptions

This section discusses the assumptions adopted in Sections 2.1–2.4. They will be considered in order of appearance, thus we first turn to Assumption 1.

In this paper it is assumed that the matrix of explanatory variables is non-stochastic. Nonetheless, some extensions to allow X_n to be random are possible. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which ε_n, $n \in \mathbb{N}$, are defined. One possible idea, which we briefly outline, is to ensure that for each $n \in \mathbb{N}$ the requirements imposed on X_n are satisfied on a set A_n such that $\mathbb{P}(\Omega \setminus A_n) = \mathrm{o}(1)$. If we further suppose that assumptions made on ε_n are satisfied conditionally on X_n,[17] then our theory implies that the Gaussian QML estimator $\hat{\theta}_n$, described in Section 2.2, is consistent. Indeed, by Theorem 2.1 we have $\lim_{n\to\infty} \mathbb{P}\left[\|\hat{\theta}_n - \theta_0\| > \delta \mid X_n\right] = 0$ almost surely, for any $\delta > 0$. Then, by Lebesgue's dominated convergence theorem $\hat{\theta}_n$ is consistent.

[16] The symbol $\oplus$ denotes the matrix direct sum, i.e. $A \oplus B$ is the block diagonal matrix with consecutive blocks A, B.

[17] In particular, $\mathbb{E}[\varepsilon_n \mid X_n] = 0$ and $\mathbb{E}\left[\varepsilon_n\varepsilon_n^\mathrm{T} \mid X_n\right] = \sigma^2\mathrm{I}_n$. Also note that X_n does not need to be independent of ε_n, cf. Assumption E2 in Shi, W. and Lee, L. F. (2017).

However, establishing asymptotic distribution of $\hat{\theta}_n$ is, in general, possible only conditionally on X_n.

Let us note that Assumption 1, used for the consistency argument, does not require the sequence $\frac{1}{n}X_n^{\mathrm{T}}X_n$ to be convergent. Instead, our reasoning stipulates that this sequence is merely bounded. This does not imply that elements of X_n should be bounded in absolute value, as it is explicitly assumed in e.g. Lee, L. F. (2004) or Lee, L. F. and Yu, J. (2010a). The necessity of non-singularity of $\left(X_n^{\mathrm{T}}X_n\right)$ is straightforward, as regressors should not be correlated. Furthermore, it is also natural to require that the regressor values are "not too small" asymptotically, so that they can provide a sufficient amount of information about the slope parameter in the model. Note that our assumption that $\left\|(X_n^{\mathrm{T}}X_n)^{-1}\right\| = \mathrm{O}(\frac{1}{n})$ is not far from the well-known condition necessary for consistency of OLS estimator for non-spatial regression, i.e. $\|(X_n^{\mathrm{T}}X_n)^{-1}\| = \mathrm{o}(1)$.

For the sake of simplicity of presentation, Assumption 2 stipulates for a homoskedastic error term. However, in the case of heteroskedastic innovations, similar results can be obtained by applying a variance normalising transformation. See also Liu S. F. and Yang Z. (2015), where an idea for handling some types of unknown heteroskedasticity in QML estimation is presented. Note that elements of the error term do not need to be identically distributed nor fully independent. However, the error term is required to be a projected unitary transformation of a vector of quadruple independent variables, thus, in general, its components are merely uncorrelated. The distinction between independence and absence of correlation is all the more relevant as the innovations are not assumed to be Gaussian.

Interpretation of Assumption 3 is straightforward. The invertibility of the spatial difference operator $\Delta_n(\lambda)$ allows one to find the closed form of the spatial lag model (2.1) with respect to the dependant variable. This, among other things, facilitates interpretation of model parameters, in particular, by the use of direct and indirect effects, as defined in Le Sage, J. and Pace, R. K. (2009), page 74.

Assumption 4 spells out two important boundedness requirements. Firstly, it stipulates for the parameter space Λ to be a compact subset of $\mathbb{R}^d$. Note that the space does not have to be connected nor have a dense interior. This implies that our analysis provides better grounds for specifications of Λ in which singularity points of $\Delta(\lambda)$ are geometrically avoided. Moreover, let us consider a situation in which the model specification includes a restriction, possibly nonlinear, on regressive parameters $\lambda_1, \ldots, \lambda_d$, given by an equation of the form $F(\lambda_1, \ldots, \lambda_d) = 0$, for some function $F\colon \Lambda \to \mathbb{R}$. Such restriction is naturally equivalent to limiting the parameter space to the set of roots of F, which typically is a $(d-1)$-dimensional hypersurface embedded in Λ. The arguments of the standard asymptotic analysis are null in this setting, as they require the true value $\lambda_0 = (\lambda_1^0, \ldots, \lambda_d^0)$ to be an internal point of Λ. That is, $\lambda_0 \in \mathrm{Int}_{\mathbb{R}^d}\Lambda$, however $\mathrm{Int}_{\mathbb{R}^d}\Lambda = \emptyset$, c.f. Lee, L. F. (2004) or Li, K. (2017). Our theorems, on the other hand, immediately imply that the Gaussian QML estimator for those models is consistent. With some additional provisions, it is also asymptotically normal.

Secondly, Assumption 4 gives the crucial condition imposed on the spatial weight matrix $\mathbf{W}_n = \langle W_{n,1}, \ldots, W_{n,d}\rangle^{\mathrm{T}}$ in order to make it eligible for our asymptotic analysis. This paper stipulates that the following set of matrices $W_{n,r}$, $r \le d$, $\Delta_n(\lambda)^{-1}$, $\lambda \in \Lambda$, is bounded in the spectral norm, rather than Kelejian's norm[18] $\|\cdot\|_{\mathrm{K}} = \max\{\|\cdot\|_1, \|\cdot\|_\infty\}$. Let us note that any set of square matrices which is bounded in $\|\cdot\|_{\mathrm{K}}$-norm is also bounded in the spectral norm, as follows from Theorem C.2. That is to say, the asymptotic theory presented in this paper is indeed a generalization of the theory initiated in Lee, L. F. (2004). Moreover, it is also a proper generalization. To justify this statement it is enough to construct a spatial weight matrix W_n, $n \in \mathbb{N}$, for which $\sup_{n\in\mathbb{N}} \|W_n\| < \infty$ and $\sup_{n\in\mathbb{N}} \|W_n\|_{\mathrm{K}} = \infty$. It is quite

[18] We use this name to indicate the norm's central role in the CLT by Kelejian, H. H. and Prucha, I. R. (2001).

straightforward to give an example of such a matrix, by ensuring non-summability of one of its columns. A construction of a non-summable spatial weight matrix which cannot be easily fixed by eliminating a finite number of spatial units, is presented in Section 2.6.

Assumption 5 spells out the conditions imposed on spatial weight matrix $\mathbf{W}_n$ used to obtain consistency of the Gaussian QML estimator $\hat{\theta}_n$. It should be noted that this assumption is generally stronger than mere identification of λ. In fact, it implies that the Gaussian log-likelihood (2.2), (2.4) is asymptotically able to discriminate between different values of λ. In other words, Assumption 5 ensures that there is enough information in the observed process to decrease the estimate uncertainty for $\hat{\theta}_n$, with increasing n. The distinction between the statements **(a)** and **(b)** reflects the fact that this information can come from either the spatial autocorrelation of Y_n or via the accumulated spatial lag of regressors.

Assumption 1' extends Assumption 1 with requirements necessary for obtaining limiting distribution of the deviation $d_n = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$. Intuitively, the limiting distribution can be normal, regardless of the original distribution of ε_n only when none of the observations within matrices X_n and $W_{n,r}\Delta_n(\lambda_0)^{-1}X_n\beta_0$, $r \leq d$, makes an overwhelming contribution to the estimate of the corresponding slope coefficient. Let us note that this assumption is also necessary in the simple case of non-spatial least squares regression. Although implicitly, this assumption is also present in Lee, L. F. (2004) and Lee, L. F. and Yu, J. (2010a), as it is a consequence of other assumptions adopted therein (in particular, boundedness of elements of X_n).

It is known that for derivation of asymptotic distribution of d_n the conditions expressed in Assumption 2 are not sufficient, see for example Pruss, A. R. (1998). Therefore, in our Assumption 2' we adopt the standard econometric postulate that the innovations are stochastically independent within samples. Note that we still do not assume that the elements of the error term follow the same distribution. Instead, we impose the requirement of uniform integrability of the fourth powers of all components of the error terms.

Assumption 6 spells out the necessary conditions for existence of the limiting distribution variance. Note that for consistency of $\hat{\theta}_n$ the sequences $(\mathfrak{I}_n)_{n\in\mathbb{N}}$, $(\Sigma_{\mathcal{S},n})_{n\in\mathbb{N}}$ do not need to converge, let alone their limits be invertible. The requirement of invertibility of the matrix $\mathfrak{I}$ could be relaxed. However, with the present argument, it is possible to obtain only partial results on asymptotic distribution of d_n. An approach to the problem of singularity of $\mathfrak{I}$ which considers various convergence rates has been described in Lee, L. F. (2004).

2.6 Novel spatial interaction patterns – an example

In this section we formally show that the class of spatial weight matrices covered by our asymptotic analysis is a proper superset of the class of matrices absolutely summable in rows and columns. Indeed, in view of Theorem C.2, boundedness in $\|\cdot\|_{\mathrm{K}}$ norm implies boundedness in the spectral norm. Below, we construct an example of a row-standardized spatial weight matrix W_n which is bounded in the spectral norm and in which no column is absolutely summable. Since a row-standardized matrix is naturally bounded in row sums, our construction relies on its column sums being unbounded. As a result, the spatial interaction pattern described by our matrix allows spatial units to influence a possibly unbounded number of neighbours.[19] To maintain simplicity of the argument we allow each spatial unit to be itself affected by a limited number of units.

Let us define sets $B_1 = \{2\}$, $B_2 = \{3,4\}$, $B_3 = \{5,6,7\}$ and further $B_k = \{\max B_{k-1} + l\}_{1\leq l\leq k}$, for $k \geq 4$. Clearly, sets B_k, $k \geq 1$, are mutually disjoint and $\bigcup_{k=1}^{\infty} B_k = \mathbb{N} \setminus \{1\}$. Thus,

[19] We should note that a modification to our construction is possible to make each spatial unit influence only a finite number of neighbours.

each number $i \geq 2$ uniquely determines a pair $(k(i), l(i))$ defined by $i \in B_{k(i)}$ and $l(i) = i - \min B_{k(i)} + 1$. In other words, $l(i)$ is the ordinal number of i in the increasing sequence of elements of $B_{k(i)}$. Note that $i > l(i)$, for any $i \geq 2$. Let us define an infinite matrix $\tilde{W}$ with all elements equal to zero except for $\tilde{W}_{1,2} = 1$ and $\tilde{W}_{i,l(i)} = \frac{1}{k(i)}$, $\tilde{W}_{i,i+1} = 1 - \frac{1}{k(i)}$, for all $i \geq 2$.

Firstly, no column of $\tilde{W}$ is absolutely summable. To justify this fact observe that if $j \geq 1$ is a column number and $k \geq j$, then there exists an $i \in B_k$ such that $j = l(i)$ and $\tilde{W}_{ij} = \frac{1}{k}$. Thus $\sum_{i=1}^{\infty} \tilde{W}_{ij} \geq \sum_{k=j}^{\infty} \frac{1}{k} = \infty$. Secondly, $\|\tilde{W}\| \leq 1 + \frac{\pi}{\sqrt{6}}$.[20] Indeed, with $\tilde{W}^{\mathrm{U}} = \left[\tilde{W}_{ij}\mathbb{I}_{\{i<j\}}\right]_{i,j\leq n}$ and $\tilde{W}^{\mathrm{L}} = \left[\tilde{W}_{ij}\mathbb{I}_{\{i>j\}}\right]_{i,j\leq n}$ we can write $\tilde{W} = \tilde{W}^{\mathrm{U}} + \tilde{W}^{\mathrm{L}}$. Then, we have $\|\tilde{W}^{\mathrm{U}}\| \leq \|\tilde{W}^{\mathrm{U}}\|_{\mathrm{K}} = 1$, as an indirect consequence Theorem C.2. Next, let us denote the columns of $\tilde{W}^{\mathrm{L}}$ with c_j, $j \geq 1$. The vectors $c_j \in \mathfrak{l}_2$ are orthogonal, as for each c_j, the set of indices of its non-zero elements is $l^{-1}(\{j\})$, and those sets are mutually disjoint for $j \geq 0$. Moreover, $\|c_j\|^2 = \sum_{k=j}^{\infty} \frac{1}{k^2} \leq \frac{\pi^2}{6}$. For any $x \in \mathfrak{l}_2$, by Bessel's inequality, we have $\|(\tilde{W}^{\mathrm{L}})^{\mathrm{T}} x\|^2 = \sum_{j=1}^{\infty} |c_j^{\mathrm{T}} x|^2 \leq \frac{\pi^2}{6}\|x\|^2$. Since $\|\tilde{W}^{\mathrm{L}}\| = \|(\tilde{W}^{\mathrm{L}})^{\mathrm{T}}\|$, this yields $\|\tilde{W}^{\mathrm{L}}\| \leq \frac{\pi}{\sqrt{6}}$.

Finally, we will define the matrix $W_n = [W_{n;ij}]_{i,j\leq n}$, $n \in \mathbb{N}$. Namely, let us set $W_{n;ij} = \tilde{W}_{ij}$, for $i, j \leq n$, with the exception of $W_{n;n,n-1} = 1 - \frac{1}{k(n)}$. Notice that $\|W_n\| \leq \|\tilde{W}\| + 1$. Moreover, for $n \geq 4$ the matrix W_n is row-standardized. Indeed, we have $\sum_{j=1}^{n} W_{n,1j} = W_{1,2} = 1$ and for any $2 \leq i \leq n$ we have $\sum_{j=1}^{n} W_{n,ij} = \frac{1}{k(i)} + 1 - \frac{1}{k(i)} = 1$. This implies that the non-singularity set for all $\Delta_n(\lambda) = \mathrm{I}_n - \lambda W_n$, $n \in \mathbb{N}$, is non-trivial, as it contains, at least, the interval $(-1, 1)$.

We anticipate that spatial interaction patterns which are not necessarily absolutely summable will be particularly useful in the case of the infill asymptotics, where spatial interaction with asymptotically increasing number of neighbours is more natural. However, an application to the case of increasing domain asymptotics is also possible. In particular, the argument of this section may be related to a class of inverted distance spatial weight matrices, as it relies on the divergence of the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$. Assume that the strength of the potential interaction of units i and j is proportional to $\mathrm{dist}\,(i,j)^{-\gamma}$, for some $\gamma > 0$. Let $N(j, \delta)$ denote the number of neighbours i influenced by unit j such that $\mathrm{dist}\,(i,j) \approx \delta$. When $N(j,\delta) \geq \mathrm{const} \cdot \delta^{\gamma-1}$, the columns sum for region j in the implied spatial weight matrix is asymptotically at least $\lim_{\delta\to\infty} \int_0^{\delta} N(j,\tilde{\delta})\tilde{\delta}^{-\gamma}\,\mathrm{d}\tilde{\delta} = \infty$. Thus, the standard asymptotic analysis of Lee, L. F. (2004) cannot be applied. However, if, at the same time $N(j,\delta) = \mathrm{O}(\delta^{2\gamma-1-\epsilon})$, for some $\varepsilon > 0$, then this interaction pattern may still lead to a spatial weight matrix which is bounded in the spectral norm, as $\lim_{\delta\to\infty} \int_0^{\delta} N(j,\tilde{\delta})(\tilde{\delta}^{-\gamma})^2\,\mathrm{d}\tilde{\delta} < \infty$.

3 Application to higher-order general fixed effects model

3.1 The elimination technique

This section provides a theoretical illustration of the utility of our improved asymptotic analysis. First, we describe a new fixed effect elimination scheme for the high-order SAR model. Then, from the theorems of Section 2, we derive statements on the large sample behaviour of the resulting QML estimator. We should note that in simple cases of constant number of fixed effect dummies that distinguish non-overlapping groups of observations, a consistent, asymptotically normal QML estimator can be obtained by concentrating out the nuisance parameters. However, in general case a careful approach is necessary. The argument presented below gener-

[20] Spectral (induced) norm is also defined for infinite matrices, which are understood as operators on $\mathfrak{l}_2$ – the Hilbert space of all square summable infinite sequences, equipped with the natural inner product, i.e. the sum of products of corresponding coordinates.

alises that developed in Olejnik, A. and Olejnik, J. (2017) and also extends the ideas for fixed effect elimination described in Lee, L. F. and Yu, J. (2010a,b) and Lee, L. F. *et. al.* (2010). In those papers the asymptotic analysis of QML estimators relies on the fact that the demeaning operator matrix is row and column absolutely summable. By using the virtues of our improved analysis we are able to account for a wider class of demeaning operators. Namely, we consider model transformations which are bounded in the spectral norm, rather than in the $\|\cdot\|_{\mathrm{K}}$-norm; and any non-zero projection matrix has unit spectral norm. We also make use of the fact that the requirements expressed in our Assumption 4, unlike in the standard analysis of Lee, L. F. (2004), are invariant under unitary transformations of the spatial weight matrix.

Let us consider a modified version of the SAR model specification (2.1) in which an additional term of fixed effects is introduced. Namely, we now turn to the following specification

$$Y_n = \lambda^{\mathrm{T}}\mathbf{W}_n Y_n + X_n\beta + \Phi_n\mu + \varepsilon_n, \tag{3.1}$$

where ε_n satisfies Assumption 2', the term Φ_n is a matrix of κ fixed effects and $\mu \in \mathbb{R}^\kappa$ is the corresponding vector parameter. The number of columns in Φ_n is allowed to change with sample size i.e. $\kappa = \kappa(n)$. Although in typical applications the columns of Φ_n are dummy variables distinguishing non-overlapping groups of observations, here, no such formal requirement is imposed.[21] In applied spatial econometrics it is common to eliminate fixed effects by means of the demeaning procedure, see e.g. Elhorst, J. P. (2014). This approach can be understood as a simple projection on the space spanned by the columns of Φ_n, thus, it is closely related to the famous Frisch-Waugh theorem, see Baltagi, B. H. (2005). Our technique extends this idea in two crucial aspects. Firstly, the fixed effect term is eliminated together with its higher-order spatial lags. Secondly, the transformed model is further projected onto a lower dimensional space,[22] in order to avoid the concerns regarding singularity of the resulting variance, as expressed in Anselin, L. *et. al.* (2006), page 641.

Let $\mathcal{K}_n = \mathcal{K}_n(\Phi_n) \subset \mathbb{R}^n$ be the Krylov subspace generated by iterating $\mathbf{W}_n$ on the columns of the matrix Φ_n. That is, $\mathcal{K}_n$ is the smallest subspace $H \subset \mathbb{R}^n$ satisfying $\Phi_n\alpha \in H$, for any $\alpha \in \mathbb{R}^\kappa$, and $W_{n,r}h \in H$, for any $h \in H$, $r \leq d$. In still other words, $\mathcal{K}_n$ is the smallest $\mathbf{W}_n$-invariant subspace containing columns of Φ_n. Our idea is to filter out those vector components of both Y_n and X_n which lie in $\mathcal{K}_n$. Under the assumption that the Krylov space $\mathcal{K}_n$ is sufficiently small or, equivalently, its orthogonal complement $\mathcal{K}_n^\perp$ is sufficiently rich, we can obtain a consistent QML estimator of $\theta = \left(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \sigma^2\right)^{\mathrm{T}}$.

Let $n_* = n - \dim \mathcal{K}_n$ and fix an $n_* \times n$ matrix $\mathbf{F} = \mathbf{F}_n$ whose rows form an orthonormal basis of $\mathcal{K}_n^\perp$. It is easy to observe that $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is the projection onto $\mathcal{K}_n^\perp$ and $\mathbf{F}\mathbf{F}^{\mathrm{T}} = \mathrm{I}_{n_*}$. Moreover, we have $\|\mathbf{F}\| = 1$, whenever $n_* > 0$.

Note that, by definition, $\mathbf{F}\mathcal{K}_n = \{0\}$ and, in particular, $\mathbf{F}\Phi_n = 0$. Moreover, as $\mathrm{I}_n - \mathbf{F}^{\mathrm{T}}\mathbf{F}$ projects onto $\mathcal{K}_n$, we have

$$\mathbf{F}\lambda^{\mathrm{T}}\mathbf{W}_n\left(\mathrm{I}_n - \mathbf{F}^{\mathrm{T}}\mathbf{F}\right) = 0, \quad \text{for any } \lambda \in \mathbb{R}^d. \tag{3.2}$$

Denote $Y_n^* = \mathbf{F}Y_n$, $X_n^* = \mathbf{F}X_n$ and $\varepsilon_n^* = \mathbf{F}\varepsilon_n$. Transforming the specification (3.1), by using (3.2), we obtain

$$\begin{aligned} Y_n^* = \mathbf{F}Y_n &= \mathbf{F}\lambda^{\mathrm{T}}\mathbf{W}_n Y + \mathbf{F}X_n\beta + \mathbf{F}\Phi_n\mu + \mathbf{F}\varepsilon_n \\ &= \mathbf{F}\lambda^{\mathrm{T}}\mathbf{W}_n\mathbf{F}^{\mathrm{T}}\mathbf{F}Y + \mathbf{F}X_n\beta + \mathbf{F}\varepsilon_n \\ &= \lambda^{\mathrm{T}}\mathbf{W}_n^* Y_n^* + X_n^*\beta + \varepsilon_n^*, \end{aligned} \tag{3.3}$$

[21] Columns of Φ_n are allowed to be arbitrary vectors, as long as the relevant assumptions of this section hold. In particular, the groups of observations implied by the fixed effects design are allowed to overlap.

[22] This is very similar to the approach of Lee, L. F. and Yu, J. (2010a), therein referred to as the Helmert transformation.

where $\mathbf{W}_n^* = \left\langle W_{n,r}^* \right\rangle_{r\le d}$ with $W_{n,r}^* = \mathbf{F}W_{n,r}\mathbf{F}^{\mathrm{T}}$, $r \le d$.

It is easy to observe that ε_n^* satisfies Assumption 2 with ε_n^* substituted for ε_n and $\bar{\varepsilon}_n$ replaced with present ε_n. The crucial observation, however, is that Assumptions 3 and 4 are satisfied when $\mathbf{W}_n^*$ is substituted for $\mathbf{W}_n$. Indeed, note that with $\Delta_n^*(\lambda) = \mathrm{I}_{n_*} - \lambda^{\mathrm{T}}\mathbf{W}_n^*$ we have $\Delta_n^*(\lambda)\mathbf{F}\left(\mathrm{I}_n - \lambda^{\mathrm{T}}\mathbf{W}_n\right)^{-1}\mathbf{F}^{\mathrm{T}} = \mathrm{I}_{n_*}$, by (3.2). Thus, $\Delta_n^*(\lambda)$ is invertible and $\Delta_n^*(\lambda)^{-1} = \mathbf{F}\left(\mathrm{I}_n - \lambda^{\mathrm{T}}\mathbf{W}_n\right)^{-1}\mathbf{F}^{\mathrm{T}}$. Lastly, observe that $\|W_{n,r}^*\| \le \|W_{n,r}\|$, for each $r \le d$, and $\|\Delta_n^*(\lambda)^{-1}\| \le \|\Delta_n(\lambda)^{-1}\|$, for every $\lambda \in \Lambda$.

In order to ensure proper identification of parameters in the transformed specification (3.3) we adopt the following assumptions.

Assumption 1Φ The matrix $(X_n^*)^{\mathrm{T}}X_n^*$ is invertible and we have

$$\sup_{n\in\mathbb{N}} \left\|n_*^{-1} \cdot (X_n^*)^{\mathrm{T}}X_n^*\right\| < \infty \quad \text{and} \quad \sup_{n\in\mathbb{N}} \left\|n_* \cdot ((X_n^*)^{\mathrm{T}}X_n^*)^{-1}\right\| < \infty.$$

Assumption 5Φ For every $\lambda_1, \lambda_2 \in \Lambda$, such that $\lambda_1 \neq \lambda_2$, at least one of the statements **(a)**, **(b)** below is satisfied:

(a) $\liminf_{n\to\infty} \frac{1}{\sqrt{n_*}D_n^*(\lambda)}\|\Delta_n^*(\lambda_1)\Delta_n^*(\lambda_2)^{-1}\|_{\mathrm{F}} > 1$,

(b) $\liminf_{n\to\infty} \frac{1}{\sqrt{n_*}}\|M_{X_n^*}\Delta_n^*(\lambda_1)\Delta_n^*(\lambda_2)^{-1}X_n^*\beta\| > 0$, for every $\beta \in \mathbb{R}^k$,

where $D_n^*(\lambda) = |\det\left(\Delta_n^*(\lambda_1)\Delta_n^*(\lambda_2)^{-1}\right)|^{1/n_*}$.

Assumption 7Φ We have $\lim_{n\to\infty} n_* = \infty$.

Finally, we can apply the construction of the Gaussian QML estimator from Section 2.2 to the model specification 3.3. Namely, let us set $\hat{\beta}_n^* = \hat{\beta}_n^*(\hat{\lambda}_n^*)$ and $\hat{\sigma}_n^{2*} = \sigma_n^{2*}(\hat{\lambda}_n^*)$, where

$$\hat{\beta}_n^*(\lambda) = \left((X_n^*)^{\mathrm{T}}X_n^*\right)^{-1}(X_n^*)^{\mathrm{T}}\Delta_n^*(\lambda)Y_n^*,$$
$$\hat{\sigma}_n^{2*}(\lambda) = \frac{1}{n_*}\|\Delta_n^*(\lambda)Y_n^* - X_n^*\hat{\beta}_n^*(\lambda)\|^2, \quad \lambda \in \Lambda,$$

and $\hat{\lambda}_n^*$ is a maximiser of $\ln|\det\Delta_n^*(\lambda)| - \frac{n_*}{2}\ln\left(\hat{\sigma}_n^{2*}(\lambda)\right)$ over $\lambda \in \Lambda$. Note, that under normality of the original ε_n the estimators $\hat{\lambda}_n^*$, $\hat{\beta}_n^*$ and $\hat{\sigma}_n^{2*}$ are exact ML estimators. That is to say, they maximize the Gaussian log-likelihood function for θ given Y_n^* and X_n^*, i.e.

$$\log L_n^*(\theta) = -\frac{n_*}{2}\ln\left(2\pi\sigma^2\right) + \ln|\det\Delta_n^*(\lambda)| - \frac{1}{2\sigma^2}\left\|\Delta_n^*(\lambda)Y_n^* - X_n^*\beta\right\|^2. \tag{3.4}$$

3.2 Results on the asymptotic behaviour

The following result is an immediate consequence of Theorem 2.1 applied to the transformed specification (3.3).

Theorem 3.1. *Under* Assumptions 2', 3, 4, 1Φ, 5Φ *and* 7Φ *the QML estimators* $\hat{\lambda}_n^*$, $\hat{\beta}_n^*$ *and* $\hat{\sigma}_n^{2*}$ *are consistent estimators of* λ, β *and* σ^2 *respectively.*

Establishing asymptotic normality of the quantity $\sqrt{n_*}\left(\hat{\theta}_n^* - \theta_0\right)$ requires a slightly more delicate argument than the mere application of Theorem 2.3. The main difficulty is that the components of $\mathbf{F}\varepsilon_n$ do not have to be independent, even if the original ε_n is, unless normality of the error term is assumed, c.f. the proof of Theorem 2 in Lee, L. F. and Yu, J. (2010a). However, using the virtues of the improved asymptotic analysis, this goal can also be achieved by a straightforward argument.

Assumption 1'Φ Assumption 1Φ holds and each column of the matrices $\mathbf{F}^{\mathrm{T}}X_n^*$, $\mathbf{F}^{\mathrm{T}}W_{n,r}^*\Delta_n^*(\lambda_0)^{-1}X_n^*\beta_0$, $r \leq d$, is a member of the linear space $\Xi^* \subset \prod_{n=1}^{\infty}\mathbb{R}^n$, with $\Xi^* = \{(x_n)_{n\in\mathbb{N}} \colon x_n = (x_{n,i})_{i\leq n}$ and $\max_{n,i} x_{n,i}^2 =$

Assumption 6Φ Let θ_0 be the true value of parameter $\theta = \left(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \sigma^2\right)^{\mathrm{T}}$ in specification (3.1). For the matrices $\mathfrak{I}_n^* = -\mathbb{E}_{\theta_0}\frac{\partial^2 \ln L_n^*}{\partial\theta^2}(\theta_0)$ and $\Sigma_{\mathcal{S}^*,n} = \mathbb{E}_{\theta_0}(\mathcal{S}_n^*)^{\mathrm{T}}\mathcal{S}_n^*$, where $\mathcal{S}_n^* = \frac{\partial \ln L_n^*}{\partial\theta}(\theta_0)$, $n \in \mathbb{N}$, the following limits exist: $\mathfrak{I}^* = \lim_{n\to\infty}\frac{1}{n_*}\mathfrak{I}_n^*$ and $\Sigma_{\mathcal{S}^*} = \lim_{n\to\infty}\frac{1}{n_*}\Sigma_{\mathcal{S}^*,n}$. Moreover, the matrix $\mathfrak{I}^*$ is non-singular.

Theorem 3.2. *Under* Assumptions 1'Φ, 2', 3, 4, 5Φ, 6Φ *and* 7Φ *for the QML estimator* $\hat{\theta}_n^* = \left(\hat{\lambda}_n^{*\mathrm{T}}, \hat{\beta}_n^{*\mathrm{T}}, \hat{\sigma}_n^{2*}\right)^{\mathrm{T}}$ *described in* Section 3.1, *the asymptotic distribution of* $\sqrt{n_*}\cdot(\hat{\theta}_n - \theta_0)$ *is the multivariate normal distribution with zero mean and variance* $(\mathfrak{I}^*)^{-1}\Sigma_{\mathcal{S}^*}(\mathfrak{I}^*)^{-1}$.

Proof. The proof relies on the very same argument (with L_n substituted for L_n^*) as the proof of Theorem 2.2, up to the point where our CLT is used to deduce asymptotic normality of the linear-quadratic form $\frac{1}{\sqrt{n_*}}\mathcal{S}^*\alpha = Q_n - \mathbb{E}\,Q_n$, with α as previously, $Q_n = x_n^{\mathrm{T}}\varepsilon_n^* + (\varepsilon_n^*)^{\mathrm{T}}A_n\varepsilon_n^* = x_n^{\mathrm{T}}\mathbf{F}\varepsilon_n + (\varepsilon_n)^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}A_n\mathbf{F}\varepsilon_n$, where $x_n = \frac{1}{\sqrt{n_*}\sigma_0^2}\left(X_n^*a + b^{\mathrm{T}}\mathbf{W}_n^*\Delta_n^*(\lambda_0)^{-1}X_n^*\beta_0\right)$ and $A_n = \frac{1}{\sqrt{n_*}\sigma_0^2}\left(\frac{c}{2\sigma_0^2}\mathrm{I}_{n_*} + b^{\mathrm{T}}\mathbf{W}_n^*\Delta_n^*(\lambda_0)^{-1}\right)$. Note that by Assumptions 1'Φ and 4 we have $(\sqrt{n_*}\cdot\mathbf{F}^{\mathrm{T}}x_n)_{n\in\mathbb{N}} \in \Xi^*$, $\|\mathbf{F}^{\mathrm{T}}A_n\mathbf{F}\|^2 = \mathrm{O}(1/n_*)$, $\|\mathbf{F}^{\mathrm{T}}x_n\|^2 = \mathrm{O}(1)$, $\|\mathbf{F}^{\mathrm{T}}A_n\mathbf{F}\|_{\mathrm{F}}^2 = \mathrm{O}(1)$ and by Assumption 6Φ we have $\lim_{n\to\infty}\mathrm{Var}\,\frac{1}{\sqrt{n_*}}\mathcal{S}^*\alpha = \alpha^{\mathrm{T}}\Sigma_{\mathcal{S}^*}\alpha$. If $\alpha^{\mathrm{T}}\Sigma_{\mathcal{S}^*}\alpha > 0$ then Theorem C.1 can be used to deduce that $\frac{1}{\sqrt{n_*}}\mathcal{S}^*\alpha$ converges in distribution to $\mathbf{N}\big(0, \alpha^{\mathrm{T}}\Sigma_{\mathcal{S}^*}\alpha\big)$.

The rest of the proof proceeds by exactly the same argument as Theorem 2.2 with $\mathrm{P} = \mathrm{I}_{k+d+1}$. $\square$

3.3 Relation to the ordinary demeaning procedure

The fixed effect elimination approach described in Section 3.1 can be reformulated to use the classical demeaning operator $M_{\mathcal{K}} = \mathbf{F}^{\mathrm{T}}\mathbf{F}$ rather than the $\mathbf{F}$ matrix itself. Firstly, let us denote $Y_n^\dagger = \mathbf{F}^{\mathrm{T}}Y_n^*$, $X_n^\dagger = \mathbf{F}^{\mathrm{T}}X_n^*$ and $\mathbf{W}_n^\dagger = \langle\mathbf{F}^{\mathrm{T}}W_{n,r}^*\mathbf{F}\rangle_{r\leq d}$. Transforming the log-likelihood function in (3.4) we obtain

$$\log L_n^\dagger(\theta) = -\frac{n_*}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|\mathrm{pdet}\left(M_{\mathcal{K}} - \lambda^{\mathrm{T}}\mathbf{W}_n^\dagger\right)\right| - \frac{1}{2\sigma^2}\left\|Y_n^\dagger - \lambda^{\mathrm{T}}\mathbf{W}_n^\dagger - X_n^\dagger\beta\right\|^2, \quad (3.5)$$

where $\mathrm{pdet}(A)$ denotes pseudo-determinant, i.e. the product of all non-zero singular values of matrix A. The function in (3.5) is a proper Gaussian log-likelihood of θ, given $Y_n^\dagger = M_{\mathcal{K}}Y_n$ and $X_n^\dagger = M_{\mathcal{K}}X_n$, with respect to n_*-dimensional Lebesgue measure on $\mathcal{K}^\perp$. Furthermore, notice that $M_{\mathcal{K}} = P_{\mathcal{K}^\perp}$ is an identity on the space $\mathcal{K}^\perp$, thus $\Delta_n^\dagger(\lambda) = M_{\mathcal{K}} - \lambda^{\mathrm{T}}\mathbf{W}_n^\dagger$ is a proper spatial difference operator on $\mathcal{K}^\perp$. Numerically, values of $L_n^\dagger$ in (3.5) and L_n^* in (3.4) are the same, however, an advantage of using $L_n^\dagger$ is that not only $Y_n^\dagger$ and $X_n^\dagger$ but also the matrix $\mathbf{W}_n^\dagger$ does not depend on a particular choice of $\mathbf{F}$.

By the determinant formula for block matrices we have

$$\begin{aligned}\det\Delta_n(\lambda) &= \det\left(\begin{bmatrix}\mathbf{F}\\ \mathbf{E}\end{bmatrix}\cdot\Delta_n(\lambda)\cdot\left[\mathbf{F}^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}}\right]\right)\\ &= \det\left(\mathbf{F}\Delta_n(\lambda)\mathbf{F}^{\mathrm{T}}\right)\cdot\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right)\\ &= \mathrm{pdet}\left(\Delta_n^\dagger(\lambda)\right)\cdot\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right),\end{aligned}$$

where $\mathbf{E}$ is a matrix of orthonormal columns spanning $\mathcal{K}$. Thus, given the value of $\det\Delta_n(\lambda)$, computation of the value of $\mathrm{pdet}\left(\Delta_n^\dagger(\lambda)\right)$ in (3.5) amounts to evaluation of the numerical value of the determinant of $\dim\mathcal{K}\times\dim\mathcal{K}$ matrix $\mathbf{E}\Delta_n\mathbf{E}^{\mathrm{T}}$. Lastly, Theorems 3.1 and 3.2 imply that maximising $L_n^\dagger$ with respect to θ gives consistent, asymptotically normal estimates.

For some specifications the value of $\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right)$ can be found analytically. This is true, for example, in two special cases: spatial fixed effects and time fixed effects, as considered in Lee, L. F. and Yu, J. (2010a). In paricular, in a balanced panel setting, with $n = N\cdot T$, a time invariant vector of matrices $\mathbf{W}_n = \bar{\mathbf{W}}_N\otimes \mathrm{I}_T$ and a usual matrix Φ_n of spatial unit dummy variables, the space $\mathrm{span}(\Phi_n)$ is already $\mathbf{W}_n$-invariant. Thus we have $\mathcal{K} = \mathrm{span}(\Phi_n)$. Finally, it can be also observed that $\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right) = \det\left(\mathrm{I}_N - \lambda^{\mathrm{T}}\bar{\mathbf{W}}_N\right)$. Similarly, if the matrices in $\mathbf{W}_n$ are additionally row-normalized and the matrix Φ_n incorporates both time and spatial fixed effects, we have $\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right) = (1-\sum_{r\le d}\lambda_r)^{T-1}\det\left(\mathrm{I}_N - \lambda^{\mathrm{T}}\bar{\mathbf{W}}_N\right)$, with $\lambda = (\lambda_r)_{r\le d}$. Moreover, we note that in the case of time fixed effects specification, it can be shown that $\det\left(\mathbf{E}\Delta_n(\lambda)\mathbf{E}^{\mathrm{T}}\right) = (1-\sum_{r\le d}\lambda_r)^{T}$.

We'll briefly take a closer look at the individual fixed effect case. A standard approach to ML estimation is to maximise the ordinary log-likelihood with Y_n and X_n replaced with $Y_n^\dagger$ and $X_n^\dagger$, i.e. to maximise the expression

$$-\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|\det\Delta_n(\lambda)\right| - \frac{1}{2\sigma^2}\left\|Y_n^\dagger - \lambda^{\mathrm{T}}\mathbf{W}_n^\dagger - X_n^\dagger\beta\right\|^2, \tag{3.6}$$

as described in e.g. Elhorst, J. P. (2014), Section 3.1.1. Note that for the spatial fixed effects specification we have $n_* = NT - N$ and $\mathrm{pdet}\left(\Delta_n^\dagger(\lambda)\right) = \det\Delta_n(\lambda)^{\frac{T-1}{T}}$. By concentrating out both β and σ^2 from (3.5) and (3.6), we can observe that both expressions give the same value for maximisers in variables λ and β. However, the QML estimate of σ^2 from (3.5) is $\frac{T}{T-1}$ times larger than that maximising (3.6). This readily implies a multiplicative bias correction for the standard approach estimate of σ^2. Furthermore, Theorem 3.2 implies that, up to the factor, the estimates are also asymptotically normal. This bias correction has been originally suggested in Lee, L. F. and Yu, J. (2010a). However, their proof of the asymptotic normality relies on a false statement concerning the Taylor expansion of the gradient of the log-likelihood. Our Theorem 3.1 may be the first to formally prove validity of their result, by yielding it as its special case.[23]

4 Conclusions

In this paper we have introduced an improved analysis of the asymptotic behaviour of the well known Gaussian QML estimator for higher-order SAR models. Among other things, our approach allows one to consider a wider range of distributions for the vector of innovations. In particular, weaker conditions on the existence of moments are imposed and, as a result, heavier tailed distributions may be considered. Moreover, elements of the error term do not need to be identically distributed as long as their excess kurtosis is uniformly bounded, see Assumptions 2 and 2' for precise formulation. Furthermore, we have addressed the problem of the estimator's consistency and asymptotic distribution under the general form of the parameter space for the autoregressive parameter.

We have argued that our asymptotic analysis covers a fundamentally larger class spatial weight matrices in model specification. Importantly, this makes it possible to account for spatial interaction patterns with larger amount of spatial dependence. Additionally, as has been demonstrated by the sample application in Section 3, our approach also amplifies the theoretical usability of the asymptotic analysis. In particular, it extends the set of possible

[23]We should note that an adaptation of our reasoning to the case of a SAR model with a spatially autocorrelated error term should not pose excessive difficulty. However, to retain simplicity of the argument, we focus on the specification (2.1).

model transformations that can be applied to the original model specification, without violating the crucial boundedness requirement expressed in Assumption 4. As an example, we have developed statements on large sample behaviour of QML estimates under higher-order SAR general fixed effect specifications. Those results would not be possible without the virtues of our improved asymptotic analysis.

We believe that the large sample analyses of Gaussian QML estimators based on the original ideas in Lee, L. F. (2004), could benefit from applying our improved theorems, either directly or with minor modifications. Furthermore, similar ideas, in particular our Theorem C.1, can be used to reconsider large sample statements on estimators other than QML, for example GMM or 2SLS, see Lee, L. F. (2007), Lee, L. F. and Liu, X. (2010) and Lee, L. F. and Yu, J. (2014). We should also mention that we have made every effort to keep our reasoning mathematically rigorous. As a result, we have avoided some of the oversights present in the argument of the standard analysis. For example, we properly derive the asymptotic distribution based on the Cramér-Wald theorem. Moreover, our proof does not rely on the existence of the Lagrange remainder in the Taylor expansion theorem.[24] Because of this deficiency, the original reasoning in Lee, L. F. (2004) might be considered unsatisfactory and our proof may constitute the first formally complete one.

A Remarks

Proof of Remark 2.1. For any sequence of $n \times n$ matrices $(A_n)_{n\in\mathbb{N}}$ define $\|(A_n)_{n\in\mathbb{N}}\|_{\mathcal{A}} = \sup_{n\in\mathbb{N}} \|A_n\|$. The set $\mathcal{A} = \{(A_n)_{n\in\mathbb{N}} \colon \|(A_n)_{n\in\mathbb{N}}\|_{\mathcal{A}} < \infty\}$, equipped with element-wise addition and multiplication, as well as the norm $\|\cdot\|_{\mathcal{A}}$, is a unital Banach algebra.[25] By Proposition 1.7 in Takesaki, M. (1979), the set $G(\mathcal{A})$ of all invertible elements in $\mathcal{A}$ is open in $\mathcal{A}$. The map $\Delta \colon \mathbb{R}^d \to \mathcal{A}$, given by $\Delta(\lambda) = (\mathrm{I}_n - \lambda^T \mathbf{W}_n)_{n\in\mathbb{N}}$, is continuous, thus the pre-image $V = \Delta^{-1}(G(\mathcal{A}))$ is open in $\mathbb{R}^d$. The norm $\|\cdot\|_{\mathcal{A}}$ is a continuous function on $\mathcal{A}$ and the map $G(\mathcal{A}) \ni A \mapsto A^{-1}$ is also continuous, see (Takesaki, M., 1979, Corrolary 1.8). Thus, the map $D \colon V \to \mathbb{R}$, $D(\lambda) = \|(\mathrm{I}_n - \lambda^{\mathrm{T}} \mathbf{W}_n)^{-1}_{n\in\mathbb{N}}\|_{\mathcal{A}}$, is continuous and the pre-image $U = D^{-1}((-1, K_\Lambda + 1))$ is open in V, thus open in $\mathbb{R}^d$. By Assumptions 3 and 4 we have $\Lambda \subset U$. □

Remark A.1 Let $U_\Lambda \supset \Lambda$ be the set given in Remark 2.1. The function $\log L_n$, given in (2.2), is thrice differentiable in each $\lambda \in U_\Lambda$. To simplify notation let us denote $\tilde{W}^\lambda_{n,r} = W_{n,r}\Delta_n^{-1}(\lambda)$ and $\varepsilon_n(\lambda, \beta) = Y_n - \lambda^{\mathrm{T}} \mathbf{W}_n Y - X_n\beta$. First order partial derivatives of $\log L_n$ are

$$\frac{\partial \ln L_n(\theta)}{\partial \lambda} = \left[-\operatorname{tr} \tilde{W}^\lambda_{n,r} + \frac{1}{\sigma^2} \varepsilon_n(\lambda,\beta)^{\mathrm{T}} W_{n,r} Y_n \right]_{r\le d},$$

$$\frac{\partial \ln L_n(\theta)}{\partial \beta} = \frac{1}{\sigma^2} \varepsilon_n(\lambda,\beta)^{\mathrm{T}} X_n,$$

$$\frac{\partial \ln L_n(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|\varepsilon_n(\lambda,\beta)\|^2 .$$

The second and third order partial derivatives are

$$\frac{\partial^2 \ln L_n(\theta)}{\partial \lambda^2} = \left[-\operatorname{tr}\left(\tilde{W}^\lambda_{n,r_1} \tilde{W}^\lambda_{n,r_2} \right) - \frac{1}{\sigma^2} \left(W_{n,r_1} Y_n \right)^{\mathrm{T}} W_{n,r_2} Y_n \right]_{r_1,r_2 \le d},$$

$$\frac{\partial^2 \ln L_n(\theta)}{\partial \lambda \partial \beta} = \left[-\frac{1}{\sigma^2} X_n^{\mathrm{T}} W_{n,r} Y_n \right]_{r\le d},$$

[24] Recall that the Lagrange reminder in Taylor series expansion of a vector valued function is not available. The function $f : [0,1] \to \mathbb{C}$, $f(t) = e^{it}$, $t \in [0,1]$, can serve as a counterexample. See also Feng, C. *et. al.* (2014). Instead, our technique makes use of an original bound (inequalities (2.9), (2.10)).

[25] It is a complete normed space with a submultiplicative norm and $(\mathrm{I}_n)_{n\in\mathbb{N}}$ as identity.

$$
\begin{aligned}
\frac{\partial^2 \ln L_n(\theta)}{\partial\lambda\partial(\sigma^2)} &= \left[-\frac{1}{\sigma^4}\varepsilon_n(\lambda,\beta)^{\mathrm{T}} W_{n,r} Y_n\right]_{r\le d},\\
\frac{\partial^2 \ln L_n(\theta)}{\partial\beta^2} &= -\frac{1}{\sigma^2} X_n^{\mathrm{T}} X_n,\\
\frac{\partial^2 \ln L_n(\theta)}{\partial\beta\partial(\sigma^2)} &= -\frac{1}{\sigma^4}\varepsilon_n(\lambda,\beta)^{\mathrm{T}} X_n,\\
\frac{\partial^2 \ln L_n(\theta)}{\partial(\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\left\|\varepsilon_n(\lambda,\beta)\right\|^2,\\
\frac{\partial^3 \ln L_n(\theta)}{\partial\lambda^3} &= \left[-\operatorname{tr}\left(\tilde W^{\lambda}_{n,r_1}\tilde W^{\lambda}_{n,r_3}\tilde W^{\lambda}_{n,r_2} + \tilde W^{\lambda}_{n,r_1}\tilde W^{\lambda}_{n,r_2}\tilde W^{\lambda}_{n,r_3}\right)\right]_{r_1,r_2,r_3\le d},\\
\frac{\partial^3 \ln L_n(\theta)}{(\partial\lambda)^2\partial(\sigma^2)} &= \left[\frac{1}{\sigma^4}\left(W_{n,r_1}Y_n\right)^{\mathrm{T}} W_{n,r_2}Y_n\right]_{r_1,r_2\le d},\\
\frac{\partial^3 \ln L_n(\theta)}{\partial\lambda\partial\beta\partial(\sigma^2)} &= \left[\frac{1}{\sigma^4}X_n^{\mathrm{T}} W_{n,r}Y_n\right]_{r\le d},\\
\frac{\partial^3 \ln L_n(\theta)}{(\partial\beta)^2\partial(\sigma^2)} &= \frac{1}{\sigma^4}X_n^{\mathrm{T}} X_n,\\
\frac{\partial^3 \ln L_n(\theta)}{\partial\beta\partial(\sigma^2)^2} &= \frac{2}{\sigma^6}X_n^{\mathrm{T}}\varepsilon_n(\lambda,\beta),\\
\frac{\partial^3 \ln L_n(\theta)}{\partial(\sigma^2)^3} &= -\frac{n}{\sigma^6} + \frac{3}{\sigma^8}\left\|\varepsilon_n(\lambda,\beta)\right\|^2.
\end{aligned}
$$

Moreover, the following derivatives vanish

$$
\frac{\partial^3 \ln L_n(\theta)}{\partial\lambda\partial\beta^2},\quad \frac{\partial^3 \ln L_n(\theta)}{\partial\beta\partial\lambda^2},\quad \frac{\partial^3 \ln L_n(\theta)}{\partial\beta^3}.
$$

Partial derivatives which are not mentioned are given by symmetry.

Remark A.2 Let $e_n = e_n(\theta)$, $\theta = \left(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \sigma^2\right)^{\mathrm{T}}$, be an element of any of the matrices representing first, second or third order derivatives of the function $\log L_n$, c.f. Remark A.1. Then, under Assumptions 1 and 4, e_n is of the form

$$
e_n = \varepsilon_n^{\mathrm{T}} A_n \varepsilon_n + x_n^{\mathrm{T}}\varepsilon_n + z_n,
$$

where A_n, x_n, z_n are non-random continuous functions of θ, for which there exist a universal continuous function $K(\beta,\sigma^2)$, independent of n, satisfying $\|A_n\|^2 \le K(\beta,\sigma^2)$ and $\|x_n\|^2, \|z_n\|^2 \le n\cdot K(\beta,\sigma^2)$.

Proof. We will say that a random vector $E_n = E_n(\theta) \in \mathbb{R}^n$ is an amenable vector if $E_n = A_n\varepsilon_n + z_n$, where A_n and z_n are some non-random continuous functions of θ as in the statement of the theorem. Note that ε_n, X_n, $\varepsilon(\lambda,\beta)$, $W_{n,r}Y_n$ are amenable vectors. Then, by examining formulas in Remark A.1, it can be observed that all entries e_n are of the form $e_n = z_n^{\mathrm{T}} z_n + f(\sigma^2)E_n^{\mathrm{T}} F_n$, where E_n, F_n are amenable vectors and z_n is a non-random vector as in the statement of the theorem and f is a continuous function of σ^2. Changing the choice of K if necessary, we complete the proof. □

Remark A.3 Let θ_0 be fixed. Under Assumptions 1 – 4 and 6 we have $\left\|\mathfrak{I} + \frac{1}{n}\frac{\partial^2 \ln L_n}{\partial\theta^2}(\theta_0)\right\| = o_{\mathbb{P}}(1)$.

Proof. In view of Assumption 6, it suffices to show that each element $e_n = e_n(\theta)$ of the matrix representing $\frac{1}{n}\frac{\partial^2 \ln L_n}{\partial \theta^2}(\theta)$ satisfies $\operatorname{Var}_{\theta_0} e_n(\theta_0) = \mathrm{o}_{\mathbb{P}}(1)$. Using Remark A.2, we have

$$\operatorname{Var}_{\theta_0}(e_n) \leq 2\operatorname{Var}_{\theta_0}\left(\frac{1}{n}\varepsilon_n^{\mathrm{T}} A_n \varepsilon_n\right) + 2\operatorname{Var}_{\theta_0}\left(\frac{1}{n} x_n^{\mathrm{T}} \varepsilon_n\right).$$

Then, it is enough to use Lemma B.2 to establish the bounds for $\operatorname{Var}_{\theta_0} e_n(\theta_0)$. □

Remark A.4 Let θ_0 be fixed and let $\tau = \{\beta \in \mathbb{R}^k \colon \|\beta\| < \|\beta_0\| + 1\} \times U_\Lambda \times (\frac{1}{2}\sigma_0^2, 2\sigma_0^2)$ with U_Λ given in Remark 2.1. Moreover, let $e_n = e_n(\theta)$ be an element of the matrix representing third order derivative of the function $\frac{1}{n}\log L_n$. Then, under Assumptions 1 – 4, the quantity $\mathbb{E}_{\theta_0}\sup_{\theta\in\tau}|e_n(\theta)|$ is bounded in $n \in \mathbb{N}$.[26]

Proof. By Remark A.2 we have

$$\begin{aligned}\mathbb{E}_{\theta_0}\sup_{\theta\in\tau}|e_n(\theta)| &\leq \frac{1}{n}\mathbb{E}_{\theta_0}\sup_{\theta\in\tau}\left(\|A_n\|\|\varepsilon_n\|^2 + \|x_n\|\|\varepsilon_n\| + \|z_n\|\right)\\ &\leq 3(1+\sigma_0^2)\max_{\theta\in\operatorname{closure}(\tau)}\sqrt{K(\beta,\sigma^2)} < \infty.\end{aligned}$$

□

Remark A.5 Let θ_0 be fixed and let $\tau = \{\beta \in \mathbb{R}^k \colon \|\beta\| < \|\beta_0\| + 1\} \times U_\Lambda \times (\frac{1}{2}\sigma_0^2, 2\sigma_0^2)$ with U_Λ given in Remark 2.1. Under Assumptions 1 – 4 we have $\sup_{\theta\in\tau}\left\|\frac{1}{\sqrt{n}}\frac{\partial \ln L_n}{\partial\theta}\right\| = \mathrm{O}_{\mathbb{P}}(1)$.

Proof. As $\mathbb{E}_{\theta_0}\frac{\partial \ln L_n}{\partial \theta} = 0$, it is sufficient to observe that, by Remark A.2 and Lemma B.2, for any element e_n of the vector $\frac{1}{\sqrt{n}}\frac{\partial \ln L_n}{\partial\theta}$ we have

$$\operatorname{Var}_{\theta_0}(e_n) \leq \frac{2}{n}\operatorname{Var}_{\theta_0}\left(\varepsilon_n^{\mathrm{T}} A_n\varepsilon_n\right) + \frac{2}{n}\operatorname{Var}_{\theta_0}\left(x_n^{\mathrm{T}}\varepsilon_n\right) \leq CK(\beta,\sigma^2)(1+\sup_{n,i}\mathbb{E}\,\bar{\varepsilon}_{n,i}^4),$$

for some constant C. Then, $\sup_{\theta\in\tau}\operatorname{Var}_{\theta_0}(e_n(\theta)) < \infty$. □

Remark A.6 The functions $\bar{R}_n \colon U_\Lambda \to \mathbb{R}$ given by formula (2.5) are differentiable. Moreover, under Assumptions 1 and 4, both $\|\bar{R}_n\|$ and $\|(\bar{R}_n)'\|$ are uniformly bounded by a constant depending only on β_0 and σ_0^2.

Proof. Boundedness of $\bar{R}_n$ follows from Assumptions 1, 4 and Remark 2.1. By Jacobi's formula, for any $\lambda \in U_\Lambda$, we have

$$\frac{1}{n}\frac{\partial}{\partial\lambda}\ln|\det\Delta_n(\lambda)| = \left(\frac{-1}{n}\operatorname{tr}\left(\Delta_n(\lambda)^{-1}W_{n,r}\right)\right)_{r\leq d}.$$

As $|\frac{1}{n}\operatorname{tr}(A)| \leq \|A\|$, for any $n\times n$ matrix A, by Remark 2.1 we have the boundedness of $\left\|\frac{1}{n}\frac{\partial}{\partial\lambda}\ln|\det\Delta_n(\lambda)|\right\|$ in $\lambda\in U_\Lambda$, $n\in\mathbb{N}$.

Let us note that for any $1 \leq r \leq d$ and $\lambda = (\lambda_j)_{j\leq d} \in U_\Lambda$ we have

$$\begin{aligned}\left|\frac{\partial}{\partial\lambda_r}\frac{\sigma_0^2}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2\right| &\leq \sigma_0^2\left\|\frac{\partial}{\partial\lambda_r}\left(S_n(\lambda)^{\mathrm{T}}S_n(\lambda)\right)\right\|\\ &= \sigma_0^2\left\|\left(W_{n,r}\Delta_n(\lambda_0)\right)^{\mathrm{T}}S_n(\lambda) + S_n(\lambda)^{\mathrm{T}}W_{n,r}\Delta_n(\lambda_0)\right\|.\end{aligned}$$

[26] Note that the integrand is measurable as e_n is continuous in θ.

Thus, by Remark 2.1, $\frac{\sigma_0^2}{n}\frac{\partial}{\partial\lambda}\|S_n(\lambda)\|_{\mathrm{F}}^2$ is bounded in $\lambda \in U_\Lambda$ and $n \in \mathbb{N}$. Similarly, for any $1 \le r \le d$ and $\lambda \in U_\Lambda$ we have

$$\frac{1}{n}\frac{\partial}{\partial\lambda_r}\|M_{X_n}S_n(\lambda)X_n\beta_0\|^2 = \\ -\frac{2}{n}\beta_0^{\mathrm{T}}X_n^{\mathrm{T}}S_n(\lambda)^{\mathrm{T}}M_{X_n}W_{n,r}\Delta_n(\lambda_0)^{-1}X_n\beta_0.$$

Again, by Remark 2.1, $\frac{1}{n}\frac{\partial}{\partial\lambda}\|M_{X_n}S_n(\lambda)X_n\beta_0\|^2$ is bounded in $\lambda \in U_\Lambda$, $n \in \mathbb{N}$.

Finally, as $\frac{1}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2 \ge \|S_n(\lambda)^{-1}\|^{-2}$, for $\lambda \in U_\Lambda$, Remark 2.1 implies that $\frac{1}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2 > \delta > 0$, $\lambda \in U_\Lambda$, for some constant δ. Thus the derivative of $\ln\left(\frac{1}{n}\|M_{X_n}S_n(\lambda)X_n\beta_0\|^2 + \frac{\sigma_0^2}{n}\|S_n(\lambda)\|_{\mathrm{F}}^2\right)$ and, as a result, the derivative of $\bar{R}_n$ itself is bounded on U_Λ. □

B Lemmata

Lemma B.1. *Let $U \subset \mathbb{R}^d$ be an open set and $A \subset U$ its compact subset. If $F\colon U \to \mathbb{R}^m$ is differentiable and $\|F\|, \|F'\| \le M < \infty$ for a constant M, then F is Lipschitz continuous on A with a constant $K_{\mathrm{L}} = K_{\mathrm{L}}(M, A)$.*[27]

Proof. Set $\delta = \inf\{\|x-\xi\|\colon x \in A \text{ and } \xi \in \mathbb{R}^d \setminus U\}$. As A is compact we have $\delta > 0$. Let $x, y \in A$. If $\|x-y\| \le \delta$, then the line segment $\overline{x,y}$ is a subset of U and $\|F(x)-F(y)\| \le \|\int_0^1 \frac{\partial}{\partial t}F(x+t(y-x))\,\mathrm{d}t\| \le M \cdot \|x-y\|$. If $\|x-y\| > \delta$, then we have $\|F(x)-F(y)\| \le \frac{2M}{\delta}\|x-y\|$. □

Lemma B.2. *Let $\varepsilon_n = (\varepsilon_{n,i})_{i\le n}^{\mathrm{T}}$ be an $n \times 1$ random vector satisfying* Assumption 2. *Let $(A_n)_{n\in\mathbb{N}}$, $(P_n)_{n\in\mathbb{N}}$ be sequences of $n \times n$ matrices satisfying $\sup_{n\in\mathbb{N}}\|P_n\| \le 1$ and $P_n = P_n^{\mathrm{T}}P_n$. Moreover, let $x_n \in \mathbb{R}^n$, for $n \in \mathbb{N}$, be non-random vectors satisfying $\|x_n\|^2 = \mathrm{O}(n)$. Then*

(a) *for $Z_n^{\mathrm{a}} = \frac{1}{n}x_n^{\mathrm{T}}A_n\varepsilon_n$, we have $\operatorname{Var} Z_n^{\mathrm{a}} \le \frac{\sigma^2}{n}\|A_n\|^2 \cdot \frac{1}{n}\|x_n\|^2$,*

(b) *for $Z_n^{\mathrm{b}} = \frac{1}{n}\varepsilon_n^{\mathrm{T}}A_n\varepsilon_n$, we have $\operatorname{Var} Z_n^{\mathrm{b}} \le \frac{3}{n}\|A_n\|^2\sup_{n,i}\mathbb{E}\,\bar{\varepsilon}_{n,i}^4$,*

(c) *for $Z_n^{\mathrm{c}} = \frac{1}{n}\varepsilon_n^{\mathrm{T}}A_n^{\mathrm{T}}P_nA_n\varepsilon_n$, we have $\operatorname{Var} Z_n^{\mathrm{c}} \le \frac{3}{n}\|A_n\|^4\sup_{n,i}\mathbb{E}\,\bar{\varepsilon}_{n,i}^4$ and $|\mathbb{E}\,Z_n^{\mathrm{c}}| \le \frac{\sigma^2}{n}\|A_n\|^2\|P_n\|_{\mathrm{F}}^2$.*

Proof. To prove part **(a)**, it is enough to observe that

$$\operatorname{Var}\left(x_n^{\mathrm{T}}A_n\varepsilon_n\right) = \mathbb{E}\,\varepsilon_nA_n^{\mathrm{T}}x_nx_n^{\mathrm{T}}A_n\varepsilon_n = \sigma^2\|A_n^{\mathrm{T}}x_n\|_{\mathrm{F}}^2 \le \sigma^2\|A_n^{\mathrm{T}}\|^2\|x_n\|^2.$$

In proving part **(b)**, we may assume $\tilde{A}_n = \mathrm{E}_n^{\mathrm{T}}A_n\mathrm{E}_n = (\tilde{a}_{ij})_{i,j\le n}$ to be symmetric, as $\left\|\frac{\tilde{A}_n^{\mathrm{T}}+\tilde{A}_n}{2}\right\| \le \|\tilde{A}_n\|$, c.f. Assumption 2. Then,

$$\mathbb{E}\left((\varepsilon_n^{\mathrm{T}}A_n\varepsilon_n)^2\right) = \mathbb{E}\left((\bar{\varepsilon}_n^{\mathrm{T}}\tilde{A}_n\bar{\varepsilon}_n)^2\right) = \sum_{i,j,k,l\le n}\tilde{a}_{ij}\tilde{a}_{kl}\,\mathbb{E}\,\bar{\varepsilon}_{n,i}\bar{\varepsilon}_{n,j}\bar{\varepsilon}_{n,k}\bar{\varepsilon}_{n,l}$$
$$= \sigma^4\cdot\left(\sum_{i\ne j}(\tilde{a}_{ii}\tilde{a}_{jj}+\tilde{a}_{ij}\tilde{a}_{ij}+\tilde{a}_{ij}\tilde{a}_{ji})\right) + \sum_{i\le n}\tilde{a}_{ii}^2\,\mathbb{E}\,\bar{\varepsilon}_{n,i}^4,$$

$$\left(\mathbb{E}\,\varepsilon_n^{\mathrm{T}}A_n\varepsilon_n\right)^2 = \left(\sum_{i,j\le n}\tilde{a}_{ij}\,\mathbb{E}\,\bar{\varepsilon}_{n,i}\bar{\varepsilon}_{n,j}\right)^2 = \sigma^4\sum_{i,j\le n}\tilde{a}_{ii}\tilde{a}_{jj}.$$

[27] In fact, it is not necessary to assume that F is bounded, yet it considerably simplifies the proof.

Thus, with $\|\mathrm{E}_n\| = 1$ by Bessels's inequality, we have

$$\operatorname{Var} \varepsilon_n^{\mathrm{T}} A_n \varepsilon_n = \mathbb{E}\left((\varepsilon_n^{\mathrm{T}} A_n \varepsilon_n)^2\right) - \left(\mathbb{E}\, \varepsilon_n^{\mathrm{T}} A_n \varepsilon_n\right)^2$$
$$= \sum_{i\le n} \tilde{a}_{ii}^2 (\mathbb{E}\, \bar{\varepsilon}_{n,i}^4 - 3\sigma^4) + 2\sigma^4 \|\tilde{A}_n\|_{\mathrm{F}}^2 \le 3n\|A_n\|^2 \sup_{n,i} \mathbb{E}\, \bar{\varepsilon}_{n,i}^4.$$

Finally, the part **(c)** follows from **(b)** and the observation that $|\mathbb{E}\, Z_n^{\mathrm{a}}| = \frac{\sigma^2}{n} \operatorname{tr} A_n^{\mathrm{T}} P_n A_n = \frac{\sigma^2}{n}\|P_n A_n\|_{\mathrm{F}}^2 \le \frac{\sigma^2}{n}\|A_n\|^2\|P_n\|_{\mathrm{F}}^2$ □

C Theorems

Theorem C.1. *Let* $(\varepsilon_n)_{n\in\mathbb{N}}$ *satisfy* Assumption 2' *and let* $x_n = (x_{n,i})_{i\le n}$, $n \in \mathbb{N}$, *be column vectors. Denote* $Q_n = \varepsilon_n^{\mathrm{T}}(x_n + A_n\varepsilon_n)$ *and assume that* $\operatorname{Var} Q_n > 0$ *for sufficiently large* $n \in \mathbb{N}$. *If* $\|x_n\|^2 + \|A_n\|_{\mathrm{F}}^2 = \mathrm{O}(\operatorname{Var} Q_n)$, $\|A_n\|^2 = \mathrm{o}(\operatorname{Var} Q_n)$ *and* $\max_{i\le n} x_{n,i}^2 = \mathrm{o}(\operatorname{Var} Q_n)$, *then* $\frac{Q_n - \mathbb{E}\, Q_n}{\sqrt{\operatorname{Var} Q_n}}$ *converges in distribution to standardised normal* $\mathbf{N}(0,1)$.

Proof. Without loss of generality we may assume that $\sigma_0^2 = 1$ and the matrices A_n, $n \in \mathbb{N}$, are symmetric. Let us denote $\varsigma_{n,i}^2 = \operatorname{Var} \varepsilon_{n,i}^2$, $\eta_{n,i}^3 = \mathbb{E}\, \varepsilon_{n,i}^3$ and $\mathcal{V}_n = \operatorname{Var} Q_n$, for $n \in \mathbb{N}$, $i \le n$. With $A_n = [a_{n,ij}]_{i,j\le n}$, setting

$$U_{n,i} = a_{n,ii}\left(\varepsilon_{n,i}^2 - 1\right) + x_{n,i}\varepsilon_{n,i} + 2\varepsilon_{n,i} \sum_{1\le j<i} a_{n,ij}\varepsilon_{n,j},$$

for $n \in \mathbb{N}$ and $1 \le i < n$,we can write $\sum_{i\le n} U_{n,i} = Q_n - \mathbb{E}\, Q_n$. Observe that $\mathbb{E}\,[U_{n,i} \mid \mathcal{F}_{n,i-1}] = 0$, with σ-fields $\mathcal{F}_{n,i} = \sigma\{\varepsilon_{n,j} \colon j \le i\}$, $1 \le i \le n$. The proof relies on the martingale difference array CLT which can be found in Hall, P. and Hyde, C. C. (1980), Corollary 3.1. According to the theorem, it is enough to show that

$$\forall_{\delta>0}\ q_n(\delta) \to 0 \quad \text{and} \quad V_n \to 1,$$

in probability, as $n \to \infty$, where

$$q_n(\delta) = \frac{1}{\mathcal{V}_n} \sum_{i\le n} \mathbb{E}\left[U_{n,i}^2 \mathbb{I}_{\{U_{n,i}^2 \ge \delta^2 \cdot \operatorname{Var} Q_n\}} \mid \mathcal{F}_{n,i-1}\right], \quad \delta > 0,$$

$$V_n = \frac{1}{\mathcal{V}_n} \sum_{i\le n} \mathbb{E}\left[U_{n,i}^2 \mid \mathcal{F}_{n,i-1}\right].$$

Denote $T_{n,i} = 2\sum_{j<i} a_{i,j}\varepsilon_{n,j}$. Direct computation reveals that

$$U_{n,i}^2 = 2a_{n,ii}x_{n,i}\varepsilon_{n,i}(\varepsilon_{n,i}^2 - 1) + 2a_{n,ii}\varepsilon_{n,i}(\varepsilon_{n,i}^2 - 1)T_{n,i}$$
$$+ 2x_{n,i}\varepsilon_{n,i}^2 T_{n,i} + \varepsilon_{n,i}^2 T_{n,i}^2 + a_{n,ii}^2(\varepsilon_{n,i}^2 - 1)^2 + x_{n,i}^2\varepsilon_{n,i}^2$$

and

$$V_n = \frac{1}{\mathcal{V}_n} \sum_{i\le n} \big(2\eta_{n,i}^3 a_{n,ii}x_{n,i} + 2\eta_{n,i}^3 a_{n,ii}T_{n,i} + 2x_{n,i}T_{n,i}$$
$$+ T_{n,i}^2 + \varsigma_{n,i}^2 a_{n,ii}^2 + x_{n,i}^2\big).$$

Since $\mathcal{V}_n = \sum_{i\le n} \mathbb{E}\, U_{n,i}^2$, $n \in \mathbb{N}$, we have

$$\mathcal{V}_n = 2\sum_{i\le n} \eta_{n,i}^3 x_{n,i} a_{n,ii} + \mathbb{E}\, T_{n,i}^2 + \sum_{i\le n} \varsigma_{n,i}^2 a_{n,ii}^2 + \sum_{i\le n} x_{n,i}^2.$$

Further, it follows that

$$(V_n - 1)^2 = \mathcal{V}_n^{-2} \cdot \left(\sum_{i \le n} \left(2\eta_{n,i}^3 a_{n,ii} T_{n,i} + 2x_{n,i} T_{n,i} + \left(T_{n,i}^2 - \mathbb{E}\, T_{n,i}^2 \right) \right) \right)^2,$$

thus $\mathbb{E}\,(V_n - 1)^2 = \frac{1}{\mathcal{V}_n^2} \mathrm{O}\left(\xi_n^{(1)} + \xi_n^{(2)} + \xi_n^{(3)} \right)$, where $\xi_n^{(1)} = \mathbb{E}\left(\sum_{i \le n} a_{n,ii} T_{n,i} \right)^2$, $\xi_n^{(2)} = \mathbb{E}\left(\sum_{i \le n} x_{n,i} T_{n,i} \right)^2$ and $\xi_n^{(3)} = \mathbb{E}\left(\sum_{i \le n} \left(T_{n,i}^2 - \mathbb{E}\, T_{n,i}^2 \right) \right)^2$.

For any numbers i and j denote $i \wedge j = \min\{i, j\}$. Let us note that for some constant $C > 0$ we have

$$\begin{aligned}
\frac{1}{16}\xi_n^{(3)} &= \mathbb{E}\left(\sum_{i \le n} \sum_{k_1, k_2 < i} a_{n,ik_1} a_{n,ik_2} \varepsilon_{n,k_1} \varepsilon_{n,k_2} - \sum_{i \le n} \sum_{k < i} a_{n,ik}^2 \right)^2 \\
&\le 2\,\mathbb{E}\left(\sum_{i \le n} \sum_{\substack{k_1, k_2 < i \\ k_1 \ne k_2}} a_{n,ik_1} a_{n,ik_2} \varepsilon_{n,k_1} \varepsilon_{n,k_2} \right)^2 \\
&\quad + 2\,\mathbb{E}\left(\sum_{i \le n} \sum_{k < i} a_{n,ik}^2 \left(\varepsilon_{n,k}^2 - 1 \right) \right)^2 \\
&\le C \sum_{i,j \le n} \sum_{k_1, k_2 \le i \wedge j} a_{n,ik_1} a_{n,ik_2} a_{n,jk_1} a_{n,jk_2} = C \sum_{i,j \le n} \left(\sum_{k \le i \wedge j} a_{n,ik} a_{n,jk} \right)^2.
\end{aligned}$$

Let $C \cdot B_n$ be the right-hand side of the above inequality. Further, we have

$$\begin{aligned}
B_n &= \sum_{i < j \le n} \left(\sum_{k \le i \wedge j} a_{n,ik} a_{n,jk} \right)^2 + \sum_{j \le i \le n} \left(\sum_{k \le i \wedge j} a_{n,ik} a_{n,jk} \right)^2 \\
&\le 2 \sum_{i,j \le n} \left(\sum_{k \le i} a_{n,ik} a_{n,kj} \right)^2 = 2 \left\| A_n \cdot \left[a_{n,ij} \mathbb{I}_{\{i<j\}} \right]_{i,j \le n} \right\|_{\mathrm{F}}^2 \\
&\le 2 \|A_n\|^2 \|A_n\|_{\mathrm{F}}^2. \qquad \text{(C.1)}
\end{aligned}$$

Moreover, by Schwartz inequality and (C.1) we have

$$\begin{aligned}
\frac{1}{4}\xi_n^{(2)} &= \mathbb{E}\left(\sum_{i \le n} x_{n,i} \sum_{k < i} a_{n,ik} \varepsilon_{n,k} \right)^2 \\
&= \left(\sum_{i,j \le n} x_{n,i} x_{n,j} \sum_{k < i \wedge j} a_{n,ik} a_{n,jk} \right) \\
&\le \sqrt{\sum_{i,j \le n} x_{n,i}^2 x_{n,j}^2} \sqrt{\sum_{i,j \le n} \left(\sum_{k \le i \wedge j} a_{n,ik} a_{n,jk} \right)^2} \le \sum_{i \le n} x_{n,i}^2 \cdot B_n^{1/2} \\
&\le \sqrt{2} \cdot \|A_n\| \|x_n\|^2 \|A_n\|_{\mathrm{F}}. \qquad \text{(C.2)}
\end{aligned}$$

Using (C.2) with $(a_{n,ii})_{i \le n}^{\mathrm{T}}$ substituted for x_n, we obtain $\xi_n^{(1)} \le 4\sqrt{2} \cdot \|A_n\| \|A_n\|_{\mathrm{F}}^3$. Finally, $\mathbb{E}\,(V_n - 1)^2 = \mathcal{V}_n^{-2} \cdot \mathrm{o}\left(\mathcal{V}_n^2\right)$ and $V_n \to 1$ in probability.

To complete the proof we will show that $q_n(\delta) \to 0$, $\delta > 0$, in probability. Since $q_n(\delta) > 0$, by Markov inequality, it is enough to obtain the convergence $\mathbb{E}\, q_n(\delta) \to 0$, for any $\delta > 0$.

Let $K > 0$. For $n \in \mathbb{N}$, $i \leq n$, set $Z_{n,i} = \varepsilon_{n,i}\mathbb{I}_{\{\varepsilon_{n,i}<K\}} - \mathbb{E}\,\varepsilon_{n,i}\mathbb{I}_{\{\varepsilon_{n,i}<K\}}$, $H_{n,i} = \varepsilon_{n,i} - Z_{n,i}$ and $u_{n,i} = 2Z_{n,i}\sum_{j<i} a_{ij}Z_{n,j} + a_{n,ii}\left(Z_{n,i}^2 - \mathbb{E}\,Z_{n,i}^2\right) + x_{n,i}Z_{n,i}$, $h_{n,i} = U_{n,i} - u_{n,i}$. It is easy to observe that both $(Z_{n,i})_{n,i}$ and $(H_{n,i})_{n,i}$ are sequences of independent zero mean variables.

As either $|u_{n,i}| \geq \frac{1}{2}|U_{n,i}|$ or $|u_{n,i}| < \frac{1}{2}|U_{n,i}| < |h_{n,i}|$, we have

$$
\begin{aligned}
\mathbb{E}\, q_n(\delta) &\leq 2\mathcal{V}_n^{-1}\left(\sum_{i\leq n}\mathbb{E}\left[u_{n,i}^2\mathbb{I}_{\{U_{n,i}^2\geq\delta^2\mathcal{V}_n\}}\right] + \sum_{i\leq n}\mathbb{E}\,h_{n,i}^2\right)\\
&\leq 2\mathcal{V}_n^{-1}\left(\sum_{i\leq n}\mathbb{E}\left[u_{n,i}^2\mathbb{I}_{\{(2u_{n,i})^2\geq\delta^2\mathcal{V}_n\}}\right] + 2\sum_{i\leq n}\mathbb{E}\,h_{n,i}^2\right)\\
&\leq 2\left(\delta\mathcal{V}_n\right)^{-2}\sum_{i\leq n}\mathbb{E}\left[(2u_{n,i})^2\cdot u_{n,i}^2\right] + 4\mathcal{V}_n^{-1}\sum_{i\leq n}\mathbb{E}\,h_{n,i}^2. \qquad \text{(C.3)}
\end{aligned}
$$

Since $\frac{1}{27}u_{n,i}^4 \leq 2^4 Z_{n,i}^4\left(\sum_{j<i} a_{n,ij}Z_{n,j}\right)^4 + a_{n,ii}^4(Z_{n,i}^2 - \mathbb{E}\,Z_{n,i}^2)^4 + x_{n,i}^4 Z_{n,i}^4$, it follows that $\sum_{i\leq n}\mathbb{E}\,u_{n,i}^4 = \mathrm{O}\left(\kappa_n^{(1)} + \kappa_n^{(2)} + \kappa_n^{(3)}\right)$, where

$$
\kappa_n^{(1)} = \mathbb{E}\sum_{i\leq n}\left(\sum_{j<i} a_{n,ij}Z_{n,j}\right)^4, \quad \kappa_n^{(2)} = \sum_{i\leq n} a_{n,ii}^4, \quad \kappa_n^{(3)} = \sum_{i\leq n} x_{n,i}^4.
$$

For some constant $C \geq 0$ we have $\kappa_n^{(1)} = C\cdot\mathbb{E}\sum_{i\leq n}\sum_{j,k<i} a_{n,ij}^2 a_{n,ik}^2$ and

$$
\frac{\kappa_n^{(1)}}{C} + \kappa_n^{(2)} \leq 2\sum_{i\leq n}\left(\sum_{j\leq i} a_{n,ij}^2\right)^2 \leq 2\left\|A_n^{\mathrm{T}}\cdot[a_{n,ij}]_{j\leq i}\right\|_{\mathrm{F}}^2 = \mathrm{o}\left(\mathcal{V}_n^2\right).
$$

Additionally, $\kappa_n^{(3)} \leq \left(\max_{i\leq n} x_{n,i}^2\right)\|x_n\|^2 = \mathrm{o}\left(\mathcal{V}_n^2\right)$. Thus, the first summand in the right-hand side of inequality (C.3) converges to 0 as $n \to \infty$, independently of K. Lastly, it is left to show that $\sup_{n\in\mathbb{N}}\mathcal{V}_n^{-1}\sum_{i\leq n}\mathbb{E}\,h_{n,i}^2$ converges to 0 as $K \to \infty$. To this end, observe that

$$
\begin{aligned}
\mathbb{E}\,h_{n,i}^2 \leq 6\,\mathbb{E}&\left(\sum_{j<i} a_{n,ij}\left(\varepsilon_{n,i}\varepsilon_{n,j} - Z_{n,i}Z_{n,j}\right)\right)^2\\
&+ 3\,\mathbb{E}\,a_{n,ii}^2\left(\varepsilon_{n,i}^2 - Z_{n,i}^2 - \left(1 - \mathbb{E}\,Z_{n,i}^2\right)\right)^2 + 3\,\mathbb{E}\,x_{n,i}^2\left(\varepsilon_{n,i} - Z_{n,i}\right)^2
\end{aligned}
$$

and, as $\varepsilon_n = Z_n + H_n$, for $j < i \leq n$ we have

$$
\begin{aligned}
\varepsilon_{n,i}\varepsilon_{n,j} - Z_{n,i}Z_{n,j} &= H_{n,i}H_{n,j} + Z_{n,i}H_{n,j} + H_{n,i}Z_{n,j},\\
\varepsilon_{n,i}^2 - Z_{n,i}^2 &= 2Z_{n,i}H_{n,i} + H_{n,i}^2.
\end{aligned}
$$

Thus, for a constant $C > 0$, we have

$$
\begin{aligned}
\frac{1}{C}\cdot\mathcal{V}_n^{-1}\sum_{i\leq n}\mathbb{E}\,h_{n,i}^2 \leq{}& \mathcal{V}_n^{-1}\|A_n\|_{\mathrm{F}}^2\cdot\sup_{i\leq n}\mathbb{E}\,H_{n,i}^2\cdot\sup_{i\leq n}\mathbb{E}\,Z_{n,i}^2\\
&+ \mathcal{V}_n^{-1}\|A_n\|_{\mathrm{F}}^2\cdot\sup_{i\leq n}\mathbb{E}\,H_{n,i}^4\\
&+ \mathcal{V}_n^{-1}\|A_n\|_{\mathrm{F}}^2\cdot\left(\sup_{i\leq n}\mathbb{E}\,H_{n,i}^4\cdot\sup_{i\leq n}\mathbb{E}\,Z_{n,i}^4\right)^{\frac{1}{2}} + \mathcal{V}_n^{-1}\|x_n\|^2\cdot\sup_{i\leq n}\mathbb{E}\,H_{n,i}^2.
\end{aligned}
$$

Note that $\sup_{i\leq n}\mathbb{E}\,Z_{n,i}^2$, $\sup_{i\leq n}\mathbb{E}\,Z_{n,i}^4$, $\mathcal{V}_n^{-1}\|A_n\|_{\mathrm{F}}^2$, $\mathcal{V}_n^{-1}\|x_n\|^2$ are bounded in $n \in \mathbb{N}$. Moreover, since the fourth powers of $\varepsilon_{n,i}$ are uniformly integrable, both $\sup_{n;i\leq n}\mathbb{E}\,H_{n,i}^2$ and $\sup_{n;i\leq n}\mathbb{E}\,H_{n,i}^4$ converge to 0, as $K \to \infty$. Finally, we conclude that $\mathbb{E}\,q_n(\delta) \to 0$, as $n \to \infty$. $\square$

Theorem C.2. *For any square matrix A we have $\|A\| \le \|A\|_{\mathrm{K}}$.*

Proof. Suppose that $\|A\| > \|A\|_{\mathrm{K}}$. With $B = \frac{1}{\|A\|^2} A^{\mathrm{T}} A$ we have

$$\|B\|_1 \le \frac{\left\|A^{\mathrm{T}} A\right\|_1}{\|A\|^2} \le \frac{1}{\|A\|^2} \left\|A^{\mathrm{T}}\right\|_1 \|A\|_1 = \frac{\|A\|_\infty \|A\|_1}{\|A\|^2} < \frac{\|A\|_{\mathrm{K}}^2}{\|A\|_{\mathrm{K}}^2} = 1.$$

This implies that the series $\mathrm{I}_n + B + B^2 + \dots$ converges to $(\mathrm{I}_n - B)^{-1}$. In particular, the matrix $\mathrm{I}_n - B$ is invertible. Thus, $A^{\mathrm{T}} A - \|A\|^2 = \|A\|^2 (B - \mathrm{I}_n)$ is also non-singular. Since $\|A\|$ is a singular value of A, $\|A\|^2$ is an eigenvalue of $A^{\mathrm{T}} A$, which is a contradiction. □

References

Anselin, L. (1988). Spatial Econometrics: Methods and Models, Kluwer Academic, Dordrecht, doi:10.1007/978-94-015-7799-1.

Anselin, L., Le Gallo, J. and Jayet, H. (2006). Spatial Panel Econometrics, in Matyas, L. and Sevestre, P. (eds.) *The Econometrics of Panel Data, Fundamentals and Recent Developments in Theory and Practice*, 3rd edition, Kluwer, Dodrecht, doi:10.1007/978-3-540-75892-1.

Badinger, H. and Egger, P. (2013). Estimation and testing of higher-order spatial autoregressive panel data error component models. *Journal of Geographical Systems* **15** (4) 453–489, doi:10.1007/s10109-012-0174-z.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*, Chichester West Sussex, England.

Bhansali, R. J., Giraitis, L. and Kokoszka, P. S. (2007). Convergence of quadratic forms with non-vanishing diagonal. *Statistics and Probability Letters* **77** 726–734, doi:10.1016/j.spl.2006.11.007.

Billingsley, P. (1995). *Probability and Measure*, 3nd ed. John Wiley & Sons, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. John Wiley & Sons, New York, doi:10.1002/9780470316962.

Cliff, A. D. and Ord, J. K. (1973). Spatial Autocorrelation, Pion, London.

Cressie, N. (1993). Statistics for Spatial Data, Wiley, New York.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*, Springer, Heidelberg, doi:10.1007/978-3-642-40340-8.

Elhorst, J. P., Lacombe, D. J. and Piras, G. (2014). On model specification and parameter space definitions in higher order spatial econometric models. *Regional Science and Urban Economics* **42** 211–220, doi:10.1016/j.regsciurbeco.2011.09.003.

Feng, C., Wang, H., Han, Y., Xia, Y. and Tu, X. M. (2014). The Mean Value Theorem and Taylors Expansion in Statistics. *The American Statistician* **67** 245–248, doi:10.1080/00031305.2013.844203.

Gupta, A. and Robinson, P. (2015). Inference on higher-order spatial autoregressive models with increasingly many parameters. *Journal of Econometrics* **186** 19–31, doi:10.1016/j.jeconom.2014.12.008.

Hájek, P. and Johanis, M. (2014). *Smooth Analysis in Banach Spaces*, De Gruyter, Berlin.

Hall, P. and Hyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press Inc., New York.

Han, X., Hsieh C. and Lee, L. F. (2017). Estimation and Model Selection of Higher-order Spatial Autoregressive Model: An Efficient Bayesian Approach. *Regional Science and Urban Economics* **63** 97–120, doi:10.1016/j.regsciurbeco.2016.12.003.

Kelejian, H. H. and Prucha, I. R. (2001). On the Asymptotic Distribution of the Moran I Test Statistic with Applications. *Journal of Econometrics* **104** 219–257 doi:10.1016/S0304-4076(01)00064-1.

Lee, L. F. (2004). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica* **72** (6) 1899–1925, doi: 10.1111/j.1468-0262.2004.00558.x.

Lee, L. F. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* **137** 489–514, doi:10.1016/j.jeconom.2005.10.004.

Lee, L. F. and Liu, J. (2010). Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory* **26** 187–230, doi:10.1017/S0266466609090653.

Lee, L. F., Liu, J. and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal* **13** 145–176, doi:10.1111/j.1368-423X.2010.00310.x.

Lee, L. F. and Yu, J. (2010a). Estimation of spatial autoregressive panel data models with fixed effects *Journal of Econometrics* **154** 165–185, doi:10.1016/j.jeconom.2009.08.001.

Lee, L. F. and Yu, J. (2010b). A spatial dynamic panel data model with both time and individual fixed effects *Econometric Theory* **26** 564–597, doi:10.1017/S0266466609100099.

Lee, L. F. and Yu, J. (2013). Near Unit Root in the Spatial Autoregressive Model. *Spatial Economic Analysis* **8** 314–351, doi:10.1080/17421772.2012.760134.

Lee, L. F. and Yu, J. (2014). Efficient GMM estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics* **180** 174–197, doi:10.1016/j.jeconom.2014.03.003.

Le Sage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*, CRC Press, Taylor & Francis Group, Boca Raton.

Li, K. (2017). Fixed-effects dynamic spatial panel data models and impulse response analysis. *Journal of Econometrics* **198** 102–121, doi:10.1016/j.jeconom.2017.02.001.

Liu S. F. and Yang Z. (2015). Modified QML estimation of spatial autoregressive models with unknown heteroskedasticity and nonnormality. *Regional Science and Urban Economics*, **52** 50–70, doi:10.1016/j.regsciurbeco.2015.02.003.

Olejnik, A. and Olejnik, J. (2017). An alternative to Partial Regression in Maximum Likelihood Estimation of Spatial Autoregressive Panel Data Model. *Statistical Review*, in press.

Paelinck, J. and Klaassen, L. (1979). Spatial Econometrics, Saxon House, Farnborough.

PÖTSCHER, M. B. and PRUCHA, I. R. (1997). *Dynamic Nonlinear Econometric Models: Asymptotic Theory*, Springer-Verlag, Berlin Heidelberg.

QU, X. and LEE, L. F. (2017). QML estimation of spatial dynamic panel data models with endogenous time varying spatial weights matrices. *Journal of Econometrics*, **197** 173–201, doi:10.1016/j.jeconom.2016.11.004.

SHI, W. and LEE, L. F. (2017). Spatial dynamic panel data models with interactive fixed effects. *Journal of Econometrics*, **197** 323–347, doi:10.1016/j.jeconom.2016.12.001.

TAKESAKI, M. (1979). *Theory of operator algebras I*, Springer-Verlag, New York.

YU, J., DE JONG, R. and LEE, L. F. (2008). Quasi-Maximum Likelihood Estimators For Spatial Dynamic Panel Data With Fixed Effects When Both n and T Are Large. *Journal of Econometrics* **146** 118–134, doi:10.1016/j.jeconom.2008.08.002.

PRUSS, A. R. (1998). A bounded N-tuplewise independent and identically distributed counterexample to the CLT. *Probability Theory and Related Fields* **111** 323–332, doi:10.1007/s004400050170.